ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute Discussion Paper No. 419

Morals and the Political Economy of Corrective Taxes

Felix Bierbrauer
Marten Ritterrath
Mattias Polborn
Georg Weizsäcker

June 2026

www.econtribute.de

UNIVERSITÄT BONN
UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866 – is gratefully acknowledged.

Morals and the political economy of corrective taxes*

Felix J. Bierbrauer   Mattias Polborn   Marten Ritterrath   Georg Weizsäcker

June 19, 2026

Abstract

We study the political economy of carbon taxes when *neoclassical consumers* take all other agents' emissions as given and *socially responsible consumers* internalize damages in a group-rule-utilitarian way, taking neoclassical consumers' behavior as given. We characterize political equilibrium taxes with a focus on deviations from first-best Pigouvian taxation. Welfare falls further if arguments on moral obligations to reduce carbon footprints polarize the debate in society. Finally, we present survey evidence that supports our theory: social responsibility correlates with lower consumption of brown goods, higher preferred carbon taxes, and support for moral arguments.

Keywords: Political economy of taxation, carbon taxes, ethical behavior, moral dissent
JEL classification: C9; D11; D72; H23.

* **Bierbrauer**: Department of Economics, University of Cologne, Germany (`bierbrauer@wiso.uni-koeln.de`); **Polborn**: Vanderbilt University, Nashville, USA (`mattias.polborn@vanderbilt.edu`); **Ritterrath**: VATT Institute for Economic Research & Department of Economics, University of Cologne, Germany (`marten.ritterrath@gmail.com`); **Weizsäcker**: Humboldt-Universität zu Berlin, Germany (`weizsaecker@hu-berlin.de`). The authors gratefully acknowledge support by the Deutsche Forschungsgemeinschaft (German Research Foundation) via Germany's Excellence Strategy – EXC2126/2 – 390838866 – and CRC TRR 190 (Project number 280092119). We benefitted from the comments of seminar audiences at WZB Berlin, TU Dortmund, University of Vienna and the Central European University and from conversations with Ingela Alger, Hans Peter Grüner, Marc Kaufmann and Sebastian Siegloch. The survey was pre-registered at https://aspredicted.org/ja9hq4.pdf.

1 Introduction

Most people agree that climate change is a pressing challenge for humanity, yet discussions about climate policy are highly acrimonious and make strong use of moral arguments, where each side portrays the other as normatively ill-guided.[1] Part of the disagreement can be captured by traditional political economy analyses of taxation, as voters vary in their economic situations and thus in their desired levels of taxation and regulatory measures. However, such differences arguably cannot justify the debate's heatedness. We therefore enrich the analysis by studying the extent to which economic agents, as consumers and citizens, are motivated by morals. We find that moral concerns interact with material concerns in a way that generates acrimony.

Our theoretical model considers three dimensions of heterogeneity. First, individuals differ in their demand of dirty goods, second, in their valuation of material climate damages, and third, in whether they perceive a moral obligation to decrease their carbon footprint, beyond material self-interest. Accordingly, our stylized economy has two groups of consumers comprised of types that differ among the first two dimensions of heterogeneity listed above, and who each choose a consumption level and vote on the level of a corrective tax: neoclassical consumers (henceforth, *N-agents*) who take prices and overall carbon emissions as given and maximize individual consumer surplus, and socially responsible consumers (*S-agents*) who, by contrast, internalize part of the damages arising from their actions. More precisely, they act as group-rule utilitarians who care about the welfare of all members of their group.

We characterize the political equilibrium taxes in this economy, and discuss how the disalignment of political preferences leads to deviations from the well-known Pigouvian prescription: tax rate equal to marginal damage. A benchmark result is that, with a symmetric distribution of carbon footprints, *N-agents, if they were alone,* would implement an efficient Pigouvian tax. If alone, S-agents would also reach an efficient outcome, with no need for a tax. By contrast, when N- and S-agents are together, the distributive effects of taxes cause a conflict between the groups, to the effect that the political equilibrium involves overtaxation from the perspective of N-agents and undertaxation from the perspective of S-agents. We then augment the model by allowing for moral activism in the form of statements that criticize/defend the N-agents' objective function. The possibility of activism boosts the acrimony even further. An equilibrium with moral activism may be Pareto-inferior to the activism-free one and involve undertaxation not only from the perspective of S-agents, but also from that of most N-agents.

[1] The acrimony is reflected in popular slogans such as *"There is no Planet B"* or *"Drill, Baby, Drill!"* that carry messages about normative objectives. Proponents of a stronger climate policy often make universalist appeals that emphasize the welfare of large sets of people, including those in future generations. Critics of climate policies often appeal to freedom of choice.

The paper's theory part involves various results that are, to the best of our knowledge, novel. We propose a microfoundation for corrective tax preferences, both for neoclassical and for socially-responsible consumers, and derive comparative statics on how equilibrium taxes change when the population share of socially responsible agents increases. Despite the fact that such an increase comes with more internalization of climate damage, we show that it can imply *lower* taxes: S-agents consume less of the dirty good and respond less to taxes, so a shift from N- to S-agents shrinks the tax base and makes consumption less elastic. Both forces lower an N-voter's marginal benefit from raising taxes, so that the decisive N-voter prefers a lower tax. Moreover, this equilibrium effect can be large enough to cause an overall *increase* in emissions. In our analysis of moral activism, we discuss equilibrium multiplicity; there may exist both an equilibrium with moral pressure by socially responsible consumers and high taxes, and an alternative low-tax-equilibrium in which this moral pressure is countered by "rebellious" neoclassical consumers pushing back. In terms of welfare, we find that an equilibrium with moral pressure and moral rebellion can be Pareto-dominated by an equilibrium with no moral activism at all.

Finally, the empirical part of the paper reports on a survey that confirms most of our assumptions and results. Several of our theoretical results rest on the premise that N-agents tend to have larger carbon footprints and perceive smaller social costs of carbon than S-agents. The survey shows that this is indeed the case. Individuals who are more similar to our model's S-agents report having a greener life-style, viewing climate change as a more serious problem, wanting carbon taxes to be larger and being more supportive of moral pressure on consumers who disregard social costs. In the remainder of the introduction, we give a more detailed overview of our analysis, followed by a discussion of related literature.

Preferences over carbon taxes. In our theory, both N- and S-agents care about the climate but face a free-rider problem.[2] For N-agents, we follow an approach that has been used in the literature on the political economy of redistributive taxation and adapt it to a setting with corrective taxes.[3] Specifically, N-agents take damages due to carbon emissions into account and value the mitigative effect of a corrective Pigouvian tax. In addition, there is a distributive effect as the tax revenue is rebated. An N-agent's tax preference balances the mitigation and the redistribution effects with her loss of consumer surplus when taxes go up. The interplay of the Pigouvian and the distributive component typically implies that many N-agents want to deviate from a

[2] We also allow for the willingness to pay for damage mitigation to be zero.

[3] Seminal references are Roberts (1977) and Meltzer and Richard (1981). Bierbrauer et al. (2022) and Berliant and Boyer (2024) discuss more recent contributions to the political economy of taxation.

pure Pigouvian benchmark. For instance, an N-agent with a below-average carbon footprint wants taxes that are higher than this benchmark.

S-agents are ethically motivated. In an environment composed entirely of such agents, they would adhere to a Kantian prescription: *Behave such that an optimal social outcome would result if everybody behaved this way.* However, in an environment that also includes neoclassical consumers, the prescription is modified: *Take as given that N-agents maximize their own preferences. Behave such that an optimal social outcome results for the S-agents if they all behaved this way.* This modified rule nests both individual payoff maximization and the adherence to the Kantian prescription as special cases. The shifter between these extreme cases is the population share of socially responsible consumers. If there is only one such consumer, her behavior is the same as the behavior of N-agents. By contrast, if everybody else also is an S-agent, then the modified rule stipulates Kantian behavior.[4] The modified rule involves an antagonism: There is an in-group of socially responsible agents whose welfare is taken into account, whereas the out-group's welfare is not.[5] Clearly, this group-rule-utilitarian calculus internalizes some of the social costs of carbon: differently from an individual, the whole group of S-agents has a noticeable impact emissions. Therefore, S-agents show a behavioral response not only to changes of (tax-inclusive) consumer prices, but also to changes in the social cost of carbon, while N-agents only react to prices.

The payoff that S-agents reach as a group depends on the corrective tax, and we infer their tax preferences from the value function of their group-rule utilitarian optimization problem. The derivative of the value function gives the marginal benefit of higher taxes. Again, there is a distributive and a Pigouvian component, but they differ from the N-agents' tax preferences. The Pigouvian motive is entirely due to the fact that the N-agents show a behavioral response to taxation. (The S-agents' behavior is, by definition, optimal for the members of their group – it needs no correction.) Importantly, the distributive effect of the tax-and-rebate mechanism leads to an under- or overtaxation relative to the pure Pigouvian benchmark: if the average S-agent has a smaller carbon footprint than the average N-agent, then S-agents prefer high taxes to take advantage of the rebate, beyond the pure Pigouvian correction.

We look at several stylized special cases for the parameters of our general setup. One

[4]Related studies on moral behavior analyze how Kantian prescriptions are traded off against other objectives, including material incentives. Brekke et al. (2003) take a similar approach to ours in their translation from outcome utilities to a Kantian prescription, embedding it into policy regimes (but not political equilibria like ours). Ellingsen and Mohlin (2026) and Alger (2026) examine circumstances under which Kantian motives can dominate other motives.

[5]In part A.8 of the Appendix, we explore universalism as an alternative model, which incorporates also the welfare implications for neoclassical consumers. Our choice to give more emphasis to group-rule-utilitarianism follows our interest in highlighting the antagonism that can arise from other-regarding preferences. Alger (2026) considers related effects, like aversion to behindness, that limit universalist motives.

such special case is that all N-agents have the same consumption preferences and damage assessments and the same is true for all S-agents, so that there is a representative agent for each group,[6] and we further assume that the S-agent has a weakly lower carbon footprint and a weakly higher damage estimate than the N-agent, with one of these inequalities strict. We finally assume that N-agents outnumber S-agents. For this case, we consider a rise in the population share of S-agents (while staying a minority), and find that the politically decisive N-agent indeed prefers lower taxes than before the change, for distributive reasons. While the additional S-agents consume less than the N-agents whom they replace, the tax decrease leads to a higher consumption, especially among continuing N-agents. This is a political economy version of the waterbed effect: the demand for the dirty good shrinks, causing a drop of its tax-inclusive price, so that the demand for the dirty good bounces back up. We show that the bounce back can be so pronounced that country-wide emissions increase.

Moral pressure and moral rebellion. Next, we allow S-agents to exert moral pressure on N-agents to reduce their carbon footprints. The latter then choose whether to comply with this moral pressure or to reject it, to which we refer as moral rebellion. We take the imposition of moral pressure as having a cost, e.g., because one has to assume an unpleasant role in private interactions and remind others of their moral obligations. Yet, there also is a benefit: neoclassical consumers do not want to be frowned at, so they reduce their consumption of the dirty good when they fear critical reactions. S-agents weigh the costs and benefits to determine the intensity of their moral pressure using, again, a group-rule utilitarian calculus.

We show that moral pressure is a substitute for taxes: The lower the tax rate (and the higher the N-agents' consumption), the stronger is the optimal moral pressure. N-agents adjust their tax preferences accordingly, understanding that there is now the additional benefit that higher taxes crowd out moral pressure. If the decisive voter is an N-agent, political equilibrium taxes thus increase and carbon emissions decrease. Thus, moral pressure pays off for S-agents, not only directly, but also politically.

However, we then allow N-agents to engage in moral rebellion and reject the pressure, which requires a costly effort but generates the benefit of unhindered consumer surplus maximization. Here, we assume a strategic complementarity: rejecting moral pressure is less costly when more people rebel. As in other models, complementarity can give rise to equilibrium multiplicity, and we analyze conditions under which an equilibrium in which N-agents simply accept moral pressure exists alongside an equilibrium with widespread moral rebellion. In the latter, moral rebellion is more attractive for high-

[6] Our general setup has a joint distribution of preferences for the dirty good and views of the social cost of carbon among N-agents and a separate distribution among S-agents.

intensity consumers. Hence, the equilibrium partitions N-agents into a set of rebellious heavy consumers and a set of light consumers who can still be reached by moral pressure.

The set of influenceable light consumers expands with higher taxes. Therefore, S-agents now perceive moral pressure and taxes as complementary. They intensify moral pressure when taxes are high, creating an additional cost: higher taxes crowd in moral pressure. If the decisive voter is an N-agent, the political equilibrium taxes may therefore be lower and carbon emissions higher, relative to the baseline with no moral pressure and no moral rebellion. As a result, moral rebellion pays off for N-agents, not only directly, but also politically. Via a numerical example we illustrate these effects and demonstrate that, in comparison to the initial equilibrium without any moral activism, everyone is worse off; it can be bad for all if politics is turned into a moral battle.

Survey. Finally, we present a survey in which we measure climate concerns, the importance of carbon-intensive consumption, socially responsible behavior, and carbon tax preferences. The survey sheds light on how close people are to the stylized categories of N- and S-agents of our theory.[7] These two stylized types are distinct in their consumption behavior, notably their behavioral responses to changes in consumer prices and emission damages. N-agents respond to the former but not to the latter, whereas S-agents respond to both. Moreover, some of our formal propositions focus on an environment in which S-agents want higher taxes than N-agents, due to either their smaller average carbon footprints or their larger damage estimates. Our survey allows us to see whether this is an empirically relevant environment, by examining how the respondents' stated consumption patterns correlate with tax preferences, damage assessments, and attitudes towards moral pressure and moral rebellion.

We run an online survey on Prolific with representative samples of US and UK respondents, stratified by political affiliation. Our main measures of socially responsible behavior are the *Damage Response*, the stated adjustment of car usage to an increase in the social cost of carbon at fixed prices, and a domain-general measure of the self-perceived trade-off between consumption value and social costs, *Price vs. Harm*. We also elicit proxies for the importance of carbon intensive consumption and climate concern, preferences over carbon taxes, and attitudes towards moral pressure.

We document three main findings, each consistent with our theory. First, the behavioral distinction shows up in the data: respondents with a stronger *Damage Response* are more concerned about climate change, place less weight on carbon-intensive consumption, and weigh harm more heavily relative to price. Second, more socially responsible

[7] There is an emerging literature that uses survey methods to explore views on climate change and on policies addressing it, see in particular Dechezleprêtre et al. (2025) or Fabre et al. (2025) and the references therein. To this literature we add a survey that is guided by our political economy model.

behavior (as measured by a larger *Damage Response*) is correlated with a preference for higher carbon taxes, even when controlling for brown consumption and climate concern. Third, more socially responsible respondents are more likely to approve of moral pressure and less likely to view it as intrusive, while respondents for whom carbon-intensive consumption is important show the opposite pattern. Approval of moral pressure, in turn, correlates with higher preferred taxes.

Related Literature. A rich literature on the political economy of redistributive taxes focuses on how changes in income inequality transmit via the political process into changes in redistributive policies. A well-known result is the Meltzer-Richard (1981) hypothesis that an increase in inequality, measured by the gap between average and median income, leads to a larger and more redistributive government.[8] A contribution of this paper is to derive a version of this result for corrective taxes: The gap between the median and the average carbon footprint explains whether there is over- or undertaxation - relative to a pure Pigouvian tax.

We also move beyond this extension in our analysis of how moral pressure, moral rebellion and carbon taxes jointly emerge as endogenous political equilibrium outcomes. The work of Bénabou and Tirole (2006) and Alesina and Angeletos (2005) and this paper have in common that "opinions" – views on the origins of economic success in their case, views on moral obligations to reduce carbon footprints in ours – matter for equilibrium taxes. Our work is related to Bénabou and Tirole (2006) in that there is an explicit attempt of one group to influence the views of another group.[9] We move beyond their approach in that we give those who are being influenced a choice to accept or to reject this influence.

We model socially responsible consumers as group-rule utilitarians.[10] Group-rule utilitarianism has previously been used in political economy models with endogenous turnout, as a way around the paradox of voting, see Coate and Conlin (2004) and Feddersen and Sandroni (2006). To the best of our knowledge, there is no previous work that uses group-rule utilitarianism to formalize moral behavior in markets. Instead, the literature has explored Kantian morality and other-regarding preferences.

[8] The Meltzer-Richard hypothesis often fails empirically, and this has motivated alternative positive theories of the welfare state, e.g. by Bénabou and Tirole (2006) or Alesina and Angeletos (2005).

[9] In Bénabou and Tirole (2006), these are parents who seek to influence their children's beliefs about the returns to educational effort. In our case, socially responsible consumers seek to influence the views of selfish payoff maximizers on the acceptability of large carbon footprints. In doing so, socially responsible consumers distinguish an in-group of like-minded individuals and an out-group of individuals whose behavior warrants a correction. Thereby, our analysis also speaks to the large literature on political polarization (Fiorina and Abrams, 2008; Abramowitz and Webster, 2016; Krasa and Polborn, 2014; Callander and Carbajal, 2022; Gennaioli and Tabellini, 2025).

[10] A related specification is *Homo Moralis* in Alger and Weibull (2013) who is also an intermediate version of a Kantian and a neoclassical agent. The difference to group-rule utilitarianism is that the ethical component of the Homo Moralis preferences is universalist.

The Kantian approach uses a counterfactual as a normative benchmark: morality requires to follow a rule that would yield the socially optimal outcome if all individuals complied. Models of Kantian morality assume that it is costly for individuals to violate this norm, see Alger and Weibull (2013) and Roemer (2015) for foundational work and van Leeuwen and Alger (2024) or Capraro and Rand (2018) for lab experiments with a focus on Kantian morality. Implications of Kantian morality for climate policy are analyzed in Eichner and Pethig (2021).

For the rich literature on experiments about other-regarding preferences, such as altruism or inequity aversion, see Fehr and Schmidt (1999), Bolton and Ockenfels (2000) or Charness and Rabin (2002) for important early references, or Fehr and Charness (2025) for a recent review. In most of these experiments, participants can affect other participants' payoffs, giving other-regarding preferences a chance to play out. By contrast, individuals with other-regarding preferences are indistinguishable from selfish ones in Dufwenberg et al. (2011) where players interact via the price mechanism in large markets. The same is true in our analysis. Individuals who do not belong to the group of socially responsible consumers may still have other-regarding preferences that may shape their tax preferences, but these other-regarding preferences do not show in the behavior as a consumer.[11]

An alternative approach for the analysis of large markets has recently been developed by Kaufmann et al. (2024) where an individual's impact on the equilibrium quantity is non-negligible even when there are many market participants. Therefore, individuals who care about others internalize some of the social damages from carbon emissions even in the absence of corrective taxes. Broadly, in their framework and also in ours, more socially responsible behavior reduces overall emissions, for given taxes. In our framework, taxes are endogenous, however. An increased share of socially responsible consumers may lead to lower taxes and ultimately to higher emissions.

We view group-rule utilitarianism as an approach that is complementary to the literature on Kantian morality and other-regarding preferences. Our ethical agents care for the welfare of other ethical agents and, like in the Kantian literature, their behavior is in line with the solution of a collective action problem. However, our ethical agents consider only the welfare of their own group of ethical agents. Their normative standard thus changes with the composition of the economy. For instance, their response to changes in the social cost of carbon depends on the population share of neoclassical

[11] Models of social preferences have the property that individuals are selfish in the absence of externalities, see Bierbrauer and Netzer (2016): When individuals have a choice between A and B and when A and B have the same consequences for everybody else, they choose the alternative that yields the larger payoff for themselves. When the participant in a market system with many participants takes aggregate outcomes, i.e. prices and quantities, as given, then the individual's choices maximize consumer surplus, irrespective of whether or not the individual has other-regarding preferences.

consumers. This malleability of moral standards is absent from Kantian treatments, but broadly consistent with the evidence on social preferences in lab experiments. In particular, group utilitarianism has a flavor of conditional cooperation (Fischbacher and Gächter, 2010; Thöni and Volk, 2018) in that it sets a standard for moral behavior by referencing an in-group of moral people, while selfish payoff maximizers form the out-group and warrant a costly corrective activity. Consistent with this, many experiments also show a willingness to sacrifice own payoff for the purpose of punishing others (Fehr and Gächter, 2000).[12]

Outlook. The next section introduces our baseline model of an economy with neoclassical and socially responsible agents who consume and vote. Their respective tax preferences are characterized in Section 3, while political equilibrium taxes are analyzed in Section 4. Sections 5 and 6 extend the basic setup. Exerting moral pressure becomes possible in Section 5, rejecting it in Section 6. Section 7 uses examples to show that equilibria with moral activism can be Pareto-dominated by those without moral activism. We report on our survey in Section 8, and the final section contains concluding remarks. Formal proofs are relegated to the Appendix.

2 The Model

There is an atmospheric externality; CO2 emissions from the production of a dirty good contribute to global warming. We consider a country that is inhabited by a continuum of individuals who differ in their demand for the dirty good and, possibly, also in their views on the desirability to reduce the country's aggregate emissions. Moreover, two subgroups of the population differ in how they behave as consumers of the dirty good. *Neoclassical consumers* (henceforth, N-agents) take both prices and aggregate consumption as given when acting as consumers. They equalize their marginal utility from the dirty good with its tax-inclusive price. As we explain in the next section, in their role as voters, N-agents also have preferences over carbon taxes. However, their political preferences are not revealed by their behavior as consumers.

The second group consists of *socially responsible consumers* (henceforth, S-agents). They differ from N-agents because they, motivated by ethical concerns, internalize environmental damages. Below, we formalize their consumption choices as resulting from

[12] We formally treat moral pressure as a behavioral intervention or "nudge" that makes neoclassical consumers behave as if they were facing a higher price, without actually paying it; see Farhi and Gabaix (2020) for an analysis of nudges and behavioral biases in the context of optimal tax models. List et al. (2023) discuss "nudges as moral taxes" and "guilt per unit of consumption induced by a (...) nudge." To be clear, this uses the term "nudge" for every non-price intervention. "Moral pressure" is not an example of a nudge in the sense of Thaler and Sunstein (2009), a mere change of the choice architecture that helps biased agents to make better choices without having a negative effect on others.

a rule-utilitarian calculus. Subsequently, we turn to their political preferences.

We let θ parameterize the marginal utility of consuming the dirty good. More formally, let x be a generic consumption level of the dirty good and denote consumption utility by $u(x,\theta)$. We denote the marginal utility of consumption by $u_x(\cdot)$ and assume that it is increasing in θ, i.e. $u_{x\theta}(\cdot) \geq 0$. For tractability, we sometimes impose the assumption that the function u is iso-elastic; i.e. $u(x,\theta) = \theta^{\frac{1}{\varepsilon}} \frac{1}{1-\frac{1}{\varepsilon}} x^{1-\frac{1}{\varepsilon}}$, for $\varepsilon > 1$. An individual's attitude towards the country's emissions is given by a damage function $\bar{x} \mapsto D(\bar{x},\delta)$, where $\bar{x}$ denotes aggregate or per-capita consumption of the dirty good and δ is a parameter. We let $D_{\bar{x}\delta}(\cdot) > 0$ so that a higher value of δ implies larger perceived marginal damages.[13] We also let $\delta = 0$ imply that $D(\bar{x},\delta) = 0$ for all $\bar{x}$. For tractability, we sometimes assume that the damage function D is linear; i.e. $D(\bar{x},\delta) = \delta\, \bar{x}$. Below, when we have fully specified preferences over corrective taxes, we expand on the interpretation of δ. There is a joint distribution of θ and δ among N-agents, and one among S-agents. Our model allows both for the possibility that these distributions differ or are the same. We will refer to stylized special cases to illustrate the theory's implications. Corrective taxes increase the consumer price of the dirty good, given by $q(t) = k + t$, where k is the marginal cost of production and t is the tax on the dirty good. Tax revenue is rebated lump sum.

2.1 Neoclassical consumers

N-agents benefit from the tax revenue that is redistributed in a lump-sum fashion and they are harmed by the damages that are associated with the country's emissions. However, they take damages and tax revenue as exogenously given and choose x to maximize $u(x,\theta) - qx$. The solution, denoted by $x^*(t,\theta)$, satisfies the first order condition

$$u_x(x^*(t,\theta),\theta) \quad = \quad q(t). \tag{1}$$

Equation (1) has implications for the comparative statics of consumption: N-types respond to a change in taxes, but not to changes of marginal damages δ. This is akin to an assumption of price-taking behavior. A single individual cannot move the aggregate variable $\bar{x}$ and therefore takes $D(\bar{x},\delta)$ as exogenously given when deciding on consumption. We denote indirect utility by $v(q(t),\theta)$. For later reference, note that by Roy's identity $v_t(q(t),\theta) = -x^*(q(t),\theta)$, where the subscript t indicates a partial derivative with respect to the tax rate.[14] The per-capita consumption of N-types is

[13]The extent of the atmospheric externality depends also on the volume of emissions in the rest of the world X. We treat X as fixed in the analysis that follows and, for ease of notation, do not include it as an argument of the damage function.

[14]We assume that preferences are quasi-linear in the consumption of an unspecific numeraire good, with the implication that the marginal utility of income is 1.

henceforth denoted by

$$\bar{x}_N(t) = \mathbf{E}_N[x^*(q(t),\theta)] \ ,$$

where the expectations operator $\mathbf{E}_N$ indicates an aggregation in the set of N-types.

2.2 Socially responsible consumers

A Kantian approach to socially responsible consumption could be captured by the following principle: *Consume according to the rule that generates the socially optimal outcome if followed by every one.* However, our S-agents understand that N-agents will not behave according to this prescription. Therefore, they take the behavior of N-agents as given and focus on the welfare implications of rule-based behavior for ethical agents: *Consume according to the rule that maximizes welfare for S-agents if followed by all S-agents.*[15] This amounts to group rule-utilitarianism, where the rule then takes the form of a function $\theta \mapsto x_S(t,\theta)$ that is chosen to maximize

$$R(t) \ + \ \mathbf{E}_S[u(x_S(t,\theta),\theta) - q \ x_S(t,\theta) - D(\bar{x}(t),\delta)] \ .$$

Here, $R(t)$ is tax revenue and the operator $\mathbf{E}_S$ indicates an aggregation over the set of S-agents. Per-capita tax revenue can also be written as $R(t) = t \ \bar{x}(t)$ where

$$\bar{x}(t) = \phi_S \ \bar{x}_S(t) + \phi_N \ \bar{x}_N(t)$$

is per-capita consumption, $\phi_S = 1 - \phi_N$ is the population share of S-agents, $\bar{x}_S(t)$ is per-capita consumption among S-agents and $\bar{x}_N(t)$ is per-capita consumption among N-agents. The S-agents' problem can therefore also be written as: Choose $\theta \mapsto x_S(t,\theta)$ to maximize

$$t \ \phi_N \ \bar{x}_N(t) \ \ + \ \mathbf{E}_S[u(x_S(t,\theta),\theta)] - \big(k + \phi_N \ t\big) \ \bar{x}_S(t) - \mathbf{E}_S[D(\bar{x}(t),\delta)] \ . \qquad (2)$$

Optimal choices satisfy the first order condition

$$u_x(x_S(t,\theta),\theta) \quad = \quad k \ + \ \phi_N \ t \ + \phi_S \ \mathbf{E}_S[D_x(\bar{x}(t),\delta)] \ . \qquad (3)$$

Thus, S-agents internalize the damages they impose on other S-agents and they are responsive to taxes. When almost everybody is an S-agent ($\phi_S \approx 1$), then S-agents hardly respond to taxes and are very responsive to damages. In contrast, if they are mostly surrounded by N-agents, S-agents behave almost like N-agents. For later reference, note that, with a linear damage function, the consumption-choices of S-agents are

[15] In the Appendix, we also discuss *universalism.* U-agents feel socially responsible for every one, so they also care for the well-being of N-agents.

as if they were N-agents facing a consumer price of

$$q_S(t) \quad = \quad k \, + \, \phi_N \, t \, + \phi_S \, \bar{\delta}_S \quad \text{ where } \quad \bar{\delta}_S := \mathbf{E}_S[\delta] \; .$$

In this case, the consumption of an S-agent with type θ is given by $x^*(q_S(t), \theta)$.

3 Preferences over taxes

3.1 Neoclassical consumers

The policy preference of an N-agent with type (θ, δ) is represented by the utility function

$$U(t, \theta, \delta) \quad = \quad R(t) + v(q(t), \theta) - D(\bar{x}(t), \delta) \; ,$$

which describes her preferences in her role as a citizen who evaluates carbon taxes. Such taxes affect her via three channels: (i) welfare derived as a consumer of the dirty good, as reflected by $v(q(t), \theta)$, (ii) tax revenue, captured by $R(t)$, and (iii) the damages $D(\bar{x}(t), \delta)$ caused by the per-capita consumption of the dirty good $\bar{x}(t)$. The ideal tax rate for the agent solves $U_t(t, \theta, \delta) = 0$, where

$$U_t(t, \theta, \delta) \quad = \quad \bar{x}(t) - x^*(q(t), \theta) \quad + \quad \Big(t - D_{\bar{x}}(\bar{x}(t), \delta)\Big) \; \bar{x}'(t) \; .$$

The first term, $\bar{x}(t) - x^*(q(t), \theta)$, captures the distributive effect: A tax increase mechanically increases tax revenue (which is redistributed lump sum) proportional to $\bar{x}(t)$. This benefit is counterbalanced by a reduction in consumption utility; by Roy's identity, this loss equals $-x^*(q(t), \theta)$. The net effect is positive for individuals with below-average consumption of the dirty good and vice versa. Behavioral responses to a tax increase affect $U_t(t, \theta, \delta)$ via the second term, $\big(t - D_{\bar{x}}(\bar{x}(t), \delta)\big) \; \bar{x}'(t)$. Again, there is a netting of two effects. Consumption decreases, which leads to a revenue loss given by $t \; \bar{x}'(t) < 0$, and a reduction in damages given by $-D_{\bar{x}}(\bar{x}(t), \delta) \; \bar{x}'(t) > 0$, which adds to the benefits of a tax increase. Note that $U_{t\theta}(\cdot) < 0$ and $U_{t\delta}(\cdot) > 0$, so that N-types' ideal taxes are decreasing in θ and increasing in δ.

For N-agents, there is a clear distinction between private choices and collective choices. Even if an N-type demands substantial taxes in the political process, this does not affect her consumption choice. More formally, δ is a shifter for an N-agent's marginal rate of substitution between her own numeraire good consumption and the per-capita consumption of the dirty good in the country; i.e. it is a measure of the willingness to pay for a reduction of the country's overall emissions. If this willingness to pay is high, the N-agent wants a high corrective tax, while consuming as much as another N-agent with a lower δ (who therefore wants a lower corrective tax) but has the same marginal utility of consumption. Observing N-agents private choices therefore does not reveal their political preferences.

An individual who thinks that domestic emissions have no relevant impact on global emissions (e.g. because the own country is small, or because of carbon leakage) and who also has purely instrumental preferences (i.e., only global emissions matter), has $\delta = 0$, but might have a positive willingness to pay for a reduction of *global* emissions. An individual with $\delta > 0$ cares about the country's carbon footprint, for instrumental or other reasons. All these possibilities are consistent with our framework. The survey in Section 8 documents that, for most individuals in the US and the UK, δ is bounded away from zero.[16]

Single-peaked preferences. The following Proposition 1 states sufficient conditions for preferences over taxes to be single-peaked. Single-peakedness implies, in particular, that the ideal tax of an N-agent with type (θ, δ) is well-defined.

Proposition 1 *Let* $u(x,\theta) = \theta^{\frac{1}{\varepsilon}} \frac{1}{1-\frac{1}{\varepsilon}} x^{1-\frac{1}{\varepsilon}}$ *, for* $\varepsilon > 1$*, and let the damage function be linear. N agents' preferences are single-peaked if one the following statements holds:*

1. *There are only neoclassical consumers:* $\phi_N = 1$*.*

2. *The cost parameter* k *exceeds a critical value* $\hat{k}$*, the population share* ϕ_S *does not exceed a critical value* $\hat{\phi}_S$*,* $\theta \in \Theta = [0,1]$*, and* $\bar{\theta}_N > \frac{1}{2}$*, where* $\bar{\theta}_N$ *is the average value of* θ *among* N*-agents.*

Finding sufficient conditions for single-peakedness is more demanding when there is a mix of N- and S-agents. As we show in the proof of Proposition 1, an N-agent has an ideal tax for the subpopulation of N-agents and another ideal tax for the subpopulation of S-agents. The ideal tax for the mixed population lies between the two, with the implication that there is a range of taxes where type (θ, δ) is confronted with a trade-off between overtaxing one group and undertaxing the other one. The conditions in Proposition 1 are needed to ensure that, for any given pair (θ, δ), the function $t \mapsto U(t, \theta, \delta)$ is concave over this range.

[16] Assuming that individuals may differ in their willingness to pay taxes for a reduction of country-wide emissions contrasts with normative approaches that identify a *social* cost of carbon by aggregating world-wide damages and discounting them at a particular rate. Individual heterogeneity may arise for a variety of different reasons, including, but not limited to (i) differences in own exposure to climate damages, (ii) differences in altruism towards the victims of climate change, in particular, in different countries (iii) differences in discount rates or risk aversion, or (iv) differences in beliefs about the severity of carbon leakage.

How do N-agents' political preferences depend on the population share of S-agents? The function $t \mapsto U(t, \theta, \delta)$ depends parametrically on the population share of S-agents, ϕ_S. To make this explicit we expand our notation and write $t \mapsto U(t, \phi_S, \theta, \delta)$.

Proposition 2 *Suppose that damage functions are linear. Consider an N-agent with type (θ, δ). Let $t \leq \delta$ and $t < \bar{\delta}_S$. Suppose that for all such t, $\bar{x}_N(t) \leq \bar{x}_S(q_S(t))$.*[17] *Then there is $\hat{\phi}_S$ so that, for $\phi_S < \hat{\phi}_S$, $U_{t\phi_S}(t, \phi_S, \theta, \delta) \quad < \quad 0$.*

Proposition 2 considers the political preferences of a neoclassical consumer of type (θ, δ). It starts from an initial situation in which taxes are low: They neither exceed the marginal damages δ that are relevant to this individual, nor the damages $\bar{\delta}_S$ that are relevant to socially responsible consumers. It then shows that the marginal benefits of raising taxes are decreasing in the population share of S-agents. This is driven by two forces: S-agents consume less than N-agents and they are less responsive to tax changes. Thus, higher taxes bring less revenue and they are less effective in mitigating damages when more S-agents are around. Proposition 2 can be interpreted as a crowding-out result: A larger share of S-agents lowers the N-agents' ideal taxes.

3.2 Socially responsible consumers

Denote by $W^S(t)$ the value function associated with the group rule-utilitarian consumption problem in (2). S-agents' welfare depends on the tax level t. The derivative $W^S_t(t)$ gives the S-agents' marginal benefit from a tax increase. By the envelope theorem,

$$W^S_t(t) \quad = \quad \phi_N(\bar{x}_N(t) - \bar{x}_S(t)) \ + \phi_N \left(t - \mathbf{E}_S[D_x(\bar{x}(t), \delta)] \right) \bar{x}'_N(t) \ .$$

The first term in this sum captures the S-agents' distributive benefit; S-agents benefit if they consume less than N-agents, and this effect is the more important the larger is ϕ_N. The second term captures the corrective effect of taxation; when taxes rise, N-agents consume less ($\bar{x}'_N(t) < 0$), and if the tax rate falls short of the marginal damages perceived by the S-agents ($t < \mathbf{E}_S[D_x(\bar{x}(t), \delta)]$), then the marginal corrective taxation benefits are positive. Observe that S-agents want to correct only the N-agents' demand for dirty goods, while their own demand needs no correction.

Proposition 3 *Suppose that $\bar{\theta}_N \geq \bar{\theta}_S$. Let $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \ \frac{1}{1-\frac{1}{\varepsilon}} \ x^{1-\frac{1}{\varepsilon}}$, for $\varepsilon > 1$. Suppose that damage functions are linear. Then S-agents have single-peaked preferences. The peak t^*_W equals $\bar{\delta}_S$ if $\bar{\theta}_N = \bar{\theta}_S$ and exceeds $\bar{\delta}_S$ if $\bar{\theta}_N > \bar{\theta}_S$.*

[17] This is ensured if $\bar{\theta}_N \geq \bar{\theta}_S$ and $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \ \frac{1}{1-\frac{1}{\varepsilon}} \ x^{1-\frac{1}{\varepsilon}}$, for $\varepsilon > 1$.

When $\bar{\theta}_N > \bar{\theta}_S$, the S-agents' ideal tax involves overtaxation relative to the Pigouvian benchmark with $t = \bar{\delta}_S$. The S-agents then want to take advantage of the N-agents' larger demand for dirty goods.

How do the political preferences of S-agents depend on the population share of S agents? The function $t \mapsto W_S(t)$ which captures the political preferences of S-agents depends in a parametric way on the population share of S-agents, ϕ_S. Again, to make this explicit we expand our notation and write $t \mapsto W_S(t, \phi_S)$.

Proposition 4 *Let $t \leq \bar{\delta}_S$. Suppose that $\bar{\theta}_N \geq \bar{\theta}_S$ and $u(x,\theta) = \theta^{\frac{1}{\varepsilon}} \, \frac{1}{1-\frac{1}{\varepsilon}} \, x^{1-\frac{1}{\varepsilon}}$, for $\varepsilon > 1$. Then, $W^S_{t\phi_S}(t) < 0$.*

From the perspective of S-agents, the benefits from higher taxes are smaller when their own population share is larger. Both the corrective and the redistributive benefits of higher taxes are due to N-agents. If there are fewer of them, it becomes less attractive to tax. Note that N-agents share this view. By Proposition 2, an N-agent's marginal benefit from taxes also decreases when there are more S-agents.

4 Political equilibrium

We do not adopt a specific model of the political process, but take the shortcut of equating the median voter's preferred tax policy with the political equilibrium outcome.[18] In the following, we characterize political equilibrium taxes for various special cases of our general setup: The purpose is to clarify how political equilibrium taxes are shaped by (i) heterogeneity in the demand for dirty goods, (ii) heterogeneity in the assessment of damages and (iii) the mix between neoclassical and socially responsible consumers.

4.1 Country A: Neoclassical consumers only

Heterogeneity only in the demand for dirty goods. Consider a country that is inhabited only by N-types. Suppose, moreover, that these agents have a common understanding of the relevant damages, i.e., $\delta = \delta_N$ for all N-agents. The ideal tax rate for a type θ agent then solves $U_t(t, \theta, \delta_N) = 0$, where

$$U_t(t,\theta,\delta) \quad = \quad \bar{x}(t) - x^*(q(t),\theta) \;\; + \;\; \Big(t - D_{\bar{x}}(\bar{x}(t),\delta_N)\Big)\, \bar{x}'(t) \; .$$

[18] There are alternative specifications that could be used. When preferences are single-peaked, the median voter's preferred policy is a Condorcet winner, i.e., it wins a majority against any other alternative. One implication is that the median voter's preferred policy is the equilibrium outcome in a Downsian model of two party competition. Another implication is that the median voters' ideal policy is the (only) stable outcome when citizens can propose an alternative to the status quo and initiate a referendum. By Black's (1948) theorem, the majority is rule transitive when preferences are single-peaked. Therefore, a sequence of pairwise votes "converges" to the median voter's ideal policy.

The decisive voter is the person with the median value of θ, denoted by θ_M. If the distribution of consumption levels is symmetric so that median and average consumption coincide ($\bar{x}(t) = x^*(q(t), \theta)$), then $t = D_{\bar{x}}(\bar{x}(t), \delta_N)$, i.e., the median voter's preferred tax rate is the standard "Pigouvian tax."

If the distribution is asymmetric, then distributive effects must be considered in addition to the Pigouvian motive. If the median voter consumes less of the dirty good than the average consumption level, then she/ he benefits from the distributive effects of higher taxes, and the political equilibrium involves overtaxation relative to the Pigouvian benchmark, and vice versa. Formally, if $x^*(q(t), \theta) < \bar{x}(t)$, then $U_t(t, \theta_M, \delta_N) = 0$, implies $t > D_{\bar{x}}(\bar{x}(t), \delta_N)$, and vice versa. This result is related to the well-known Meltzer-Richard hypothesis in the political economy of income taxation where equilibrium taxes are predicted to be increasing the gap between mean and median income. Here, an analogous force is embedded into an analysis of corrective taxation.

Heterogeneity in the demand for dirty goods and in damage assessments. Suppose now that there is a non-degenerate joint distribution of θ and δ among N-types. Since preferred tax rates are decreasing in θ and increasing in δ, those with the lowest θ and the highest δ, prefer the highest tax rate, and those with the highest θ and the lowest δ prefer the lowest tax rate.

For example, let $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \frac{1}{1-\frac{1}{\varepsilon}} x^{1-\frac{1}{\varepsilon}}$, for $\varepsilon > 1$, and let $\overline{\theta}_N := \mathbf{E}_N[\theta]$. Suppose that damage functions are linear. The preferred tax rate of an N-agent of type (θ, δ) is given by

$$t^*(\theta, \delta) \quad = \quad \frac{\frac{1}{\varepsilon}\left(1 - \frac{\theta}{\overline{\theta}_N}\right) k + \delta}{1 - \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\overline{\theta}_N}\right)} \,. \tag{4}$$

The optimal tax is decreasing in θ and increasing in δ. Formula (4) also displays the inverse elasticities logic that is familiar from optimal tax theory: By how much an agent with type $\theta < \bar{\theta}_N$ wants to raise taxes above δ is inversely related to the elasticity ε.

4.2 Country B: Socially responsible consumers only

Consider a country composed only of S-agents. It follows from the discussion in Section 3.2 that the welfare realized by ethical agents does not depend on the tax rate when everyone is an ethical agent.[19] The consumption of S-agents then follows from

$$u_x(x_S(\theta), \theta) \quad = \quad k \; + \mathbf{E}_S[D_x(\bar{x}(t), \delta)] \;,$$

i.e. everyone fully internalizes the social costs of the dirty good.

[19] Formally, for all t, $W_t^S(t) = 0$ if $\phi_S = 1$.

Comparing countries A and B. For ease of comparison, suppose that all agents agree on the damages (i.e., $\delta = \bar{\delta}$ for all N- and S-agents). Also suppose that the distribution of θ is the same among both N- and S-agents. The outcome in country B with only ethical agents is a normative benchmark and may be considered the best outcome one may hope for. Remarkably, if mean and median consumption coincide, Country A with only selfish agents reaches the same outcome via a Pigouvian tax. If mean and median consumption diverge, then Country A will have higher or lower taxes and, correspondingly, a somewhat different allocation.

4.3 Country C: A mix of N- and S-agents

We now turn to country C with a mixed population consisting of N- and S-agents. For tractability, we make the following assumptions: Let the damage function be linear and suppose that there is common group-wide understanding among N- and S-agents, respectively, of the relevant damages, i.e. $\delta = \delta_N$ for all N-agents, and $\delta = \delta_S$ for all S-agents. Furthermore, the distributions of consumption preferences are degenerate: For all N-agents, $\theta = \theta_N$, and for all S-agents, $\theta = \theta_S$. Assume that $\theta_N \geq \theta_S$ and $\delta_N \leq \delta_S$. These assumptions ensure that, when $t \leq \delta_S$, the representative N-agent consumes more of the dirty good than the representative S-agent, $\bar{x}_N(t) \geq \bar{x}_S(t)$, with equality only if $\theta_N = \theta_S$ and $t = \delta_S$. Hence, carbon taxes are redistributing from the representative N-agent to the representative S-agent. We summarize the implications in the following Corollary that we state without proof.

Corollary 1 *Let $0 < \phi_N < 1$, $\theta_N \geq \theta_S$ and $\delta_N \leq \delta_S$.*

1. *The representative N-agent's preferred tax t_N^* is weakly smaller than δ_N.*
2. *The representative S-agent's preferred tax t_S^* is weakly larger than δ_S.*
3. *Their preferred taxes coincide if and only if $\theta_N = \theta_S$ and $\delta_N = \delta_S$.*

Suppose that $\theta_N > \theta_S$ so that the distributive consequences of corrective taxes favor S-agents and harm N-agents. Then, relative to their respective Pigouvian benchmarks, N-agents want undertaxation and S-agents want overtaxation because of the presence of the other group. If the median voter is an N-agent, then there is undertaxation in equilibrium, both relative to the Pigouvian benchmarks of N-agents, $t_N^P = \delta_N$, and of S-agents, $t_S^P = \delta_S$. The corollary's third point shows that a mix of N- and S-agents is not a problem per se. Ethical diversity does not spur political conflict when the representative N- and S-agent have the same consumption benefits and damage assessment.

Increasing the population share of S-agents. Suppose that at least one of the inequalities $\theta_N \geq \theta_S$ and $\delta_N \leq \delta_S$ holds strictly and consider what happens if some N-agents are replaced by S-agents. Clearly, for a given tax, the demand for the dirty good goes down. But the tax is not given; by Proposition 2, the political equilibrium tax goes down (assuming that N-agents remain in the majority). This countervailing effect pushes dirty good consumption upwards. To see this formally, we write the equilibrium tax rate as a function of the population share of S-agents, $t^{eq}(\phi_S)$, and denote its derivative by $t^{eq}_{\phi_S}(\phi_S) < 0$. Equilibrium consumption of the dirty good is then given by

$$\bar{x}^{eq}(\phi_S) \quad := \quad \phi_S \; \bar{x}_S(q_S(t^{eq}(\phi_S), \delta_S)) + (1 - \phi_S) \; \bar{x}_N(q(t^{eq}(\phi_S))) \; ,$$

and its derivative equals

$$\begin{aligned} \bar{x}^{eq}_{\phi_S}(\phi_S) \quad = \quad & -\Big(\bar{x}_N(q(t^{eq}(\phi_S))) - \bar{x}_S(q_S(t^{eq}(\phi_S)), \delta_S) \Big) \\ & + t^{eq}_{\phi_S}(\phi_S) \; \Big(\phi_S \; \bar{x}_{S,q_S}(\cdot) + (1 - \phi_S) \; \bar{x}_{N,q}(\cdot) \Big) \; , \end{aligned}$$

where the first line contains the mechanical effect when N-agents are replaced by S-agents, and the second line contains the effect due to the tax change.[20] The partial derivatives $\bar{x}_{S,q_S}(\cdot)$ and $\bar{x}_{N,q}(\cdot)$ are both negative so that the second line is positive. This decomposition raises the question which of these effects is dominant. Figure 1a illustrates a setting where the second line dominates: Having more socially responsible agents then increases equilibrium emissions.

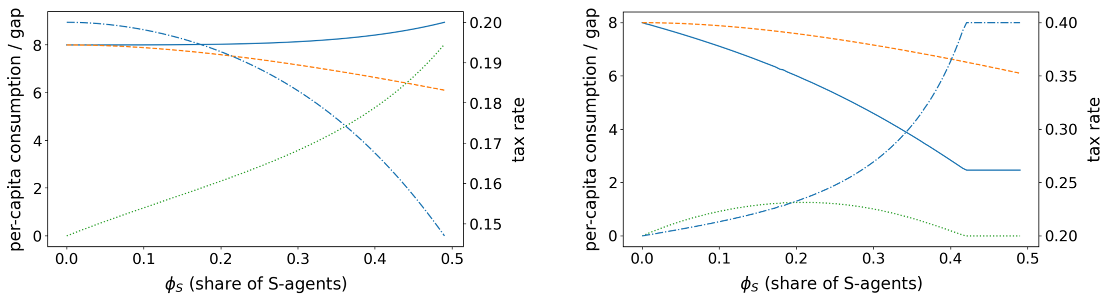
Figure 1a shows a possibility result: For some parameters, when socially responsible agents become more numerous, country-wide emissions go up. There are also other possibilities, as is illustrated by Figure 1b, which is based on a specification where δ is heterogeneous in the N-population, and θ is equally distributed across the two populations, with $\bar{\theta}_N = \bar{\theta}_S$.[21]

5 Moral pressure

As shown in the previous section, the equilibrium emissions can be excessive from the perspective of S-agents. We now extend the model and give S-agents the opportunity to express their discontent with the carbon footprints of N-Agents. We are particularly interested in how such behavior affects tax preferences and, ultimately, the political

[20] Kaufmann et al. (2024) discuss a related, but conceptually different effect that arises with an upward sloping supply function: If some consumers reduce their demand, this lowers prices and leads to demand adjustments by others that partially, but not fully, offset the initial reduction. Here, by contrast, the dirty good is produced with constant returns to scale so that the supply curve is horizontal. Changes of consumer prices are therefore entirely due to tax changes.

[21] Appendix A.9.2 shows what Figure 1b looks like under the alternative assumption that $\bar{\theta}_N > \bar{\theta}_S$.



(a) Emissions increase

(b) Emissions decrease

Figure 1: Emissions and the population share of S-agents

Notes: Equilibrium per-capita consumption of the dirty good as a function of ϕ_S. The blue solid line is equilibrium aggregate consumption $\bar{x}^{eq}(\phi_S)$. The orange dashed line is the counterfactual $\bar{x}(\phi_S)$ holding the tax fixed at $t^{eq}(\phi_S{=}0) = 0.200$, which isolates the composition channel. The green dotted line is the mechanical consumption gap $\bar{x}_N(t^{eq}) - \bar{x}_S(t^{eq})$. The blue dash-dotted line (right axis) is the equilibrium tax $t^{eq}(\phi_S)$. Panel (a): representative-agent specification with $\delta_S = 0.40$, $\delta_N = 0.20$, $\theta_N = \theta_S = 0.5$. Panel (b): $\theta \sim \text{Uniform}[0, 1]$ in both groups, $\delta_N \sim \text{Uniform}[0.1, 0.3]$, $\delta_S \sim \text{Uniform}[0.3, 0.5]$. Common parameters: $\varepsilon = 2$, $k = 0.05$.

equilibrium. Our approach is motivated both by climate protests in the streets, such as the Fridays-for-Future movement, but also by concerns raised in private conversations, e.g. on car use or short-distance flights. In such conversations, those who drive or fly may be frowned at, especially when less carbon-intense alternatives are available.

We will model an attempt of S-agents to establish and uphold a social norm to "worry about the climate and to reduce one's own carbon footprint." Our approach is inspired by a literature that uses group-rule utilitarianism to model turnout in elections. In this literature, it is assumed that voting is costly and that ethical voters who are politically like-minded determine a cutoff level of voting costs. Those with costs below the cutoff are then morally obliged to vote, whereas those with higher costs are excused. The cutoff is a measure of how hard the group is fighting to win the election. If the stakes are high, then the cutoff is also high and almost everybody should turn out. Otherwise, there is more tolerance of "indifference-based abstention;" see Coate and Conlin (2004), Feddersen and Sandroni (2006) or Bierbrauer et al. (2022).

Upholding the social norm is costly; it requires to participate in protests or to take an unpleasant role in private conversations, namely to criticize others for their lifestyle. S-agents are heterogeneous in how costly it is for them to engage in such activities. We assume that S-agents apply a rule-utilitarian calculus not only to their consumption of dirty goods, but also to how confrontational they are in their fight for the social norm. Whether such norm enforcement pays off depends on how N-agents respond. We assume that N-agents reduce their dirty good consumption when they receive moral pressure from S-agents. From the S-agents' perspective, this is the benefit from fighting harder. We now introduce a formalism that captures these considerations.

Denote by σ the fraction of S-agents who participate in upholding the social norm. We assume that S-agents' cost is an *iid* draw from a uniform distribution on the unit interval. Those with a cost below σ call out norm violations, those with a cost above σ feign tolerance. For S-agents as a group, the total cost associated with a participation rate of σ is $C(\sigma) = \int_0^\sigma c\, dc = \frac{1}{2}\,\sigma^2$.

S-agents create a "shaming" function $x \mapsto S(x, \phi_S\,\sigma)$, where x is dirty good consumption and $\phi_S\,\sigma$ is the population share of agents who "sanction" norm violations. We assume that this function is increasing in the first argument and has a positive cross-derivative; that is, the shame associated with, say, an extra flight is increasing when more people show their disapproval. For convenience, we impose a linear functional form in what follows and let $S(x, \phi_S\,\sigma) = x\ \phi_S\,\sigma$. N-agents then choose x to maximize

$$u(x, \theta) - q\,x - S(x, \phi_S\,\sigma) \tag{5}$$

The first order condition characterizing their behavior then is

$$u_x(\cdot) \quad = \quad q(t) + \phi_S\,\sigma\ .$$

The solution of the consumer choice problem is $x^*(q(t) + \phi_S\,\sigma, \theta)$. Thus, N-agents now behave as if they were facing a higher price than the tax-inclusive price $q(t)$.[22] S-agents understand that N-agents respond to moral pressure in this way. They choose σ to minimize their total costs from damages and social norm enforcement. Assuming a linear damage function, this objective can be written as

$$\bar{\delta}_S\ \phi_N\ \bar{x}_N(q(t) + \phi_S\,\sigma) \quad + \quad C(\sigma)\ , \quad \text{where} \quad \bar{\delta}_S := E_S(\delta)\ . \tag{6}$$

It is easy to verify that (6) is decreasing in σ for σ close to 0. Thus, the optimum involves $\sigma > 0$. Moreover, with iso-elastic consumption utility, the objective is globally convex. Thus, either the S-agents go all in and choose $\sigma = 1$, or they choose an interior σ implicitly defined by the rearranged first-order condition

$$\sigma \quad = - \quad \bar{\delta}_S\ \phi_N\ \phi_S\ \bar{x}_{N,q}\big(q(t) + \phi_S\,\sigma\big)\ . \tag{7}$$

Lemma 1 below, which we state without proof, shows that the S-agents' pressure goes down if taxes are higher.[23] Intuitively, if N-agents already reduce their demand because of high taxes, there is less need for moral pressure. Lemma 1 also shows that, while σ

[22] We follow Farhi and Gabaix (2020) in that behavioral interventions or "nudges" – here: the moral pressure exerted by S-agents – make people behave as if they were confronted with a higher tax. A nudge differs from a tax, however, in that it does not generate tax revenue.

[23] The Lemma follows from a straightforward application of the implicit function theorem.

falls when t rises, the tax-and-norm-inclusive-price $q(t) + \phi_S\, \sigma^*_t(t)$ increases in t. Thus, while S-agents become more tolerant when taxes go up, the total effect of a higher tax is still a reduction of the dirty goods demand of N-agents.

Lemma 1 *Let* $u(x, \theta) = \theta^{\frac{1}{\varepsilon}}\ \frac{1}{1-\frac{1}{\varepsilon}}\ x^{1-\frac{1}{\varepsilon}}$ *, for* $\varepsilon > 1$*. Let* $\sigma^*(t)$ *be the value of* σ *that solves (7). Then* $\sigma^*_t(t) < 0$ *and* $1 + \phi_S\, \sigma^*_t(t) > 0$*.*

We now revisit the N-agents' tax preferences and analyze how they are affected by the S-agents' moral pressure.

Proposition 5 *Let* $u(x, \theta) = \theta^{\frac{1}{\varepsilon}}\ \frac{1}{1-\frac{1}{\varepsilon}}\ x^{1-\frac{1}{\varepsilon}}$ *, for* $\varepsilon > 1$*. Consider an* N*-agent of type* (θ, δ) *and an ideal tax of* $t^*(\theta, \delta) \leq \delta$ *when* $\sigma = 0$*. Denote by* $t^*_P(\theta, \delta)$ *this agents' ideal tax with endogenous moral pressure. There is* $\hat{\phi}_S$ *so that* $\phi_S < \hat{\phi}_S$ *implies* $t^*_P(\theta, \delta) > t^*(\theta, \delta)$*.*

Why do N-agents revise their ideal taxes upwards when they get pressure? They now evaluate tax policy using the function

$$U(t, \theta, \delta) \quad = \quad (t - \delta)\ \bar{x}\big(t, \phi_S\, \sigma^*(t)\big) \ + \ \max_x \Big\{ u(x, \theta) - \big(q(t) + \phi_S\, \sigma^*(t)\big)\ x \Big\} \ ,$$

where $\bar{x}\big(t, \phi_S\, \sigma^*(t)\big) \ = \ \phi_N \ \mathbf{E}_N[x^*(q(t) + \phi_S\, \sigma^*(t)), \theta] \ + \ \phi_S \mathbf{E}_N[x^*(q_S(t), \theta]$. By the envelope theorem, the marginal benefit of higher taxes is given by

$$U_t(t, \theta, \delta) \ = \ \bar{x}(\cdot) \ - \ x^*(\theta, \cdot)\Big(1 + \hat{\sigma}^*_t(t)\Big) \ + \ \Big(t - \delta\Big)\ \bar{x}_t(\cdot) \ . \tag{8}$$

Higher taxes decrease consumer surplus. When $\hat{\sigma}^*_t(t) < 0$, the loss from higher taxes is dampened because moral pressure goes down at the same time. The preferred tax rates of N-agents are therefore larger in the presence of moral pressure.[24]

The premise of Proposition 5 is $t^*(\theta, \delta) \leq \delta$ when $\sigma = 0$.[25] As shown in Section 4.3, in a country with representative N- and S-agents, $\theta_N \geq \theta_S$ and $\delta_N \leq \delta_S$, the representative N-agent's preferred tax is bounded from above by δ_N, so that the premise of Proposition 5 is satisfied. If the representative N-agent is, moreover, the decisive median voter, then the political equilibrium tax goes up in response to moral pressure. Importantly, this upward revision does not arise because the median voter now internalizes a social norm to reduce the own carbon footprint. She is not turned into an S-agent, but rather understands that the S-agents' moral pressure is costly for her, and that these costs decrease when taxes are higher.

[24] This force interacts, however, with the fact that $\bar{x}(\cdot)$, $x^*(\cdot)$ and $\bar{x}_t(\cdot)$ depend on $\sigma^*(t)$. Under the premises of the Proposition, and as formally shown in the Appendix, these effects reinforce each other so that an N-agent's preferred tax with moral pressure is larger than without.

[25] Proposition 5 states only a sufficient condition, so preferred taxes may be higher also when its premises do not apply.

6 Moral rebellion

Right-wing populism often rejects "moralizing" in politics and argues that policy-making should be driven more by national self-interest and be less constrained by (leftist) ethical principles. In the following, we present an attempt to introduce such a "rebellion against moralizing" into the model of Section 5. While that model links moral pressure by ethical agents to political equilibrium outcomes, the pressured agents can respond only by adjusting their consumption of the dirty good and by recalibrating their tax preferences.

We now give them the opportunity to reject the demands of ethical agents. Specifically, we assume that, when taking their consumption decisions, N-agents make a binary choice $\gamma \in \{0, 1\}$ whether to accept ($\gamma = 0$) or reject ($\gamma = 1$) the social norm. Rejecting the social norm comes with a (non-monetary) cost, but less so if this is very common among N-agents. Formally, consider an N-agent who expects that (i) σ is the share of S-agents upholding the social norm, (ii) the share of N-agents rejecting the social norm equals $\bar{\gamma}$. Given this expectation, the N-agent chooses $\gamma \in \{0, 1\}$ and the consumption of the dirty good x to maximize

$$u(x, \theta) \; - \; q(t)\, x \; - \; \phi_S\, \sigma\, (1 - \gamma)\, x \; - \; \frac{1}{\bar{\gamma}}\, \gamma \, .$$

Moral rebellion is more attractive for high θ types. An N-agent rejects the social norm if the gain in consumption utility outweighs the cost; this is the case when

$$v(q(t), \theta) - v(q(t) + \phi_S\, \sigma, \theta) \quad \geq \quad \frac{1}{\bar{\gamma}} \, . \tag{9}$$

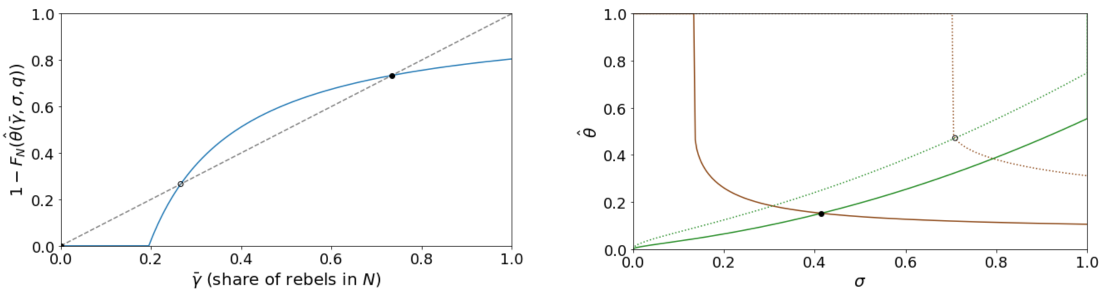
Using Roy's identity, the left-hand side can be written as $\int_0^\sigma x^*(q + \phi_S\, s, \theta) ds$, an expression that is increasing in θ. Therefore, there is a cutoff type $\hat{\theta}(\bar{\gamma}, \sigma, q)$ so that all N-agents with $\theta > \hat{\theta}(\cdot)$ choose $\gamma = 1$ and all N-agents with $\theta < \hat{\theta}(\cdot)$ choose $\gamma = 0$. Monotone comparative statics arguments can be applied to the value $\hat{\theta}$ that satisfies (9) with equality. This yields the following observations (i) $\hat{\theta}_{\bar{\gamma}}(\cdot) < 0$, revealing a complementarity amongst the choices of N-agents: if the share of N-agents rejecting the social norm goes up, the cutoff $\hat{\theta}$ goes down so that rejecting pays off for more N-agents; (ii) $\hat{\theta}_\sigma(\cdot) < 0$, more pressure from S-agents implies that more N-agents find it attractive to reject the social norm; and (iii) $\hat{\theta}_q > 0$; higher taxes reduce the consumption of the dirty good and fewer N-agents reject the social norm.

Equilibrium in the N-agents' coordination game. Given σ, N-agents play a coordination game in which, in equilibrium, all N-agents have correct expectations and their choice of γ is individually optimal, given their expectations about $\bar{\gamma}$. Given σ and

q, this requires that $\bar{\gamma}$ solves the following fixed point equation

$$\bar{\gamma} \quad = \quad 1 - F_N(\hat{\theta}(\bar{\gamma}, \sigma, q)) \; , \tag{10}$$

where F_N is a cumulative distribution function, with density denoted by f_N, that describes the distribution of θ among N-agents. Thus, equation (10) says that $\bar{\gamma}$ is equal to the mass of N-agents with θ above the cutoff $\hat{\theta}(\cdot)$. There may be multiple solutions of (10) in which case there are multiple equilibria in the coordination game among N-agents (see Figure 2a for an illustration). Note first that $\bar{\gamma} = 0$ "solves" the fixed point equation in (10).[26] This implies, in particular, that the political equilibrium that was characterized in Section 5 continues to be a political equilibrium here: There is an equilibrium in which all N-agents accept the moral pressure and adjust their consumption accordingly. To see that there is also scope for an equilibrium with moral rebellion note that, at $\bar{\gamma} = 1$, the left hand side of (10) exceeds the right-hand side as soon as F_N has positive mass to the left of $\hat{\theta}(1, \sigma, q)$.[27] If, moreover the slope of the right-hand side exceeds 1 at $\bar{\gamma} = 0$, there have to be multiple equilibria. Finally, if the right-hand side is a concave function of $\bar{\gamma}$, then there is exactly one further stable equilibrium.[28]



(a) Rebellion in Response to Moral Pressure

(b) How Taxes Affect Moral Pressure and Moral Rebellion in Equilibrium

Figure 2: Equilibrium Analysis of Moral Pressure and Rebellion

Notes: Panel (a) shows the solutions of equation (10), with moral pressure intensity fixed at its equilibrium level σ^{eq}. The blue solid curve plots $1 - F_N(\hat{\theta}(\bar{\gamma}, \sigma^{eq}, q))$; its intersections with the grey dashed 45-degree line are equilibria. The filled marker denotes the stable interior fixed point and the open marker the unstable one; the marker at $\bar{\gamma} = 0$ is the trivial fixed point at which all N-agents accept the moral pressure. Panel (b) illustrates Lemma 3 by depicting the best-response functions of N-agents (brown) and S-agents (green); their intersection is a Nash equilibrium. Solid lines correspond to a low tax rate ($t = 0.02$) and dotted lines to a high tax rate ($t = 0.08$). The filled marker is the Nash equilibrium at the low tax rate, the open marker that at the high tax rate. Parameters: $\varepsilon = 2$, $k = 0.05$, $\varphi_S = 0.20$; $\theta \sim$ Uniform$[0, 1]$ in both groups; $\delta_N \sim$ Uniform$[0.1, 0.3]$, $\delta_S \sim$ Uniform$[0.3, 0.5]$. In panel (a), $t = 0.05$.

[26]More formally, equation (9) defines $\hat{\theta}$ only for $\bar{\gamma} > 0$. Note that $\bar{\gamma} \to 0$ implies that $\hat{\theta} \to \infty$ and hence that $1 - F_N(\hat{\theta}) \to 0$.

[27]In Words: When everybody expects that all N-agents reject the moral claims of S-agents, then there are some N-agents for whom rejecting does not pay off. This contradicts the existence of an equilibrium in which all N-agents reject.

[28]An equilibrium is called stable when the slope of the right-hand side is smaller than 1, with the implication that $1 + f_N(\cdot) \, \hat{\theta}_{\bar{\gamma}}(\cdot) \quad > \quad 0$.

The following Lemma that we state without proof states the comparative statics properties of a stable interior fixed point.[29]

Lemma 2 *Given σ and q, let $\bar{\gamma}^*(\sigma, q)$ be a stable, interior fixed point. Then $\hat{\gamma}^*_\sigma(\sigma, q) > 0$ and $\hat{\gamma}^*_q(\sigma, q) < 0$.*

According to Lemma 2, in a stable equilibrium with moral rebellion by some N-agents, the intensity of rebellion increases in σ, i.e. in the moral pressure of S-agents, and decreases in the tax rate which drives the consumer price q up.

S-agents' best response: Adjusting the moral pressure to the fraction of non-rebellious N-agents. S-agents understand that their moral pressure now only affects the behavior of those N-agents with a type below $\hat{\theta}$, so they minimize

$$\frac{1}{2}\,\sigma^2 \quad + \quad \delta_S\,\phi_N \int_0^{\hat{\theta}} x^*(q + \phi_S\,\sigma, \theta)\; f_N(\theta)\; d\theta\;. \tag{11}$$

Again, assume an interior solution $\sigma^*(q, \hat{\theta})$ that is characterized by a first order condition

$$\sigma^* \quad = \quad -\delta_S\,\phi_N\,\phi_S \int_0^{\hat{\theta}} x^*_q(q + \phi_S\,\sigma^*, \theta)\; f_N(\theta)\; d\theta\;. \tag{12}$$

It is easy to see that σ^* is increasing in $\hat{\theta}$: the more people respond to pressure, the larger is the incentive for S-agents to exert it. As already discussed in Section 5, σ^* is decreasing in q; when consumer prices are high, there is less of a need for a social norm.

Moral pressure and moral rebellion as a Nash-equilibrium. For a given price q, we define a Nash equilibrium as a pair $(\sigma^{eq}, \hat{\theta}^{eq})$ with the following properties: First, given that $\hat{\theta} = \hat{\theta}^{eq}$, σ^{eq} is a best response for S-agents, i.e. σ^{eq} is a minimizer of the costs in (11). Second, given that $\sigma = \sigma^{eq}$, $\hat{\theta}^{eq}$ is an equilibrium outcome in the coordination game amongst N-agents; formally, there is a solution $\bar{\gamma}^*(\sigma^{eq}, q)$ to the fixed-point equation in (10) so that

$$v(q, \hat{\theta}^{eq}) - v(q + \phi_S\,\sigma^{eq}, \hat{\theta}^{eq}) \quad = \quad \frac{1}{\bar{\gamma}^*(\sigma^{eq}, q)}\;.$$

Using the function $(\bar{\gamma}, \sigma, q) \mapsto \hat{\theta}(\bar{\gamma}, \sigma, q)$ introduced above, this condition can also be stated as

$$\hat{\theta}^{eq} \quad = \quad \hat{\theta}(\bar{\gamma}^*(\sigma^{eq}, q), \sigma^{eq}, q)\;.$$

[29] The Lemma follows from a straightforward application of the implicit function theorem, upon using the comparative statics properties of $\hat{\theta}$ and the requirement of stability.

An equilibrium can also be described graphically as the intersection of two functions (see Figure 2b). Here, the "best-response function" of N-agents is $\sigma \mapsto \hat{\theta}(\bar{\gamma}^*(\sigma, q), \sigma, q)$ The arguments above imply that this function is downward sloping since

$$\frac{d\hat{\theta}(\cdot)}{d\sigma} \quad = \quad \hat{\theta}_{\bar{\gamma}}(\cdot)\, \bar{\gamma}^*_\sigma(\cdot) \,+\, \hat{\theta}_\sigma(\cdot) \quad < \quad 0\,.$$

The S-agents' best-response function is given by the function $\hat{\theta} \mapsto \sigma^*(q, \hat{\theta}, q)$ which solves (12) and is upward sloping. A Nash equilibrium is a pair $(\sigma^{eq}, \hat{\theta}^{eq})$ at which these functions intersect.

How does the Nash equilibrium change when taxes go up? When the tax t, or, equivalently, the price q increases, both best response functions shift. Specifically, the N-agents' best response function shifts upwards since

$$\frac{d\hat{\theta}(\cdot)}{dq} \quad = \quad \hat{\theta}_{\bar{\gamma}}(\cdot)\, \bar{\gamma}^*_q(\cdot) \,+\, \hat{\theta}_q(\cdot) \quad > \quad 0\,.$$

S-agents' best-response function shifts downwards: for given $\hat{\theta}$, the optimal σ is lower when taxes are higher. Consequently, the function $t \mapsto \hat{\theta}^{eq}(t)$ is increasing. The sign of $t \mapsto \hat{\sigma}^{eq}(t)$ is not a priori clear. There are two countervailing effects: If $\hat{\theta}$ moves up, more people can be reached and this increases the S-agents' incentives to exert moral pressure. On the other hand, if taxes go up, people consume less and this reduces the need for moral pressure. With an additional assumption, iso-elastic consumption utility, we can determine which effect dominates (see Figure 2b for an illustration).

Lemma 3 *Let* $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \frac{1}{1-\frac{1}{\varepsilon}} x^{1-\frac{1}{\varepsilon}}$ *, for* $\varepsilon > 1$*. Consider a Nash equilibrium* $(\sigma^{eq}, \hat{\theta}^{eq}))$ *with* $0 < \sigma^{eq} < 1$*. Then,* $t \mapsto \hat{\sigma}^{eq}(t)$ *is increasing.*

Implications for tax preferences. The possibility that $t \mapsto \sigma^{eq}(t)$ may be increasing contrasts with the results of Section 5. When there is no rebellion this function is decreasing by Lemma 1. This has implications for tax preferences. With rebellion, N-agents may revise their ideal taxes downwards, rather than upwards.

Proposition 6 *Let* $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \frac{1}{1-\frac{1}{\varepsilon}} x^{1-\frac{1}{\varepsilon}}$ *, for* $\varepsilon > 1$*. Consider an* N*-agent of type* (θ, δ) *and an ideal tax of* $t^*(\theta, \delta) \simeq \delta$ *in the basic model with* $\sigma(t) = 0$*, for all* t*. Denote by* $t^*_P(\theta, \delta)$ *this agent's ideal tax in an equilibrium without rebellion. Suppose that there is an equilibrium with rebellion so that* $\theta < \hat{\theta}^{eq}$*. Denote by* $t^*_R(\theta, \delta)$ *this agent's ideal tax in this equilibrium. There is* $\hat{\phi}_S$ *so that* $\phi_S < \hat{\phi}_S$ *implies*

$$t^*_P(\theta, \delta) \quad > \quad t^*(\theta, \delta) \quad > \quad t^*_R(\theta, \delta)\,.$$

The premise of Proposition 6 applies for an interesting special case of the general setup: When the share of S-agents is small and, among N-agents, $\theta/\bar{\theta}_N$ is sufficiently close to 1 for all θ, (see equation (4)) then the decisive voter is an N-agent with $t^*(\theta, \delta)$ close to δ in the basic model. According to the proposition, this decisive voter wants higher taxes in response to moral pressure, provided that the moral pressure is universally accepted. By contrast, the decisive voter wants lower taxes when the moral pressure is challenged by people with large carbon footprints.

Propositions 6 states sufficient conditions so that the political equilibrium tax, or, equivalently, the decisive median voter's preferred tax is smallest in an equilibrium with moral pressure and moral rebellion and largest in an equilibrium with only moral pressure. The preferred tax in the absence of any moral activism – i.e. in the basic model where individuals can only consume and vote – lies between the two. In the following Section, we complement this insight with numerical examples. In the context of these examples, we can give a complete characterization of tax preferences and equilibrium payoffs. This then enables us to compare equilibria from a welfare perspective and to identify the winners and the losers from moral activism.

7 Tax preferences and equilibrium payoffs: Examples

Figures 3 and 4 show tax preferences and equilibrium payoffs under the following assumptions: First, within the groups of N- and S-agents there is only heterogeneity in consumption preferences and no heterogeneity in the assessment of marginal damages. Specifically, it is assumed that, for all N-agents $\delta = \delta_N = 0.2$ and for all S-agents $\delta = \delta_S = 0.4$. This only serves an expositional purpose, allowing us to use simple two-dimensional graphs to plot preferred taxes and equilibrium payoffs as a function of θ.[30] Second, we assume that the parameter θ is drawn from a uniform distribution over the unit interval, both for N-agents and for S-agents. Finally, we use parameters α and β to scale, respectively, the marginal costs of engaging in moral pressure and moral rebellion to obtain graphs which are easy to interpret.

Figure 3 shows the preferred taxes of N-agents in equilibria with no moral activism (blue), moral pressure (orange) and moral pressure and moral rebellion (green). The blue line shows that, in the baseline model without moral activism, the preferred tax is a decreasing function of θ and, moreover, continuous. The green and the orange line, by contrast, display discontinuities.

When there is no moral rebellion and taxes are very low, S-agents go for maximal

[30] An example with full two-dimensional heterogeneity and three-dimensional graphs that gives rise to the same insights as Figures 3 and 4 can be found in section A.9 of the Appendix, see in particular Figures 5a–5c)

moral pressure and choose $\sigma = 1$. Among N-agents, those with very high demand for dirty goods favor exactly this low-tax range and accept the maximal moral pressure that comes with it. There is a cutoff type who is indifferent between the "smallest" tax in the set of those that come with maximal moral pressure $\sigma = 1$ and the "largest" in the set of those that involve diminished moral pressure $\sigma < 1$ and $\sigma_t < 0$.[31] Moreover, the former of these two taxes is larger than the latter, and this gap between the two taxes at the cutoff explains the discontinuity of the orange curve.[32]

When taxes are very high so that consumption levels are very low, moral rebellion is not attractive, so there is a range of high taxes for which the equilibrium value of $\hat{\theta}$ equals 1. For such taxes, the only equilibrium is one with moral pressure and no rebellion, and in this range the green and orange lines coincide. This range of high taxes, for which the equilibrium value of $\hat{\theta}$ equals 1, ends at some "smallest" cutoff tax rate. Each tax in this range is the ideal policy of some N-agent type θ, that is, the tax she would choose if she could dictate policy. In particular, the cutoff tax is the ideal tax of the cutoff type. For all larger types the ideal tax rate gives rise to rebellion. The transition from the "smallest" tax rate without rebellion to the "largest" tax rate with rebellion gives rise to a discontinuity.[33]

The green line moreover gives rise to "bunching" at the "largest" tax rate with rebellion: there is a set of types who all have their peaks here, namely those who prefer the "largest" tax with rebellion over the "smallest" one without.[34]

The median voter's preferred tax rises from 0.213 in the basic model to 0.334 under moral pressure without rebellion, consistent with Proposition 5, but falls to only 0.193 under moral pressure with rebellion, below even the basic-model level, consistent with Proposition 6. The S-agents' preferred taxes, shown as horizontal dotted lines, are constant in θ: as group rule-utilitarians, all S-agents maximize the same objective $W^S(t)$. When they can exert moral pressure, they prefer a slightly higher tax (0.433) than in the basic model (0.400).

Equilibrium payoffs. Figure 4 shows the corresponding equilibrium payoffs. The baseline equilibrium (with payoffs along the blue line) is based on a setting where people only consume and vote. Figure 4 shows how payoffs change when moral activism enters the scene (with payoffs along the orange line). The baseline equilibrium and the equilibrium with moral pressure (and no rebellion) cannot be Pareto-ranked. In going

[31] A more precise, formal language would invoke an infimum and a supremum.

[32] To see this more formally, note that, in (8), σ_t jumps from 0 to a negative value as one moves from the set of high taxes which all imply $\sigma = 1$ to the set of lower taxes which imply $\sigma > 1$.

[33] Formally, the reason is that in (8), σ_t jumps from a negative value to a positive one as one moves from the set of high taxes without rebellion to the set of lower taxes which provoke rebellion.

[34] The bunching would not occur if the equilibrium value of $\hat{\theta}$ could rise above 1.

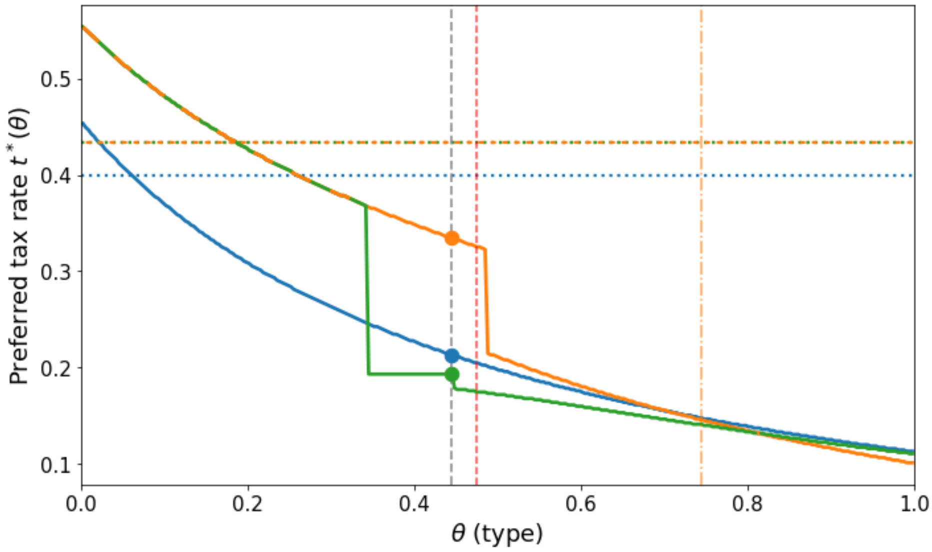


Figure 3: Preferred Tax Rates by Agent Type across Regimes

Notes: Preferred tax rate t^* as a function of θ for N-agents (solid lines) and S-agents (dotted horizontal lines) across three regimes: basic model (blue), moral pressure without rebellion (orange), and moral pressure with rebellion (green). The S-bloc's preferred taxes are $t_S^* = 0.400$ (basic), 0.433 (moral pressure), and 0.433 (rebellion). Filled markers indicate the political-equilibrium tax in each regime, determined by the median voter $\theta_M = 0.444$ (grey dashed line): $t_{eq} = 0.213$ (basic), 0.334 (moral pressure), and 0.193 (rebellion). The red dashed line marks the rebellion cutoff $\hat{\theta} = 0.474$; the orange dashed line marks the activism cutoff $\sigma = 0.745$, with $c \sim \text{Uniform}[0, 1]$ on the same scale. S-agents' preferred taxes are constant in θ because they are group rule-utilitarians who all maximize the same objective $W^S(t)$. Parameters: $\varepsilon = 2$, $k = 0.05$, $\varphi_S = 0.1$, $\delta_S = 0.40$, $\delta_N = 0.20$, $\alpha = 0.5$, $\beta = 0.2$; $\theta \sim \text{Uniform}[0, 1]$ in both groups; activism costs $c \sim \text{Uniform}[0, 1]$.

from the former to the latter, S-agents and N-agents with small carbon footprints win, while all other N-agents lose. Neither is there a Pareto-ranking for the two equilibria with moral pressure: When rebellion emerges (with payoffs along the green line), N-agents with large carbon footprints are better off, while everyone else is worse off. The sequence, however, leads to an outcome that is Pareto-dominated by the initial situation. All moral activists and also all other citizens would be better off if all they could do was to consume and to vote. Broadly, the following logic is at play: S-agents exert moral pressure, this is costly for them. Then, N-agents neutralize this pressure. Again, this is costly. The consequence is a departure from the status quo that is too small to justify the costs of this back and forth.

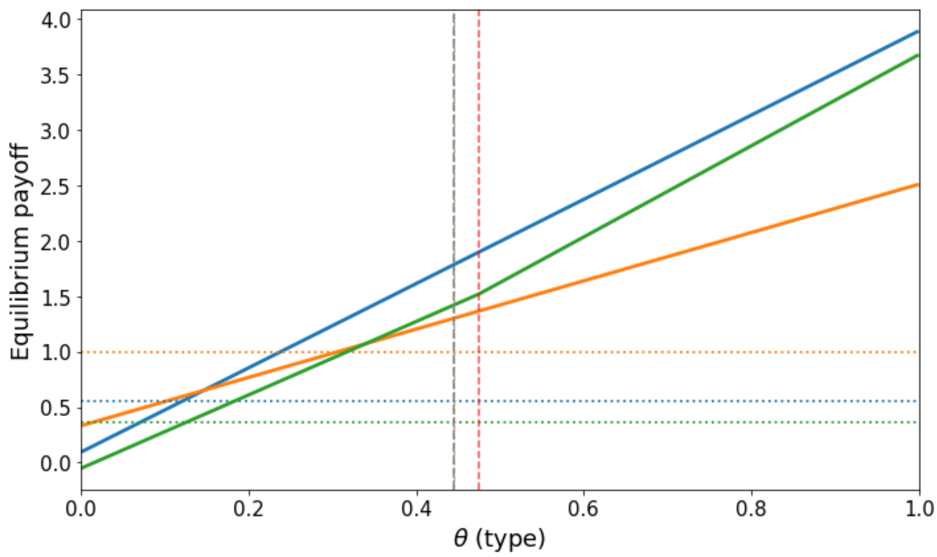


Figure 4: Political-Equilibrium Payoffs

Notes: Moral pressure and rebellion reshape N-agents' equilibrium payoffs relative to the basic model. Solid lines plot the equilibrium payoff $U(\theta)$ for N-agents and dotted horizontal lines plot the average S-agent payoff, across three regimes: basic model (blue), moral pressure only (orange), and moral pressure with rebellion (green). The vertical grey dashed line marks the median voter $\theta_M = 0.444$; the vertical red dashed line marks the rebellion cutoff $\hat{\theta} = 0.474$. Parameters: $\varepsilon = 2$, $k = 0.05$, $\varphi_S = 0.10$, $\delta_S = 0.40$, $\delta_N = 0.20$, $\alpha = 0.5$, $\beta = 0.2$; $\theta \sim \text{Uniform}[0, 1]$ in both groups.

8 Survey

Two samples of respondents, for the US and the UK, were collected through Prolific, using a representative sampling approach with emphasis on political affiliation. The survey instrument collects demographics, stated behavioral responses to price and damage changes, preferences over carbon taxes, and attitudes towards moral pressure and social responsibility. Respondents answer simple comprehension questions on the social cost of carbon and the mechanics of the carbon tax. Those who fail these checks are excluded. We also drop respondents who indicate at the end of the survey that their answers should not be used for research. After these filters, the final samples comprise 901 respondents for the US and 894 for the UK. Our survey instructions are reported in Section A.11 of the Appendix.

When reporting our survey results, we report magnitudes, but focus on the sign of correlations. For comparability, the dependent variables are left in their original units while each independent variable is standardized to have mean zero and standard deviation one. Each coefficient of interest can therefore be read as the change in the dependent variable associated with a one-standard-deviation increase in the corresponding independent variable.

8.1 Primitives

Our primary way to identify socially responsible behavior in the domain of climate change is the *Damage Response*: the self-reported adjustment of car usage in response to new information about the social cost of carbon, holding prices fixed.[35] In Table 1 we check how consumption habits and views on the severity of climate damages correlate with this measure. To make the connection to the formal model in the previous sections salient, when referring to an individual's consumption of brown goods we also use the symbol θ, and when referring to individual valuations of climate damages, we use δ. [36] Our hypothesis is that higher climate concern (δ) is associated with a stronger (more negative) damage response and that more brown consumption (θ) is associated with a weaker response.[37]

Columns (2) and (5) replaces these primitives with the *Price vs. Harm* measure, which directly elicits how respondents weigh price against harm when making consumption decisions.[38] Unlike *Damage Response*, which is specific to car usage and carbon emissions, *Price vs. Harm* is domain-general. It asks respondents how they trade off affordability against the harm that their consumption causes to others, including but not limited to environmental damages. Both variables proxy for socially responsible orientation and we expect *Price vs. Harm* to correlate negatively with *Damage Response.*

While not our primary focus about socially responsible behavior, columns (3) and (6) include the *Oil vs Tax Gap*, the difference in stated car-use reduction between a carbon-tax-induced and an oil-price-induced gasoline price increase.[39] In the model, neoclassical consumers respond only to prices and should not distinguish between the two sources (see equation (1)). A stronger oil price response relative to the tax response is indicative of socially responsible behavior (see equation (3)).

The results in Table 1 confirm these hypotheses in both countries. Climate Concern (δ) is strongly negatively associated with *Damage Response*, while Brown Consumption (θ) is positively associated with it. *Price vs. Harm* enters with the expected negative

[35] *Damage Response* is the respondent's self-reported percentage change in car usage (slider from -100 to $+100$) in response to: "Suppose scientists now estimate that the social cost of carbon from driving gasoline-powered cars has increased by 30%, while prices remain the same. How would you adjust the use of your car?" Respondents were first given a short explanation of the social cost of carbon, covering its definition and the range of damages it captures (see Section A.11 in the Appendix).

[36] Recall that, in the model, individuals with higher θ have a higher demand of the dirty good and individuals with a higher δ have a higher willingness to pay for a reduction of country-wide emissions.

[37] We elicit these measures with the following questions. *Climate Concern* (δ): "How serious of a problem do you think climate change is?" ($1 =$ Not a problem at all, $5 =$ Extremely serious). *Brown Consumption* (θ): "How important are carbon-intensive goods and activities (such as driving, air travel, or meat consumption) to your lifestyle?" ($1 =$ Not at all important, $5 =$ Essential).

[38] *Price vs. Harm*: "When you consider buying something, how do you weigh price/value (0) against whether your consumption might cause harm to others (10)?"

[39] *Oil vs Tax Gap*: difference between stated car-use reduction for a 30% gasoline price increase due to a carbon tax and one due to higher oil extraction costs (both on -100 to $+100$ sliders, so positive numbers mean a larger reduction due to an oil price price increase).

Table 1: Foundations — Correlations with Socially Responsible Behavior

	Dependent variable: Damage Response					
	UK	UK	UK	US	US	US
	(1)	(2)	(3)	(4)	(5)	(6)
Climate Concern (δ)	-5.806^{***}		-5.793^{***}	-8.449^{***}		-8.478^{***}
	(0.777)		(0.784)	(0.876)		(0.883)
Brown Consumption (θ)	2.313^{***}		2.328^{***}	1.557^{*}		1.518^{*}
	(0.811)		(0.809)	(0.876)		(0.879)
Price vs. Harm		-5.330^{***}			-4.066^{***}	
		(1.015)			(0.972)	
Oil vs Tax Gap			-0.934			-2.152^{**}
			(1.305)			(0.892)
Constant	-16.742^{***}	-16.647^{***}	-16.742^{***}	-16.633^{***}	-16.644^{***}	-16.633^{***}
	(0.820)	(0.827)	(0.820)	(0.805)	(0.842)	(0.802)
Mean of DV	-16.740	-16.647	-16.740	-16.633	-16.644	-16.633
Observations	893	890	893	900	900	900
R^2	0.060	0.045	0.062	0.110	0.025	0.117
Adjusted R^2	0.058	0.044	0.058	0.108	0.024	0.114

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored) so coefficients represent the change in the dependent variable per one-standard-deviation increase in the regressor. Mean of DV gives the mean of the dependent variable. The mean can differ across columns due to the different samples, and few respondents not answering all relevant questions. *p<0.1; **p<0.05; ***p<0.01.

sign, and the *Oil vs. Tax Gap* is significant in the US but not in the UK.

8.2 Tax Preferences

We elicit respondents' *Preferred Carbon Tax* by asking, after a brief explainer of a revenue-rebated carbon tax, how high they would like the carbon tax to be in their country, with an additional display of the implied increase in gasoline prices.[40]

According to our theory, an N-agent's preferred tax is decreasing in θ and increasing in δ. Also, much of our theory is done on the premise that S-agents want higher taxes than the average N-agent because their brown consumption is lower and their damage estimates are higher. Thus, *Damage Response*, as a proxy for socially responsible behavior, should be associated with higher preferred taxes. Note that, in a regression of *Damage Response* on preferred taxes this would be reflected in a negative coefficient, since *Damage Response* takes a negative value for S-agents. Table 2 presents such regressions, with five specifications per country: the primitives θ and δ alone; *Damage*

[40] *Preferred Carbon Tax* is the respondent's self-reported tax rate (slider in £/ton in the UK and \$/ton in the US) in response to: How high would you like the carbon tax to be in your country? (At the bottom we show you the estimated increase in gasoline prices.)" Respondents are first reminded that under the proposed carbon tax, everyone pays more for high-emission goods, but all tax revenue is returned equally to the public. Low emitters get back more than they paid."

Response alone; both together; *Price vs. Harm* alone; and θ, δ together with *Price vs. Harm*.

The results support our hypotheses. Brown Consumption θ and Climate Concern δ predict tax preferences with the expected signs. Moreover, the proxies for socially responsible behavior (*Damage Response* and *Price vs. Harm*) are robustly correlated with a higher preferred carbon tax. In the UK, the association between social responsibility and higher carbon tax preferences persists even after controlling for θ and δ (Columns (3) and (5)). This suggests that socially responsible individuals prefer higher taxes above and beyond what is explained by their lower consumption preferences and higher damage assessments alone.

Table 2: Tax Preferences — Level

	Dependent variable: Preferred Carbon Tax									
	UK	UK	UK	UK	UK	US	US	US	US	US
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Climate Concern (δ)	15.832***		14.881***		11.780***	15.788***		15.813***		11.384***
	(1.393)		(1.413)		(1.482)	(1.177)		(1.330)		(1.307)
Brown Consumption (θ)	−4.254***		−3.875***		−3.768***	−0.908		−0.913		−0.717
	(1.450)		(1.451)		(1.407)	(1.263)		(1.258)		(1.215)
Damage Response		−7.844***	−4.139***				−5.117***	0.075		
		(1.637)	(1.572)				(1.847)	(1.797)		
Price vs. Harm				14.654***	10.103***				15.241***	10.369***
				(1.484)	(1.553)				(1.321)	(1.464)
Constant	53.574***	53.589***	53.571***	53.527***	53.498***	40.221***	40.271***	40.221***	40.289***	40.262***
	(1.426)	(1.502)	(1.421)	(1.440)	(1.390)	(1.270)	(1.362)	(1.270)	(1.277)	(1.233)
Mean of DV	53.570	53.589	53.570	53.527	53.507	40.221	40.271	40.221	40.289	40.239
Observations	893	894	893	890	889	900	901	900	900	899
R^2	0.128	0.030	0.136	0.104	0.168	0.146	0.015	0.146	0.137	0.198
Adjusted R^2	0.126	0.029	0.133	0.103	0.165	0.144	0.014	0.144	0.136	0.196

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored). *p<0.1; **p<0.05; ***p<0.01.

Additional results on how respondents adjust their preferred tax in response to hypothetical local and global damage increases are reported in Appendix Tables 6 and 7. The patterns there are broadly consistent with the findings in Table 2: δ is a strong positive predictor and *Price vs. Harm* is significant in all specifications, while θ is weakly negative.[41]

8.3 Moral Pressure

We measure attitudes toward moral pressure with two items capturing, respectively, normative endorsement and normative rejection of confronting others with their brown consumption (*MP Approval* and *MP Opposition*) together with a behavioral measure of

[41] A noteworthy difference is that, in the US, *Damage Response* is positively signed (though insignificant) conditional on θ and δ in the tax-adjustment regressions, whereas it is essentially zero in the level regression.

an individual's own susceptibility to such pressure, *MP Responsiveness*.[42]

In the model, S-agents exert moral pressure on N-agents to reduce their carbon footprints. Our hypothesis is that individuals with higher δ should approve of moral pressure more, and oppose it less, as reducing others' consumption is more valuable for those highly concerned about the externality. Higher θ, in turn, should be associated with less approval and more opposition, since moral pressure raises the effective price of consumption for high-θ individuals (equation (5)). Likewise, we hypothesize that rebellion increases in θ (see section 6). Our proxies for socially responsible behavior (a more negative *Damage Response*, and larger *Price vs. Harm*) should correlate with greater approval and less opposition to moral pressure.

Table 3: Support and Opposition to Moral Pressure — UK

	Dependent variable:							
	MP Approval				MP Opposition			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Climate Concern (δ)			0.519***	0.434***			−0.486***	−0.423***
			(0.031)	(0.033)			(0.035)	(0.036)
Brown Consumption (θ)			−0.082**	−0.084***			0.136***	0.141***
			(0.034)	(0.032)			(0.035)	(0.034)
Damage Response	−0.210***		−0.085***		0.210***		0.089**	
	(0.037)		(0.032)		(0.037)		(0.035)	
Price vs. Harm		0.446***		0.279***		−0.381***		−0.217***
		(0.033)		(0.032)		(0.036)		(0.036)
Constant	3.203***	3.207***	3.203***	3.206***	2.936***	2.934***	2.937***	2.936***
	(0.036)	(0.033)	(0.032)	(0.030)	(0.037)	(0.036)	(0.034)	(0.033)
Mean of DV	3.203	3.205	3.202	3.204	2.936	2.935	2.937	2.936
Observations	892	888	891	887	893	889	892	888
R^2	0.037	0.167	0.256	0.308	0.035	0.113	0.220	0.246
Adjusted R^2	0.036	0.166	0.253	0.306	0.033	0.112	0.217	0.243

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored). *p<0.1; **p<0.05; ***p<0.01.

Table 3 presents results for the UK.[43] Our proxies for socially responsible behavior correlate with greater approval and lower opposition to moral pressure, and climate concern δ and brown consumption θ enter with the expected signs. Consistent with our theory, individuals with a higher damage assessment approve of moral pressure more strongly, while those who place greater value on brown consumption tend to oppose it.

[42] *MP Approval* is the respondent's agreement (1,=,Strongly disagree, 5,=,Strongly agree) with: "Applying moral pressure by confronting individuals with the carbon emissions of their consumption choices is justified." *MP Opposition* is the respondent's agreement on the same five-point scale with: "Moral pressure about individuals' carbon-intensive consumption is intrusive and should be avoided." *MP Responsiveness* is the respondent's answer (1,=,Never, 5,=,Very often) to: "How often do you change what you buy or consume to avoid feeling moral pressure about carbon emissions?"

[43] Since *MP Approval* and *MP Opposition* are measured on five-point Likert scales, we report OLS estimates in the main text for interpretability and comparability with our other regressions, and show in Appendix A.10.2 that all results are robust to estimating ordered probit models instead.

Corresponding US results are in Table 4. The US findings largely corroborate those for the UK. Two differences are worth noting. First, *Damage Response* loses significance for MP Approval in the US once θ and δ are controlled for (column 3), whereas it remains significant in the UK. Second, θ is not significant for MP Opposition in the US, whereas it is in the UK.

Table 4: Support and Opposition to Moral Pressure — US

	Dependent variable:							
	MP Approval				MP Opposition			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Damage Response	-0.174^{***}		0.033		0.292^{***}		0.105^{**}	
	(0.045)		(0.038)		(0.044)		(0.045)	
Price vs. Harm		0.454^{***}		0.236^{***}		-0.386^{***}		-0.158^{***}
		(0.038)		(0.041)		(0.039)		(0.040)
Brown Consumption (θ)			-0.077^{**}	-0.072^{**}			0.052	0.057
			(0.034)	(0.032)			(0.036)	(0.036)
Climate Concern (δ)			0.624^{***}	0.512^{***}			-0.565^{***}	-0.532^{***}
			(0.033)	(0.038)			(0.037)	(0.038)
Constant	3.007^{***}	3.006^{***}	3.006^{***}	3.005^{***}	3.111^{***}	3.112^{***}	3.113^{***}	3.113^{***}
	(0.038)	(0.035)	(0.033)	(0.032)	(0.039)	(0.038)	(0.034)	(0.034)
Mean of DV	3.007	3.006	3.006	3.004	3.111	3.112	3.114	3.115
Observations	901	900	900	899	899	898	898	897
R^2	0.023	0.154	0.281	0.314	0.060	0.105	0.259	0.267
Adjusted R^2	0.021	0.153	0.279	0.312	0.059	0.104	0.257	0.264

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored). All variables are defined as in Tables 1 and 3. *p<0.1; **p<0.05; ***p<0.01.

8.3.1 Moral Pressure and Tax Preferences

Our theory hypothesizes that individuals who approve of moral pressure should prefer higher carbon taxes, while those who oppose it should prefer lower taxes. Additionally, the model's rebellion mechanism implies that individuals who oppose moral pressure but nonetheless respond to it behaviorally (i.e., do not rebel) should prefer higher taxes than those who oppose and also resist it. Table 5 presents results on test of these hypotheses. Columns (1) and (4) include MP Approval and MP Opposition alongside θ and δ. Columns (2)–(3) and (5)–(6) introduce MP Responsiveness and its interaction with MP Opposition to test the rebellion channel.

The results are consistent with the first hypothesis in both countries. MP Approval is positively and significantly associated with preferred tax levels, and MP Opposition

Table 5: Moral Pressure Attitudes and Tax Preferences

	Dependent variable: Preferred Carbon Tax					
	UK	UK	UK	US	US	US
	(1)	(2)	(3)	(4)	(5)	(6)
Climate Concern (δ)	9.485***		9.647***	9.178***		7.113***
	(1.671)		(1.646)	(1.401)		(1.487)
Brown Consumption (θ)	−3.069**		−2.915**	0.012		0.019
	(1.424)		(1.448)	(1.229)		(1.202)
MP Approval	7.442***			10.911***		
	(2.378)			(2.078)		
MP Opposition	−5.606**	−13.198***	−9.639***	−1.572	−9.030***	−6.572***
	(2.319)	(1.690)	(1.831)	(2.076)	(1.461)	(1.527)
MP Responsiveness		7.565***	5.179***		11.866***	9.475***
		(1.672)	(1.689)		(1.542)	(1.598)
MP Opposition × MP Responsiveness		0.232	−0.843		4.266***	3.238**
		(1.519)	(1.518)		(1.295)	(1.301)
Constant	53.442***	53.681***	53.259***	40.090***	42.044***	41.559***
	(1.379)	(1.540)	(1.516)	(1.215)	(1.442)	(1.420)
Mean of DV	53.410	53.594	53.574	40.061	40.111	40.061
Observations	891	893	892	898	899	898
R^2	0.181	0.148	0.183	0.211	0.208	0.227
Adjusted R^2	0.177	0.145	0.178	0.207	0.206	0.222

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored); the interaction *MP Opposition* × *MP Responsiveness* is the product of the two standardized variables. The interaction *MP Opposition* × *MP Responsiveness* tests whether individuals who oppose moral pressure but are nonetheless behaviorally responsive to it prefer different tax levels. *p<0.1; **p<0.05; ***p<0.01.

enters negatively. Climate concern δ remains a strong positive correlate, and brown consumption θ is significantly negative in the UK but not in the US, consistent with the patterns in Table 2.

The rebellion results differ across countries. In the US, the interaction between MP Opposition and MP Responsiveness is positive and significant (columns 5–6), indicating that among those who oppose moral pressure, individuals who nonetheless adjust their behavior are associated with higher tax preferences, consistent with non-rebellious N-types supporting larger corrective taxes. In the UK, by contrast, the interaction is not significant, suggesting that the rebellion channel is less present in this sample.

9 Concluding Remarks

This paper analyzes preferences over carbon taxes und political equilibrium outcomes in an economy consisting of neoclassical consumers and socially responsible consumers. In a hypothetical country A inhabited only by neoclassical consumers who consume a dirty good, the resulting atmospheric externality is taken into account through a political process that determines the corrective tax. The political equilibrium tax is the median voter's preferred tax and reflects his views on damage severity, as well as a distributive consideration: it is increasing in the gap between the average and the median carbon footprint. A first-best Pigouvian tax so that the *"tax rate equals the marginal damage"* is reached in country A only if the average and the median footprint coincide. In a hypothetical country B that is inhabited only by socially responsible consumers, individuals internalize the damages due to their consumption of a dirty good so that there is no need for a corrective tax.[44]

Our analysis mainly deals with country C, a country with a mixed population consisting of neoclassical and socially responsible consumers – with focus on settings where socially responsible consumers have a lower willingness to pay for dirty goods and larger estimates of the social costs of carbon. We present extensive survey evidence that this is the empirically plausible case. Our main theoretical results characterize the tensions in country C.[45] In particular, we show that an equilibrium with moral pressure by socially responsible consumers and moral rebellion by a subset of neoclassical consumers, may give rise to a Pareto-dominated equilibrium in which carbon taxes are very low.

Our survey also shows that the distinction of the polar cases *neoclassical* and *socially responsible* is useful in that it captures attitudes towards both policy and the consumption habits of others. Individuals who are more socially responsible prefer higher corrective taxes and are more supportive of moral pressure on consumers who disregard social costs.

While this paper focuses on carbon taxes, we believe that a similar approach could be fruitful for an analysis of political economy forces also in other domains. At a more abstract level, agents in our economy take individual consumption decisions, but also have preferences over economy-wide consumption patterns. For example, individuals may wish that there are fewer imports from countries with no effective ban of child labor, or that more local products are bought. A stereotypical neoclassical consumer articulates such preferences only in her role as a citizen in the political process, while the group of socially responsible consumers, who would need only very little policy to achieve their preferred outcome if they were alone, find themselves in moral debates with neoclassical consumers.

[44] It is easy to specify stylized environments that result in the same allocation in countries A and B. For instance, if the distribution of carbon footprints is symmetric in country A, and if all individuals have the same views on what the social costs of carbon are, then country A reaches a first-best outcome via a Pigouvian tax, and country B reaches the same outcome because its inhabitants internalize the social costs of carbon even in the absence of a corrective tax.

[45] There are alternative specifications of the models' primitives, so that the same allocation results in countries A, B and C. Thus, having a mix of neoclassical and socially responsible consumers is not a sufficient condition of tensions.

References

Abramowitz, Alan I and Steven Webster, "The rise of negative partisanship and the nationalization of US elections in the 21st century," *Electoral Studies*, 2016, *41*, 12–22.

Alesina, A. and G.-M. Angeletos, "Fairness and Redistribution," *American Economic Review*, 2005, *95* (4), 960–980.

Alger, Ingela, "Norms and norm change - driven by social Kantian preferences," *Toulouse School of Economics Working Papers*, 2026, *1605.*

_ **and Jörgen W. Weibull**, "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching," *Econometrica*, 2013, *81* (6), 2269–2302.

Bénabou, R. and J. Tirole, "Belief in a just world and redistributive politics," *Quarterly Journal of Economics*, 2006, *121* (2), 699–746.

Berliant, Marcus and Pierre C. Boyer, "Politics and income taxes: Progress and progressivity," *Journal of Public Economic Theory*, 2024, *26* (4), e12704.

Bierbrauer, Felix and Nick Netzer, "Mechanism design and intentions," *Journal of Economic Theory*, May 2016, *163*, 557–603.

Bierbrauer, F.J., A. Tsyvinski, and N. Werquin, "Taxes and Turnout: When the Decisive Voter Stays at Home," *American Economic Review*, February 2022, *112* (2), 689–719.

Black, D., "On the Rationale of Group Decision-making," *Journal of Political Economy*, 1948, *56* (1), 23–34.

Bolton, Gary E and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity, and Competition," *The American Economic Review*, 2000, *90* (1), 166–193.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg, "A n economic model of moral motivation," *Journal of Public Economics*, 2003, *87*, 1967–1983.

Callander, Steven and Juan Carlos Carbajal, "Cause and effect in political polarization: A dynamic analysis," *Journal of Political Economy*, 2022, *130* (4), 825–880.

Capraro, Valerio and David G. Rand, "Do the Right Thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality," *Judgement and Decision Making*, 2018, *13* (1), 99–111.

Charness, Gary and Matthew Rabin, "Understanding Social Preferences with Simple Tests," *The Quarterly Journal of Economics*, 2002, *117* (3), 817–869.

Coate, Stephen and Michael Conlin, "A Group Rule-Utilitarian Approach to Voter Turnout: Theory and Evidence," *The American Economic Review*, 2004, *94* (5), 1476–1504.

Dechezleprêtre, Antoine, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva, "Fighting Climate Change: International Attitudes toward Climate Policies," *American Economic Review*, 2025, *115* (4), 1258–1300.

Dufwenberg, Martin, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel, "Other-Regarding Preferences in General Equilibrium," *The Review of Economic Studies*, 02 2011, *78* (2), 613–639.

Eichner, Thomas and Rüdiger Pethig, "Climate Policy and Moral Consumers," *The Scandinavian Journal of Economics*, 2021, *123* (4), 1190–1226.

Ellingsen, Tore and Erik Mohlin, "A Model of Social Duties," *Journal of Political Economy*, 2026, *134* (5), 1307–1346.

Fabre, Adrien, Thomas Douenne, and Linus Mattauch, "Majority Support for Global Redistributive and Climate Policies," *Nature Human Behaviour*, 2025, *9* (8), 1583–1594.

Farhi, Emmanuel and Xavier Gabaix, "Optimal Taxation with Behavioral Agents," *American Economic Review*, January 2020, *110* (1), 298–336.

Feddersen, Timothy and Alvaro Sandroni, "A Theory of Participation in Elections," *The American Economic Review*, 2006, *96* (4), 1271–1282.

Fehr, Ernst and Gary Charness, "Social Preferences: Fundamental Characteristics and Economic Consequences," *Journal of Economic Literature*, 2025, *63* (X), X–X.

_ **and Klaus M. Schmidt**, "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **and Simon Gächter**, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Fiorina, Morris P and Samuel J Abrams, "Political polarization in the American public," *Annu. Rev. Polit. Sci.*, 2008, *11* (1), 563–588.

Fischbacher, Urs and Simon Gächter, "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments," *American Economic Review*, 2010, *100* (1), 541–556.

Gennaioli, Nicola and Guido Tabellini, "Presidential Address: Identity Politics," *Econometrica*, 2025, *93* (6), 1937–1967.

Kaufmann, Marc, Peter Andre, and Botond Koszegi, "Understanding Markets with Socially Responsible Consumers*," *The Quarterly Journal of Economics*, 03 2024, *139* (3), 1989–2035.

Krasa, Stefan and Mattias K Polborn, "Social Ideology and Taxes in a Differentiated Candidates Framework," *American Economic Review*, 2014, *104* (1), 308–322.

List, John A., Matthias Rodemeier, Satunka Roy, and Gregory Sun, "Judging Nudging: Understanding the welfar effects of nudges versus taxes," *Unpublished Manuscript*, 2023.

Meltzer, A. and S. Richard, "A Rational Theory of the Size of Government," *Journal of Political Economy*, 1981, *89*, 914–927.

Roberts, Kevin W. S., "Voting over Income Tax Schedules," *Journal of Public Economics*, 1977, *8* (3), 329–340.

Roemer, John E., "Kantian optimization: A microfoundation for cooperation," *Journal of Public Economics*, 2015, *127*, 45–57. The Nordic Model.

Thaler, Richard H and Cass R Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New York, NY: Penguin, 2009.

Thöni, Christian and Stefan Volk, "Conditional Cooperation: Review and Refinement," *Economics Letters*, 2018, *171*, 37–40.

van Leeuwen, Boris and Ingela Alger, "Estimating Social Preferences and Kantian Morality in Strategic Interactions," *Journal of Political Economy Microeconomics*, 2024, *2* (4), 665–706.

A Appendix

A.1 Proof of Proposition 1

For later use, we note that the functional form assumption $u(x,\theta) = \theta^{\frac{1}{\varepsilon}} \, \frac{1}{1-\frac{1}{\varepsilon}} \, x^{1-\frac{1}{\varepsilon}}$, implies $x^*(q,\theta) = \theta \, q^{-\varepsilon}$, $x_q^*(q,\theta) = -\varepsilon \, \theta \, q^{-(1+\varepsilon)}$, and $v(q,\theta) \quad = \quad \theta \, \frac{1}{1-\frac{1}{\varepsilon}} \, q^{-(\varepsilon-1)}$.

A.1.1 Statement 1: Only neoclassical consumers

Suppose that $\phi_N = 1$. Also assume that damage functions are linear. Policies are then evaluated according to

$$U(t, \theta, \delta) \quad = \quad R(t) + v(k + t, \theta) - \delta\, \bar{x}(t)$$

or, equivalently,

$$U(t, \theta, \delta) \quad = \quad v(k + t, \theta) + (t - \delta)\, \bar{x}(t) \ ,$$

where $\bar{x}(t) = \bar{\theta}_N\, q^{-\varepsilon}$ and $\bar{\theta}_N$ is the average value of θ among N-agents. We have that

$$U_t(t, \theta, \delta) \quad = \quad q^{-(1+\varepsilon)}\, p(t, \theta, \delta)$$

where

$$p(t, \theta, \delta) \quad = \quad \Big\{(\bar{\theta}_N - \theta)q - \varepsilon(t - \delta)\bar{\theta}_N\Big\} \ .$$

Thus $U_t(\cdot) > 0$ if $p(\cdot) > 0$, $U_t(\cdot) = 0$ if $p(\cdot) = 0$ and $U_t(\cdot) < 0$ if $p(\cdot) < 0$. Using that $q = (k + t)$, we find that, for given θ and δ, p is an affine-linear function of t. It is now straightforward to verify the following statements:

$$p(\cdot) \ > \ 0 \quad \Longleftrightarrow \quad \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\bar{\theta}_N}\right) \ > \ \frac{t - \delta}{k + t} \ ,$$

$$p(\cdot) \ = \ 0 \quad \Longleftrightarrow \quad \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\bar{\theta}_N}\right) \ = \ \frac{t - \delta}{k + t} \ ,$$

$$p(\cdot) \ < \ 0 \quad \Longleftrightarrow \quad \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\bar{\theta}_N}\right) \ < \ \frac{t - \delta}{k + t} \ ,$$

and, finally, that $\frac{t-\delta}{k+t}$ is an increasing function of t. All this implies that preferences are single peaked. The ideal tax of an N agent with type (θ, δ) is given by

$$t^*(\theta, \delta) \quad = \quad \frac{\frac{1}{\varepsilon}\left(1 - \frac{\theta}{\bar{\theta}_N}\right)k + \delta}{1 - \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\bar{\theta}_N}\right)} \ .$$

A.1.2 Statement 2: Mixing N and S agents

An N-agent evaluates taxes t, again, according to

$$U(t, \theta, \delta) \quad = \quad v(k + t, \theta) + (t - \delta)\, \bar{x}(t) \ ,$$

but now

$$\bar{x}(t) \quad = \quad \phi_N\, \bar{\theta}_N\, q^{-\varepsilon} \ + \ \phi_S\, \bar{\theta}_S\, q_S^{-\varepsilon} \ ,$$

where $q = k + t$, as before, and $q_S = k + \phi_N \ t + \phi_S \ \bar{\delta}_S$, where $\bar{\delta}_S$ is the average of δ among S-agents. Straightforward computations yield

$$U_t(t, \theta, \delta) \quad = \quad q^{-(1+\varepsilon)} \, p_N(t, \theta, \delta) \ + \ q_S^{-(1+\varepsilon)} \, \phi_S \, \bar{\theta}_S \, p_S(t, \delta) \ ,$$

where

$$p_N(t, \theta, \delta) \quad = \quad (\phi_N \ \bar{\theta}_N - \theta)q - (t - \delta) \ \varepsilon \ \phi_N \ \bar{\theta}_N$$

and

$$p_S(t, \delta) \quad = q_S - (t - \delta) \ \varepsilon \ \phi_N \ .$$

Thus, a type (θ, δ)-agent evaluates marginal changes of t according to a weighted average of two evaluation functions with weights that, however, depend on the tax rate t. To make this explicit in terms of notation, we let $w_N(t) = q^{-(1+\varepsilon)}$, $w_S(t) = q_S^{-(1+\varepsilon)} \phi_S \bar{\theta}_S$ and write

$$U_t(t, \theta, \delta) \quad = \quad w_N(t) \, p_N(t, \theta, \delta) \ + \ w_S(t) \, p_S(t, \delta) \ .$$

Both p_N and p_S are affine functions of t. Let

$$t_N^*(\theta, \delta) \quad = \quad \frac{\frac{1}{\varepsilon}\left(1 - \frac{\theta}{\phi_N \ \bar{\theta}_N}\right) k + \delta}{1 - \frac{1}{\varepsilon}\left(1 - \frac{\theta}{\phi_N \ \bar{\theta}_N}\right)} \ .$$

It is easy to verify that $t \mapsto p_N(t, \theta, \delta)$ is a decreasing function of t, with $p_N(t, \theta, \delta) > 0$ for $t < t_N^*(\theta, \delta)$, $p_N(t, \theta, \delta) = 0$ for $t = t_N^*(\theta, \delta)$ and $p_N(t, \theta, \delta) < 0$ for $t > t_N^*(\theta, \delta)$. Analogously, let

$$t_S^*(\delta) \quad = \quad \frac{k + \phi_S \ \bar{\delta}_S + \varepsilon \ \phi_N \ \delta}{(\varepsilon - 1)\phi_N} \ .$$

It is easy to verify that $t \mapsto p_S(t, \delta)$ is a decreasing function of t, with $p_S(t, \delta) > 0$ for $t < t_S^*(\delta)$, $p_S(t, \theta, \delta) = 0$ for $t = t_S^*(\delta)$ and $p_S(t, \delta) < 0$ for $t > t_S^*(\delta)$. Thus, if we had $w_N = 1$ and $w_S = 0$, irrespective of t, then preferences would be single-peaked with a peak at $t_N^*(\theta, \delta)$. If we had $w_N = 0$ and $w_S = 1$, irrespective of t, then preferences would be single-peaked with a peak at $t_S^*(\delta)$.

All this implies that $U_t(\cdot) > 0$ when $t < t_N^*(\theta, \delta)$ and $t < t_S^*(\delta)$. Also, $U_t(\cdot) < 0$ when $t > t_N^*(\theta, \delta)$ and $t > t_S^*(\delta)$. Since U_t is continuous in t this implies that there exists some t in the intermediate range between $t_N^*(\theta, \delta)$ and $t_S^*(\delta)$ so that $U_t(t, \theta, \delta) = 0$. To establish single-peakedness we need to show that this is the only value of t so that $U_t(t, \theta, \delta) = 0$. Since U_t is itself a continuously differentiable function of t, any non-monotonicity of U over the intermediate range would be associated with another local extremum or peak, i.e. with another value of t so that $U_t(t, \theta, \delta) = 0$. In the following we provide sufficient condition for the concavity of $t \mapsto U(t,\theta,\delta)$ over the intermediate range of tax rates, i.e. we provide sufficient conditions for $U_{tt}(t,\theta,\delta) < 0$, where

$$\begin{aligned} U_{tt}(t,\theta,\delta) &= w_N'(t)\; p_N(t,\theta,\delta) + w_N(t)\; p_N'(t,\theta,\delta) \\ &\quad + w_S'(t)\; p_S(t,\delta) + w_S(t)\; p_S'(t,\delta)\ . \end{aligned}$$

Note that we have already established above that, for all t, $w_N'(t), w_S'(t), p_N'(t,\theta,\delta)$ and $p_S'(t,\delta)$ are all negative, whereas $w_N(t)$ and $w_S(t)$ are positive.

Case 1: $t_N^*(\theta,\delta) < t_S^*(\delta)$. In this case, over the intermediate range of taxes we have $p_N(t,\theta,\delta) < 0$ and $p_S(t,\delta) > 0$. This implies that

$$w_S'(t)\; p_S(t,\delta) + w_S(t)\; p_S'(t,\delta) \quad < \quad 0\ .$$

A sufficient condition for $U_{tt}(\cdot) < 0$, therefore is that also

$$w_N'(t)\; p_N(t,\theta,\delta) + w_N(t)\; p_N'(t,\theta,\delta) \quad < \quad 0\ .$$

After some algebra, $w_N'(t)\; p_N(t,\theta,\delta) + w_N(t)\; p_N'(t,\theta,\delta)$ can be written as

$$q^{-(2+\varepsilon)}\; X(t,\theta,\delta)\ ,$$

where

$$\begin{aligned} X(t,\theta,\delta) &= -\varepsilon\; k\; \Big(2\; \phi_N\; \bar{\theta}_N - \theta\Big) - \delta\; \phi_N\; \bar{\theta}_N\; \varepsilon\; (1+\varepsilon) \\ &\quad + t\; \Big(\delta\; \phi_N\; \bar{\theta}_N\; \varepsilon\; (\varepsilon - 1) + \varepsilon\; \theta\Big) \end{aligned}$$

The following assumptions are sufficient to ensure that $X < 0$ for all θ and δ: Values of θ are drawn from the unit interval, $\bar{\theta}_N > \frac{1}{2}$, $\bar{\theta}_N$ sufficiently close to 1 and k sufficiently large. These assumption imply that $2\; \phi_N\; \bar{\theta}_N - \theta > 0$. Recall that t is, by assumption, bounded between $t_N^*(\theta,\delta)$ and $t_S^*(\delta)$. Thus, by choosing k sufficiently large, we can always make sure that X is negative.

Case 2: $t_N^*(\theta,\delta) > t_S^*(\delta)$. In this case, over the intermediate range of taxes we have $p_N(t,\theta,\delta) > 0$ and $p_S(t,\delta) < 0$. This implies that

$$w_N'(t)\; p_N(t,\delta) + w_N(t)\; p_N'(t,\delta) \quad < \quad 0\ .$$

A sufficient condition for $U_{tt}(\cdot) < 0$, therefore is that

$$w_S'(t)\; p_S(t,\theta,\delta) + w_S(t)\; p_S'(t,\theta,\delta)$$

is sufficiently small. Since, this expression is linear in ϕ_S, it holds for ϕ_S sufficiently small.

A.2 Proof of Proposition 2

Recall that

$$U_t(t, \phi_S, \theta, \delta) \quad = \quad \bar{x}(t, \phi_S) - x^*(q(t), \theta) + (t - \delta)\, \bar{x}'(t, \phi_S)$$

where we have expanded the notation introduced in the main text to emphasize that both per-capita consumption $\bar{x}$ and its derivative with respect to the tax rate depend in a parametric way on the population share of S-agents, ϕ_S. To establish the Proposition, we show that, under the premises of the Proposition $\bar{x}_{\phi_S}(t, \phi_S) < 0$ and $\bar{x}'_{\phi_S}(t, \phi_S) > 0$. In words: with more S-agents the tax base shrinks, which is one force reducing the benefits from higher taxes. Also, the tax base becomes less tax sensitive, as the absolute value $\mid \bar{x}'(t, \phi_S) \mid$ goes down. Together these effects imply that $U_{t\phi_S} > 0$, when $t < \delta$. First, note that

$$\bar{x}(t, \phi_S) \quad = \quad \bar{x}_N(t) - \phi_S\Big(\bar{x}_N(t) - \bar{x}_S(q_S)\Big)$$

where $q_S = k + t + \phi_S(\bar{\delta}_S - t)$. The derivative with respect to ϕ_S equals

$$\bar{x}_{\phi_S}(t, \phi_S) \quad = \quad -\Big(\bar{x}_N(t) - \bar{x}_S(q_S)\Big) \;+\; \phi_S\, \bar{x}'_S(q_S)(\bar{\delta}_S - t)$$

which is a sum of two negative terms. Second, note that

$$\bar{x}'(t, \phi_S) \quad = \quad \bar{x}'_N(t) - \phi_S\Big(\bar{x}'_N(t) - \bar{x}'_S(q_S)(1 - \phi_S)\Big)$$

The derivative with respect to ϕ_S equals

$$\bar{x}'_{\phi_S}(t, \phi_S) \quad = \quad -\Big(\bar{x}'_N(t) + \bar{x}'_S(q_S)\Big) \;-\; \phi_S(1 - \phi_S)\, \bar{x}''_S(q_S)(\bar{\delta}_S - t)\ .$$

The first term in this sum is positive and dominates the second when ϕ_S is sufficiently small.

A.3 Proof of Proposition 3

Under the premises of the Proposition $W_t^S(t)$ can be written as

$$W_t^S(t) \quad = \quad \phi_s\, q^{-\varepsilon}\, \Big(LHS(t) - RHS(t)\Big)$$

where

$$LHS(t) \quad := \quad \bar{\theta}_N - \bar{\theta}_S \left(\frac{q}{q_S}\right)^{\varepsilon}$$

and

$$RHS(t) \quad := \quad \bar{\theta}_N \, \varepsilon \, \frac{t-\delta}{q} \, .$$

where $q = k + t$ and $q_S = k + \phi_N \; t + \phi_S \; \bar{\delta}_S$. Note that $t \mapsto LHS(t)$ is a decreasing function of t and $t \mapsto RHS(t)$ is an increasing function of t. Thus, there exists at most one level of t, henceforth denoted by t_S^*, so that $LHS(t) = RHS(t)$. Moreover, if such a t exists, then $W_t^S(t) > 0$ for $t < t_S^*$ and $W_t^S(t) < 0$ for $t < t_S^*$, which establishes single-peakedness. To show that there exists such $t_S^* >$, note that at $t = \delta$, $RHS(t) = 0$ and $LHS(t) \geq 0$, where the inequality is strict if and only if $\bar{\theta}_N > \bar{\theta}_S$. Thus, if $\bar{\theta}_N = \bar{\theta}_S$, then $t_S^* = \bar{\delta}_S$. Note that $\lim_{t \to \infty} RHS(t) = \bar{\theta}_N \; \varepsilon > \bar{\theta}_N$ and that $LHS(t)$ is bounded from above by $\bar{\theta}_N$. Thus, the functions $t \mapsto LHS(t)$ and $t \mapsto RHS(t)$ eventually cross and if $\bar{\theta}_N > \bar{\theta}_S$, then they cross at $t_S^* > \bar{\delta}_S$.

A.4 Proof of Proposition 4

Note that

$$W_t^S(t, \phi_S) \quad = \quad (1-\phi_S)\Big(\bar{x}_N(t) - \bar{x}_S(q_S)\Big) + (1-\phi_S)(t-\delta_S)\; \bar{x}_N'(t, \phi_S)$$

for $q_S = k + t + \phi_S(\delta_S - t)$. Straightforward computations yield

$$W_{t\phi_S}^S(t, \phi_S) \quad = \quad -\Big(\bar{x}_N(t) - \bar{x}_S(q_S)\Big) - (t-\delta_S)\; \Big(\bar{x}_N'(t) - (1-\phi_S)\bar{x}_S'(q_S)\Big) \, .$$

Using $t \leq \bar{\delta}_S$, $\bar{\theta}_N \geq \bar{\theta}_S$ and $u(x, \theta) = \theta^{\frac{1}{\varepsilon}} \, \frac{1}{1-\frac{1}{\varepsilon}} \, x^{1-\frac{1}{\varepsilon}}$, it is straightforward to verify that

$$\bar{x}_N(t) - \bar{x}_S(q_S) \quad > \quad 0$$

and

$$\bar{x}_N'(t, \phi_S) - (1-\phi_S)\bar{x}_S'(q_S) \quad < \quad 0 \, ,$$

which establishes the result.

A.5 Proof of Proposition 5

For some fixed but arbitrary $\sigma > 0$ and $\sigma_t < 0$, let $a = \phi_S \, \sigma$ and $b = \phi_S \; \sigma_t$. We use a and b to parameterize the function U that is used by an N-agent of type (θ, δ) to evaluate marginal changes of taxes,

$$U_t(t, \theta, \delta, a, b) \quad = \quad \bar{x}(t, a) - x^*(q(t) + a, \theta)(1 + b) - (t - \delta)\, \bar{x}_t(t, a) \; ,$$

where

$$\bar{x}(t,a) \quad = \quad \phi_N\,\bar{\theta}_N\,\big(q(t)+a\big)^{-\varepsilon} \;+\; \phi_S\,\bar{\theta}_S\,q_S(t)^{-\varepsilon}\ ,$$

$$x^*(q(t)+a,\theta) \quad = \quad \theta\,\big(q(t)+a\big)^{-\varepsilon}\ ,$$

and

$$\bar{x}_t(t,a) \quad = \quad -\varepsilon\Big(\phi_N\,\bar{\theta}_N\,\big(q(t)+a\big)^{-(1+\varepsilon)} \;+\; \phi_S\,\phi_N\,\bar{\theta}_S\,q_S(t)^{-(1+\varepsilon)}\Big)\ .$$

Henceforth, we denote the agent's ideal tax by $t^*(\theta,\delta,a,b)$. It solves the equation

$$U_t(t,\theta,\delta,a,b) \quad = \quad 0\ .$$

Obviously,

$$U_{tb}(t,\theta,\delta,a,b) \quad = \quad -x^*(q(t)+a,\theta) \quad < \quad 0\ .$$

Therefore, whatever a, moving from $b=0$ to $b<0$ increases the benefit from higher taxes and pushes the agent's ideal tax up: $t^*(\theta,\delta,a,0) < t^*(\theta,\delta,a,b)$, when $b<0$. Now, to establish the Proposition, we show that, for an agent with $t^*(\theta,\delta,0,0) \le \delta$, and a and b sufficiently small, as implied by ϕ_S small, we have

$$U_{ta}(t^*(\theta,\delta,a,b),\theta,\delta,a,b) \quad > \quad 0\ .$$

Thus, under these conditions, moving from $a=b=0$, to $a>0>b$ shifts the ideal tax up: $t^*(\theta,\delta,0,0) < t^*(\theta,\delta,a,b)$. To verify that, under the stated conditions, $U_{ta}(t^*(\cdot),\theta,\delta,a,b)>0$, note that

$$U_{ta}(t,\theta,\delta,a,b) \quad = \quad -\frac{\varepsilon}{(q(t)+a)^{1+\varepsilon}}\left(\phi_N\,\bar{\theta}_N \;-\; \theta(1+b) \;+\; (t-\delta)(1+\varepsilon)\frac{\phi_N\,\bar{\theta}_N}{q(t)+a}\right)$$

Now, for a and b close to zero, $t^*(\theta,\delta,a,b)$ is close to $t^*(\theta,\delta,0,0)$. Moreover, $t = t^*(\theta,\delta,0,0)$ implies that

$$\phi_N\,\bar{\theta}_N \;-\; \theta \;+\; (t-\delta)\varepsilon\frac{\phi_N\,\bar{\theta}_N}{q(t)}\ .$$

Therefore, the sign of $U_{ta}(t^*(\theta,\delta,a,b),\theta,\delta,a,b)$ is equal to the sign of

$$-\frac{\varepsilon}{(q(t^*(\cdot))+a)^{1+\varepsilon}}\left(-\,\theta\,b \;+\; (t^*(\cdot)-\delta)\frac{\phi_N\,\bar{\theta}_N}{q(t)+a}\right)$$

or, using that $b \simeq 0$,

$$-\frac{\varepsilon}{q(t^*(\cdot))^{1+\varepsilon}}(t^*(\cdot)-\delta)\frac{\phi_N\,\bar{\theta}_N}{q(t^*(\cdot))}$$

when a and b are close to zero. This expression is positive when $t^*(\cdot) \leq \delta$.

A.6 Proof of Lemma 3

Equation (12) implies that, in equilibrium,

$$\hat{\sigma}^{eq}(t)(q + \phi_S\, \hat{\sigma}^{eq}(t))^{1+\varepsilon} \quad = \quad \delta_S\, \phi_N\, \phi_S \int_0^{\hat{\theta}^{eq}(t)} \theta f_N(\theta)\, d\theta\ .$$

By the arguments in the main text, $t \mapsto \hat{\theta}^{eq}(t)$ is an increasing function of t. Hence, the right hand side of the above equation is increasing in t. Therefore, the left hand side needs to be an increasing function of t too. But this requires that $t \mapsto \hat{\sigma}^{eq}(t)$ is an increasing function.

A.7 Proof of Proposition 6

The proof follows from the arguments in the proof of Proposition 5, upon noting that, $U_{ta}(\cdot)$ becomes small when ϕ_S and hence a and b go to zero provided that $\mid t^*(\theta, \delta) - \delta \mid$ is small and also that $\bar{\theta}_N - \theta$ is small for all θ. Therefore, the sign of b governs the comparative statics of the preferred tax. The observations that $b < 0$ in the pressure-without-rebellion-equilibrium and $b > 0$ in the pressure-with-rebellion equilibrium then imply that $t^*_P(\theta, \delta) > t^*(\theta, \delta) > t^*_R(\theta, \delta)$ as had to be shown.

A.8 Universalists

We consider an economy with a third type of agent: Universalists. They are group-rule-utilitarians like the S-agents, but have a different objective. They maximize utilitarian welfare in the economy at large, not just in their own group.

Universalist-rule utilitarians: consumption choices. With obvious adjustments of the notation in the main text, their objective is

$$\begin{aligned} W^U \quad = \quad R(t) \quad &+ \quad \phi_N\, \mathbf{E}_N[u(x^*(q(t), \theta)) \ - \ q\, x^*(q(t), \theta)] \\ &+ \quad \phi_S\, \mathbf{E}_S[u(x_S(t, \theta)) \ - \ q\, x_S(t, \theta)] \\ &+ \quad \phi_U\, \mathbf{E}_U[u(x_U(t, \theta)) \ - \ q\, x_U(t, \theta)] \quad - \quad \mathcal{D}_U(\bar{x}(t))\ , \end{aligned}$$

where $\mathcal{D}_U(\bar{x}(t)) = \mathbf{E}_U[D(\bar{x}(t), \delta)]$. Universalists cannot control the consumption of N or S-agents. They still internalize the utility net of damages realized by other agents. This will matter for their preferences over taxes that we will characterize in the subsequent paragraph. They are paternalist, however, in that they do not respect the view of others on what the damages are. They use the damage function $\mathcal{D}_U$ to compute damages for everyone, irrespectively of whether or not the others agree that this is the right damage function. The objective W_U can equivalently be written as

$$\begin{aligned} W_U &= \phi_N\, \mathbf{E}_N[u(x^*(q(t),\theta)) \;-\; k\, x^*(q(t),\theta)] \\ &+ \phi_S\, \mathbf{E}_S[u(x_S(t,\theta)) \;-\; k\, x_S(t,\theta)] \\ &+ \phi_U\, \mathbf{E}_U[u(x_U(t,\theta)) \;-\; k\, x_U(t,\theta)] \quad - \quad \mathcal{D}_U(\cdot)\ , \end{aligned}$$

Universalists choose $\theta \mapsto x_U(\theta)$ to maximize this expression. Note that their choices do no depend on the tax rate, so that we no longer need to carry t as an argument of x_U. The first order condition characterizing the optimal choice $x^*_U(\theta)$ is

$$u_x(x^*_U(\theta),\theta) \quad = \quad k \;+\; \mathcal{D}'_U(\cdot)\ .$$

This first order condition implies, in particular, that universalist rule-utilitarians show a behavioral response to changes of marginal damages, but not to taxes.

Universalist's preferences over taxes. Their preferences over taxes follow from

$$\begin{aligned} W^U(t) \;\; = \;\; & \phi_N\, \mathbf{E}_{f_N}[u(x^*(q(t),\theta)) \;-\; k\, x^*(q(t),\theta)] \\ & + \phi_S\, \mathbf{E}_{f_S}[u(x_S(t,\theta)) \;-\; k\, x_S(t,\theta)] \\ & + \phi_U\, \mathbf{E}_{f_U}[u(x^*_U(\theta)) \;-\; k\, x^*_U(\theta)] \quad - \quad \mathcal{D}_U((1-\phi_U)\bar{x}_{-U}(t) + \phi_U \bar{x}_U)\ , \end{aligned}$$

where

$$\bar{x}_{-U}(t) = \frac{\phi_N}{1-\phi_U}\ \bar{x}_N(t) \;+ \frac{\phi_S}{1-\phi_U}\ \bar{x}_S(t)\ .$$

Note that

$$\begin{aligned} W^U_t(t) \;\; = \;\; & \phi_N\, \mathbf{E}_{f_N}[(u_x(x^*(q(t),\theta)) \;-\; k)\; x^*_t(q(t),\theta)] \\ & + \phi_S\, \mathbf{E}_{f_S}[(u_x(x_S(t,\theta)) \;-\; k)\; x_{S,t}(t,\theta)] \\ & - \mathcal{D}'_U(\cdot)\ (1-\phi_U)\ \bar{x}'_{-U}(t)\ , \end{aligned}$$

Using the first order conditions characterizing the consumption choices of neoclassical consumers and group rule utilitarians, this can be rewritten as

$$\begin{aligned} W^U_t(t) \;\; = \;\; & \phi_N\ \Big(t - \mathcal{D}'_U(\cdot)\Big)\bar{x}'_N(t) \\ & + \phi_S\Big((1-\phi_S)t + \phi_S\, \mathcal{D}'_S(\cdot) \;\; - \;\; \mathcal{D}'_U(\cdot)\Big)\bar{x}'_S(t)\ . \end{aligned}$$

<u>A special case.</u> Suppose that $\phi_S = 0$ or that $\mathcal{D}_U = \mathcal{D}_S$. Then $t = \mathcal{D}'_U(\cdot)$ implies that $W^U(t) = 0$; $t < \mathcal{D}'_U(\cdot)$ implies $W^U(t) > 0$; $t > \mathcal{D}'_U(\cdot)$ implies $W^U(t) < 0$. *Thus, the function W_U is single- peaked with a peak at $t = \mathcal{D}'_U(\cdot)$.* Thus, Universalist want a pure Pigouvian tax, with no adjustment because of distributive implications.

A.9 Additional Numerical Illustrations

A.9.1 Two-dimensional heterogeneity.

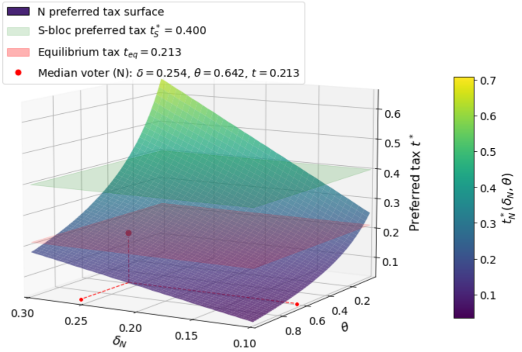
The numerical examples in the main text (Figures 3 and 4) hold the damage parameters fixed at δ_N and δ_S within the N-group and S-group and trace preferred taxes and payoffs along the consumption-taste dimension θ only. We now allow for a non-degenerate joint distribution of (θ, δ) within each group. The preferred tax of an N-agent is decreasing in θ and increasing in δ, and the median voter is the agent with the median position in the resulting two-dimensional ordering of ideal taxes (see Section 4). We replicate the three-regime comparison under such two-dimensional heterogeneity. The qualitative pattern is unchanged. The political equilibrium tax rises from the basic model to the moral-pressure regime and falls back below the basic-model level once rebellion is sustained.

We let $\delta_N \sim$ Uniform$[0.1, 0.3]$ and $\delta_S \sim$ Uniform$[0.3, 0.5]$, with $\theta \sim$ Uniform$[0, 1]$ in both groups and (θ, δ) independent within each group. Other parameters are as before ($\varepsilon = 2$, $k = 0.05$, $\varphi_S = 0.1$, $\alpha = 0.5$, $\beta = 0.2$).

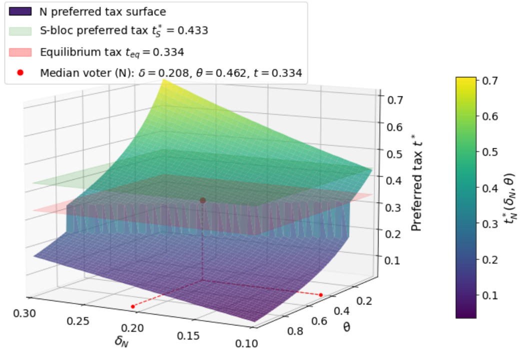
Preferred tax surfaces (Figure 5a–5c). Each of the three panels plots the N-agents' preferred tax $t^*_N(\delta_N, \theta)$ as a (colored) surface over the (δ_N, θ)-plane, the constant S-bloc preferred tax t^*_S as a green plane, and the political equilibrium tax (the median voter's ideal tax in the joint N–S population) as a red plane. The red dot marks the median voter's parameters (δ_M, θ_M) together with her preferred tax. Across the three regimes, the figures show the same qualitative pattern as in the one-dimensional numerical exercise above. Under moral pressure without rebellion, both the N-surface and the equilibrium plane shift upward relative to the basic model. Once a rebellion action is added, the N-surface kinks at the rebellion cutoff $\hat{\theta}^{eq}$, where N-agents with $\theta_N < \hat{\theta}^{eq}$ prefer no rebellion and thus higher taxes, while those with $\theta_N > \hat{\theta}^{eq}$ prefer rebellion and therefore lower taxes. In this illustration the median voter prefers rebellion and the political equilibrium tax falls below the basic-model level.

Equilibrium payoffs (Figure 5d). Figure 5d shows the analogous extension of Figure 4, where N-agents' equilibrium payoffs are plotted as surfaces over (δ, θ), evaluated at the median voter's preferred tax in each regime, alongside the average S-payoff in each regime as a horizontal plane. The comparison across regimes mirrors Figure 4. The moral-pressure surface lies above the basic-model surface for small θ but below it for high-θ types, while the rebellion surface is below the basic-model surface for the bulk of types. We recover the Pareto-inferior outcome in the rebellion equilibrium highlighted above.

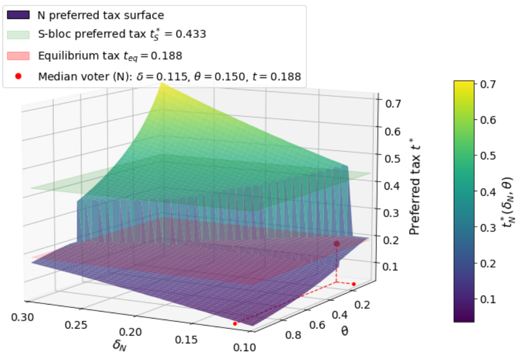
The comparative statics of moral pressure and moral rebellion that drive the main



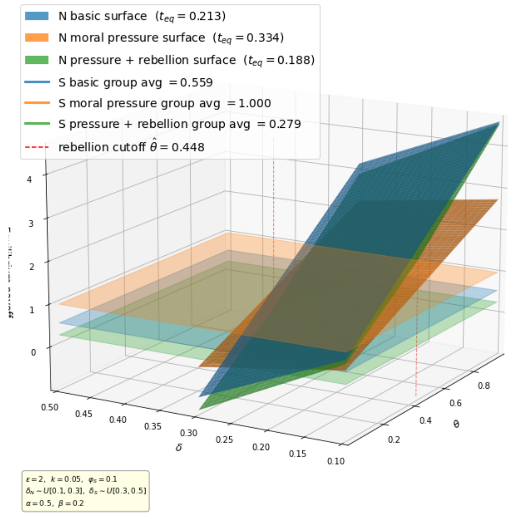
(a) Preferred Taxes: Basic Model



(b) Preferred Taxes: Moral Pressure



(c) Preferred Taxes: Moral Rebellion



(d) Political-Equilibrium Payoffs

Figure 5: Two-Dimensional Heterogeneity: Preferred Taxes and Equilibrium Payoffs
Notes: Panels (a)–(c) show N-agents' preferred tax $t_N^*(\delta_N, \theta)$ (colored surface), the S-bloc's preferred tax t_S^* (green plane), and the political-equilibrium tax t_{eq} (red plane) across the three regimes; t_{eq} equals the preferred tax of the median voter in the joint N–S population (red dot). Panel (d) plots equilibrium N-payoffs (colored surfaces) and average S-payoffs (horizontal planes); the vertical red dashed line marks $\hat{\theta}^{eq}$. Panel (d) is the two-dimensional analogue of Figure 4. Parameters: $\varepsilon = 2$, $k = 0.05$, $\varphi_S = 0.1$, $\alpha = 0.5$, $\beta = 0.2$; $\theta \sim \text{Uniform}[0, 1]$, $\delta_N \sim \text{Uniform}[0.1, 0.3]$, $\delta_S \sim \text{Uniform}[0.3, 0.5]$, with θ and δ drawn independently within each group.

propositions and the one-dimensional numerical illustration carry over.

A.9.2 Heterogeneous mean tastes: $\bar{\theta}_N > \bar{\theta}_S$

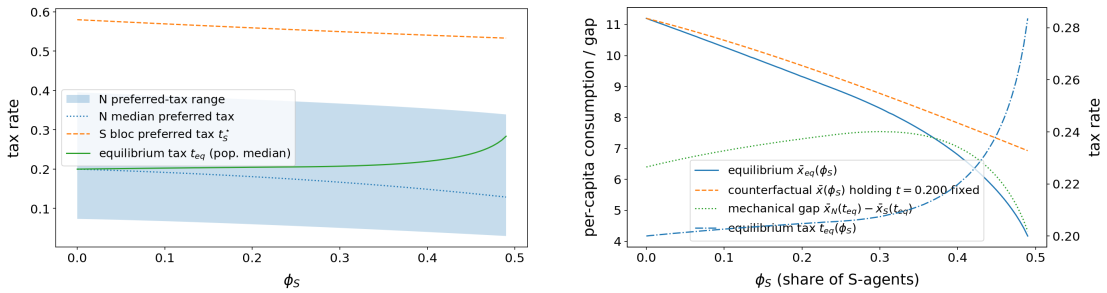
In the main text we focus on $\bar{\theta}_N = \bar{\theta}_S$, which isolates the political-economy mechanisms without bringing in the additional overtaxation motive of Proposition 3. Here we reproduce the political-equilibrium and consumption diagnostics under $\bar{\theta}_N > \bar{\theta}_S$, drawing $\theta_N \sim U[0.4, 1.0]$ ($\bar{\theta}_N = 0.7$) and $\theta_S \sim U[0.0, 0.6]$ ($\bar{\theta}_S = 0.3$). All other primitives are as in the main text.

Three features distinguish this case. First, the S-bloc's preferred tax (orange dashed in Panel (a) of Figure 6) sits strictly above $\bar{\delta}_S = 0.40$ throughout: it begins at $t^*_S(0) \approx 0.58$ and falls to ≈ 0.53 at $\phi_S = 0.5$. This is the overtaxation motive of Proposition 3 — when $\bar{\theta}_N > \bar{\theta}_S$, S-agents prefer a tax above their own Pigouvian benchmark $\bar{\delta}_S$ in order to lean on the N-bloc's larger demand for the dirty good. The negative slope is consistent with the logic of Proposition 4: as the N tax base shrinks with ϕ_S, the redistributive and corrective benefits of higher taxes that the S-bloc draws from the N-bloc both weaken, and the S-bloc's ideal tax falls.

Second, the political-equilibrium tax (green solid) rises monotonically as in the main text, but never reaches t^*_S within the plotted range. The decisive voter remains an N-agent throughout: t^*_S sits well above the upper edge of the N preferred-tax distribution, and even as the S-bloc's voting weight grows toward $\frac{1}{2}$, the cumulative-weight threshold for the decisive voter to switch to the S-bloc is not crossed for $\phi_S < \frac{1}{2}$. The visible acceleration of t^{eq} near $\phi_S = \frac{1}{2}$ reflects the population median walking through the upper end of the N preferred-tax range.

Third, in Panel (b) of Figure 6 the cross-sectional consumption gap $\bar{x}_N(t^{eq}) - \bar{x}_S(t^{eq})$ (green dotted) no longer starts at zero. With $\bar{\theta}_N \neq \bar{\theta}_S$, the two groups consume differently even when they face the same effective price, and at $\phi_S = 0$ the gap is already $(\bar{\theta}_N - \bar{\theta}_S)/(k + t^{eq})^{\varepsilon} \approx 6.4$. The price wedge $q_S - q_N = \phi_S(\bar{\delta}_S - t^{eq})$ still modulates the gap with ϕ_S, but its lower envelope is now bounded away from zero by the $\bar{\theta}$-difference.

The remaining elements of the picture, the composition channel through the dashed counterfactual, policy channel through the rising equilibrium tax, and reinforcing rather than offsetting behavioural responses across the two groups, are unchanged from the main-text scenario.



(a) Political-Equilibrium Tax

(b) Aggregate Consumption

Figure 6: Weak Waterbed Effect under $\bar{\theta}_N > \bar{\theta}_S$

Notes: Panel (a): The shaded blue band gives the cross-sectional range of N-agents' preferred taxes over the support of (θ_N, δ_N). The dotted blue line is the within-N median preferred tax. The orange dashed line is the S-bloc's preferred tax t_S^*, which exceeds $\bar{\delta}_S = 0.40$ throughout (Proposition 3) and falls strictly in ϕ_S (Proposition 4). The green solid line is the political equilibrium tax, which is the median preferred tax over the entire population. The decisive voter remains an N-agent throughout the plotted range. Panel (b): The blue solid line is equilibrium per-capita consumption $\bar{x}^{eq}(\phi_S) = (1-\phi_S)\bar{x}_N(t^{eq}) + \phi_S \bar{x}_S(q_S(t^{eq}, \phi_S))$. The orange dashed line holds the tax fixed at $t^{eq}(\phi_S{=}0) = 0.20$, isolating the composition channel. The green dotted line is the consumption gap $\bar{x}_N(t^{eq}) - \bar{x}_S(t^{eq})$, which remains strictly positive at $\phi_S = 0$ because $\bar{\theta}_N \neq \bar{\theta}_S$ even when $q_N = q_S$. The blue dash-dotted line (right axis) is $t^{eq}(\phi_S)$ from panel (a). Parameters: $\varepsilon = 2$, $k = 0.05$, $\theta_N \sim \text{Uniform}[0.4, 1.0]$, $\theta_S \sim \text{Uniform}[0.0, 0.6]$, $\delta_N \sim \text{Uniform}[0.1, 0.3]$, $\delta_S \sim \text{Uniform}[0.3, 0.5]$.

A.10 Empirical Appendix

A.10.1 Tax Response to Damage Increases

Table 2 examines preferred tax levels. Here we turn to how respondents adjust their preferred tax when told that the social cost of carbon rises by 30%, either only in their own country (*Local Tax Adjustment*, Table 6) or uniformly across all countries (*Global Tax Adjustment*, Table 7), holding the previously chosen tax as the reference point.[46]

The patterns are broadly consistent with the tax-level results in Table 2. Climate concern (δ) is strongly significant in all specifications. *Price vs. Harm* is significant everywhere, and θ is weakly negative. In both tables, the Damage Response is significant on its own but loses significance when θ and δ are included. In the US, the conditional coefficient on the Damage Response turns positive (though insignificant), unlike in the tax-level regression where it is essentially zero.

A.10.2 Ordered Probit Specification for Moral Pressure Outcomes

The moral pressure outcomes *MP Approval* and *MP Opposition* are measured on five-point Likert scales (1 = Strongly disagree, 5 = Strongly agree). In the main text Ta-

[46] *Local* (resp. *Global*) *Tax Adjustment* is the respondent's percentage change in their preferred carbon tax (slider from −100% to +100%, "−100% Abandon the tax" to "+100% Double the tax") in response to: "Now suppose that only in [your country] (resp., across all countries globally, including [your country]), the social cost of carbon … has increased by 30%. You previously chose a carbon tax of $[X] per ton. How would you adjust your preferred carbon tax rate in [your country]?"

Table 6: Tax Preferences — Responsiveness to Local Damage Information

	Dependent variable: Tax Adjustment (%, Local)									
	UK	UK	UK	UK	UK	US	US	US	US	US
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Brown Consumption (θ)	-2.560^{*}		-2.365		-2.265	-2.524^{*}		-2.684^{*}		-2.396
	(1.464)		(1.472)		(1.399)	(1.513)		(1.505)		(1.464)
Climate Concern (δ)	16.373^{***}		15.883^{***}		12.907^{***}	21.669^{***}		22.535^{***}		17.432^{***}
	(1.402)		(1.400)		(1.451)	(1.486)		(1.527)		(1.578)
Damage Response		-5.932^{***}	-2.129				-4.842^{***}	2.618		
		(1.380)	(1.301)				(1.824)	(1.652)		
Price vs. Harm				14.142^{***}	9.183^{***}				17.211^{***}	9.937^{***}
				(1.473)	(1.527)				(1.512)	(1.523)
Constant	-11.273^{***}	-11.254^{***}	-11.275^{***}	-11.078^{***}	-11.105^{***}	-15.679^{***}	-15.717^{***}	-15.679^{***}	-15.764^{***}	-15.704^{***}
	(1.292)	(1.389)	(1.291)	(1.323)	(1.261)	(1.393)	(1.559)	(1.391)	(1.459)	(1.361)
Observations	893	894	893	890	889	900	901	900	900	899
R^2	0.155	0.020	0.157	0.114	0.198	0.212	0.011	0.215	0.134	0.249
Adjusted R^2	0.153	0.019	0.155	0.113	0.195	0.210	0.010	0.212	0.133	0.246

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored). *p<0.1; **p<0.05; ***p<0.01.

Table 7: Tax Preferences — Responsiveness to Global Damage Information

	Dependent variable: Tax Adjustment (%, Global)									
	UK	UK	UK	UK	UK	US	US	US	US	US
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Brown Consumption (θ)	-2.455^{*}		-2.359		-2.184	-3.895^{**}		-3.960^{***}		-3.763^{**}
	(1.454)		(1.465)		(1.405)	(1.518)		(1.518)		(1.486)
Climate Concern (δ)	14.940^{***}		14.701^{***}		11.780^{***}	20.823^{***}		21.175^{***}		17.225^{***}
	(1.431)		(1.428)		(1.487)	(1.506)		(1.581)		(1.572)
Damage Response		-4.593^{***}	-1.041				-6.002^{***}	1.064		
		(1.438)	(1.374)				(1.824)	(1.660)		
Price vs. Harm				13.031^{***}	8.464^{***}				15.621^{***}	8.453^{***}
				(1.459)	(1.536)				(1.536)	(1.490)
Constant	-4.554^{***}	-4.563^{***}	-4.555^{***}	-4.401^{***}	-4.400^{***}	-11.625^{***}	-11.668^{***}	-11.625^{***}	-11.681^{***}	-11.620^{***}
	(1.266)	(1.353)	(1.267)	(1.291)	(1.240)	(1.351)	(1.506)	(1.352)	(1.429)	(1.329)
Observations	893	894	893	890	889	900	901	900	900	899
R^2	0.137	0.013	0.138	0.103	0.176	0.211	0.017	0.211	0.117	0.239
Adjusted R^2	0.135	0.012	0.135	0.102	0.174	0.209	0.016	0.209	0.116	0.236

Notes: Robust (HC1) standard errors in parentheses. All independent variables are standardized (z-scored). *p<0.1; **p<0.05; ***p<0.01.

bles 3 and 4, we treat these outcomes as cardinal and estimate the conditional mean by OLS with HC1 standard errors. This has the advantage of yielding coefficients that are directly interpretable as marginal effects on the response scale, and it facilitates comparison with the tax-level and tax-adjustment regressions reported elsewhere in the paper.

Because the outcomes are strictly ordinal, we also re-estimate all specifications by ordered probit, which treats the five response categories as ordered but does not impose equal spacing between them. The specifications are identical to those in Tables 3 and 4.

The ordered probit results are qualitatively identical to the OLS results (see Tables 8 and 9). Signs, significance levels, and the relative magnitudes of the coefficients are preserved across all eight specifications in both countries. We conclude that the OLS estimates reported in the main text are not driven by the cardinality assumption.

Table 8: Support and Opposition to Moral Pressure — UK (Ordered Probit)

	Dependent variable:							
	MP Approval				MP Opposition			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Damage Response	-0.221^{***}		-0.108^{***}		0.199^{***}		0.094^{**}	
	(0.036)		(0.038)		(0.036)		(0.037)	
Price vs. Harm		0.470^{***}		0.321^{***}		-0.375^{***}		-0.235^{***}
		(0.038)		(0.041)		(0.037)		(0.039)
Brown Consumption (θ)			-0.098^{***}	-0.106^{***}			0.150^{***}	0.158^{***}
			(0.037)	(0.037)			(0.036)	(0.036)
Climate Concern (δ)			0.582^{***}	0.510^{***}			-0.514^{***}	-0.457^{***}
			(0.040)	(0.042)			(0.039)	(0.041)
McFadden RÂ	0.015	0.062	0.104	0.127	0.012	0.041	0.084	0.096
Observations	892	888	891	887	893	889	892	888

Notes: Ordered probit estimates; standard errors in parentheses. All independent variables are standardized (z-scored). The dependent variables are *MP Approval* and *MP Opposition*, each measured on a five-point Likert scale (1 = Strongly disagree, 5 = Strongly agree); see Table 3 for the question wording. All covariates are defined as in Tables 1 and 3. Cut-point estimates are suppressed for readability. *p<0.1; **p<0.05; ***p<0.01.

Table 9: Support and Opposition to Moral Pressure — US (Ordered Probit)

	Dependent variable:							
	MP Approval				MP Opposition			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Damage Response	-0.157^{***}		0.044		0.263^{***}		0.102^{***}	
	(0.035)		(0.038)		(0.036)		(0.038)	
Price vs. Harm		0.447^{***}		0.258^{***}		-0.364^{***}		-0.170^{***}
		(0.037)		(0.040)		(0.036)		(0.039)
Brown Consumption (θ)			-0.092^{**}	-0.089^{**}			0.069^{*}	0.075^{**}
			(0.036)	(0.036)			(0.036)	(0.036)
Climate Concern (δ)			0.677^{***}	0.569^{***}			-0.593^{***}	-0.560^{***}
			(0.042)	(0.042)			(0.041)	(0.042)
McFadden RÂ	0.007	0.055	0.11	0.126	0.02	0.038	0.101	0.105
Observations	901	900	900	899	899	898	898	897

Notes: Ordered probit estimates; standard errors in parentheses. All independent variables are standardized (z-scored). The dependent variables are *MP Approval* and *MP Opposition*, each measured on a five-point Likert scale (1 = Strongly disagree, 5 = Strongly agree); see Table 3 for the question wording. All covariates are defined as in Tables 1 and 3. Cut-point estimates are suppressed for readability. *p<0.1; **p<0.05; ***p<0.01.

A.11 Survey Instructions

This subsection documents the full survey flow shown to respondents. Screens are presented in the order they appeared. Where a screen had a separate version for U.S. respondents, the U.S. version is shown immediately after the baseline (UK) version. Where a screen has an *adjusted* version, this shows how the screen looked after the respondent moved the slider.

Introduction and Screening

Figure 7: Introduction screen (1/2).

Purpose of the Study

You are invited to take part in a research survey conducted by researchers at the University of Cologne, the Humboldt University Berlin and the Vanderbilt University in Nashville Tennessee.

What Participation Involves

If you agree to participate, you will be asked to complete an online questionnaire. The survey will take approximately **10 minutes** to complete. There are **no right or wrong answers** —we are interested in your honest opinions.

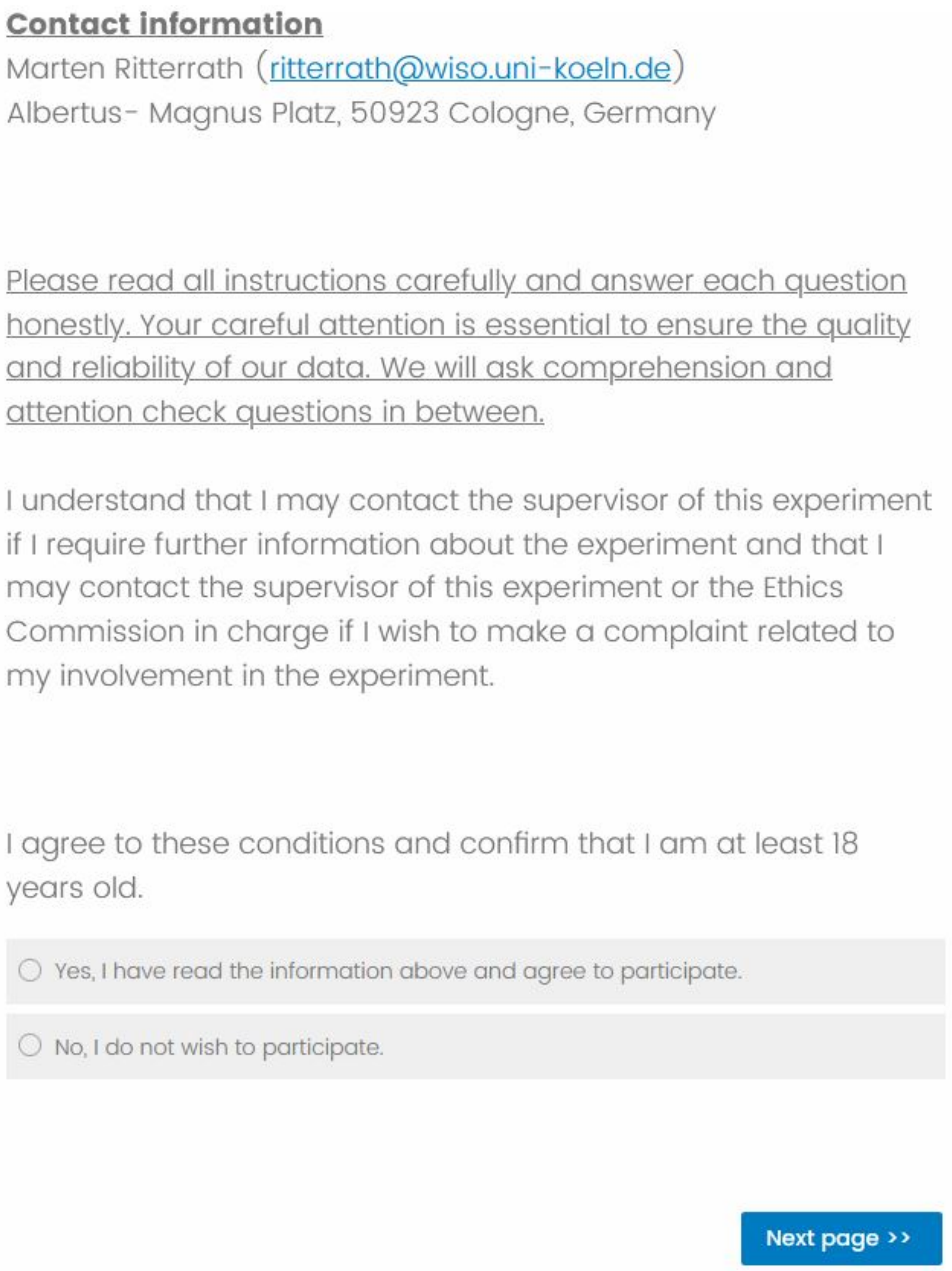
Confidentiality

All statements in this experiment are made anonymously and do not allow to draw conclusions regarding individual participants. Data will be used for scientific purposes only and will be saved only for scientific data analysis. To guarantee transparency in science and in the context of scientific cooperation projects, the collected, anonymized data may be made available for subsequent use by third parties.

Participant rights

Your participation in this experiment is voluntary. To use your data for research, you need to complete the entire experiment. You may withdraw from participation at any time during the experiment without giving any reasons. In the event of withdrawal from the experiment or failure to provide a completion code, you will not be eligible for payment.

Figure 8: Introduction screen (2/2).



Contact information

Marten Ritterrath (ritterrath@wiso.uni-koeln.de)
Albertus- Magnus Platz, 50923 Cologne, Germany

Please read all instructions carefully and answer each question honestly. Your careful attention is essential to ensure the quality and reliability of our data. We will ask comprehension and attention check questions in between.

I understand that I may contact the supervisor of this experiment if I require further information about the experiment and that I may contact the supervisor of this experiment or the Ethics Commission in charge if I wish to make a complaint related to my involvement in the experiment.

I agree to these conditions and confirm that I am at least 18 years old.

- ○ Yes, I have read the information above and agree to participate.
- ○ No, I do not wish to participate.

Next page >>

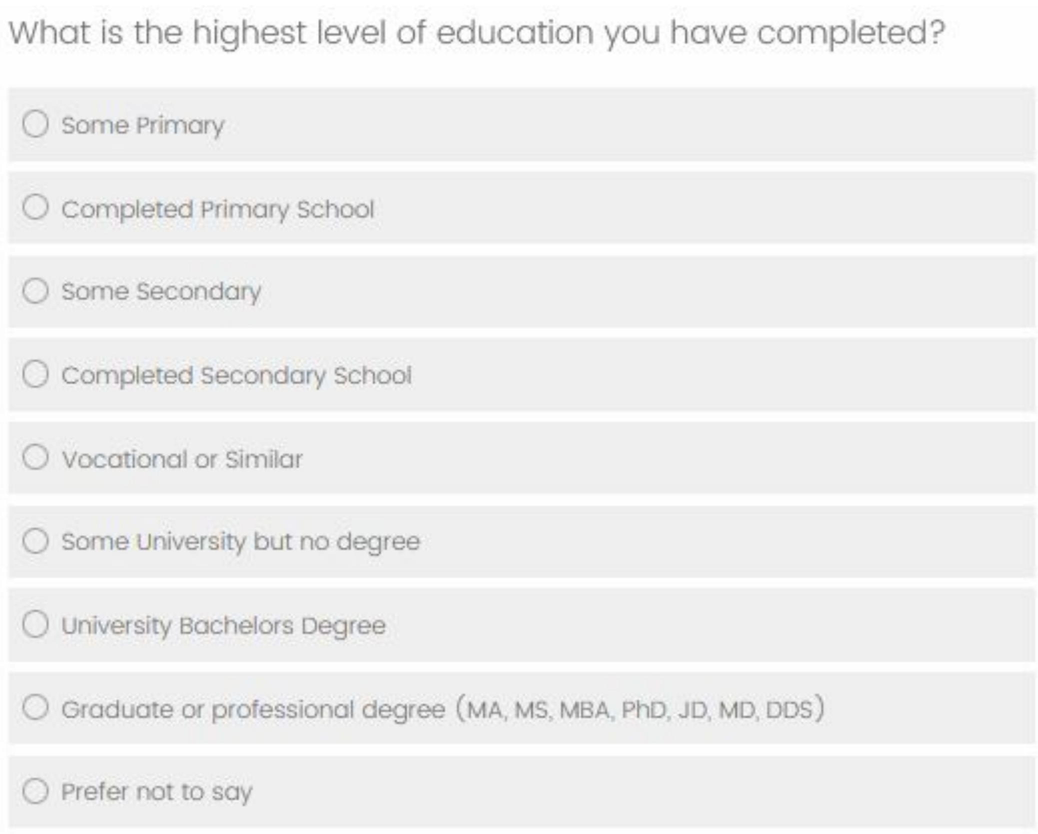
Figure 9: CAPTCHA screen.

Figure 10: Prolific ID entry.

Demographics

Figure 11: Gender question.

Figure 12: Education question (UK version).


What is the highest level of education you have completed?

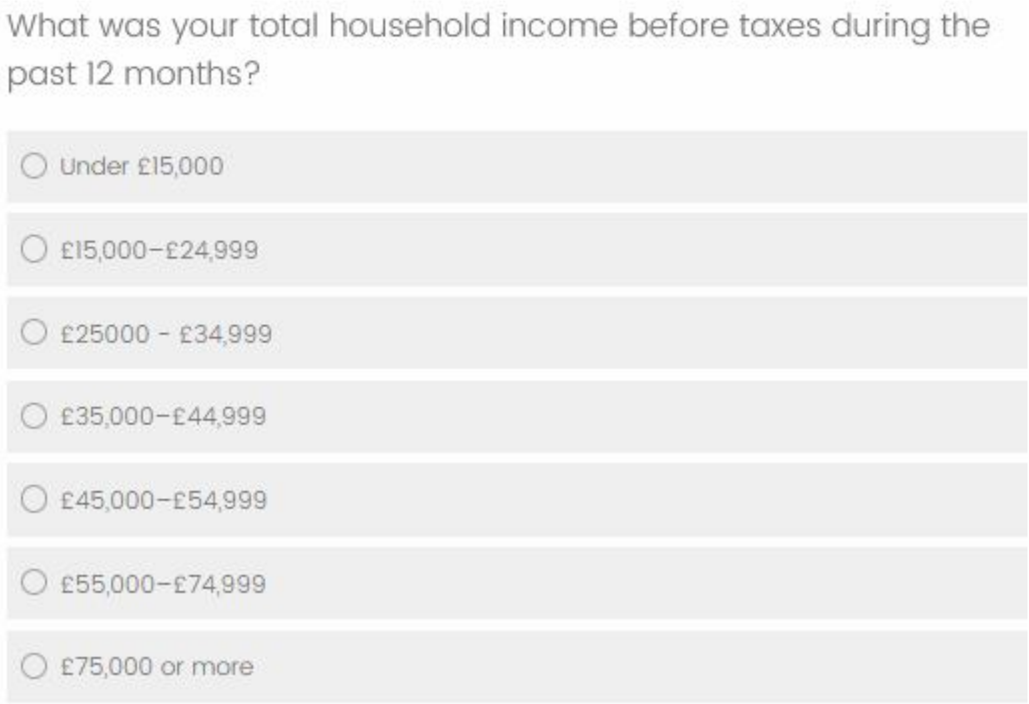
- Some Primary
- Completed Primary School
- Some Secondary
- Completed Secondary School
- Vocational or Similar
- Some University but no degree
- University Bachelors Degree
- Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS)
- Prefer not to say

Figure 13: Education question (U.S. version). The education question was adapted to reflect the U.S. educational system.


What is the highest level of education you have completed?

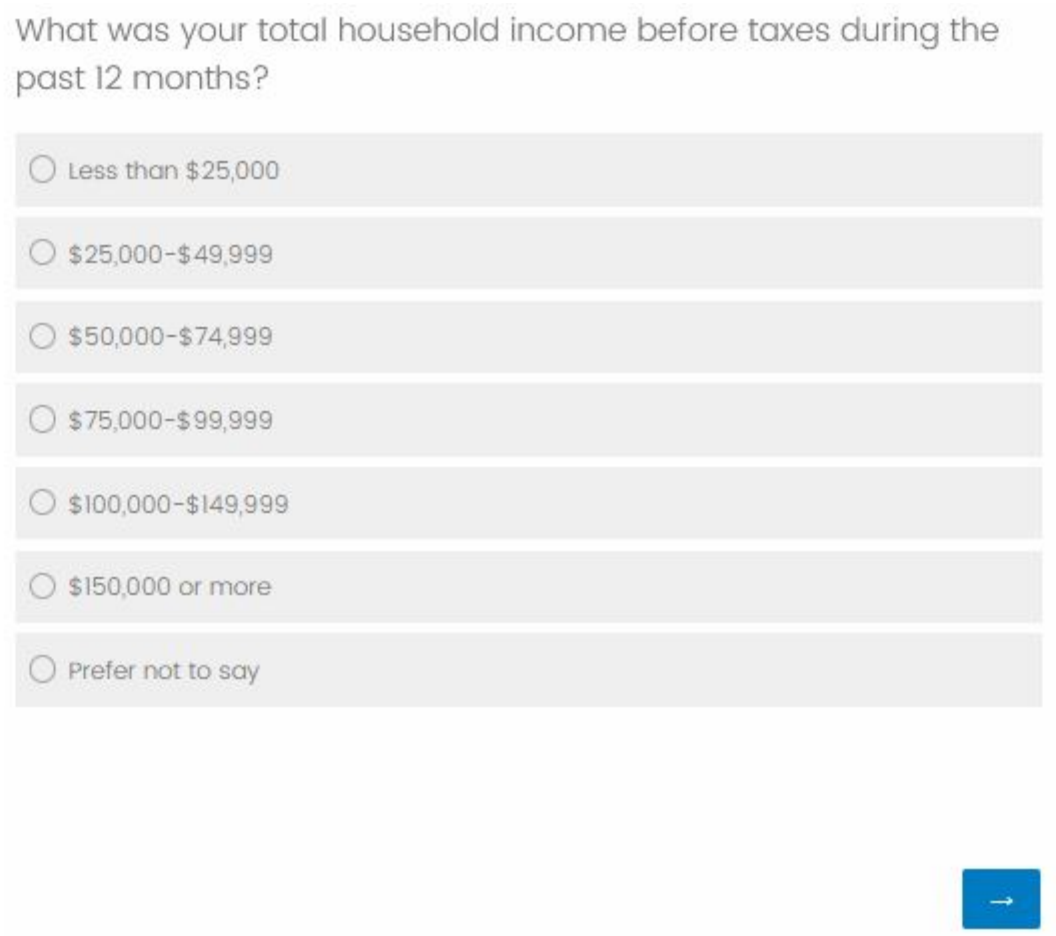
- Some high school or less
- High school diploma or GED
- Some college, but no degree
- Associates or technical degree
- Bachelor's degree
- Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)
- Prefer not to say

Figure 14: Income question (UK version).


What was your total household income before taxes during the past 12 months?

- Under £15,000
- £15,000–£24,999
- £25000 - £34,999
- £35,000–£44,999
- £45,000–£54,999
- £55,000–£74,999
- £75,000 or more

Figure 15: Income question (U.S. version). The income question was adapted to reflect U.S. income brackets (in U.S. dollars).


What was your total household income before taxes during the past 12 months?

- Less than $25,000
- $25,000-$49,999
- $50,000-$74,999
- $75,000-$99,999
- $100,000-$149,999
- $150,000 or more
- Prefer not to say

→

Car Ownership and Usage

Figure 16: Car type question.


What kind of car do you drive?
(select all that apply)

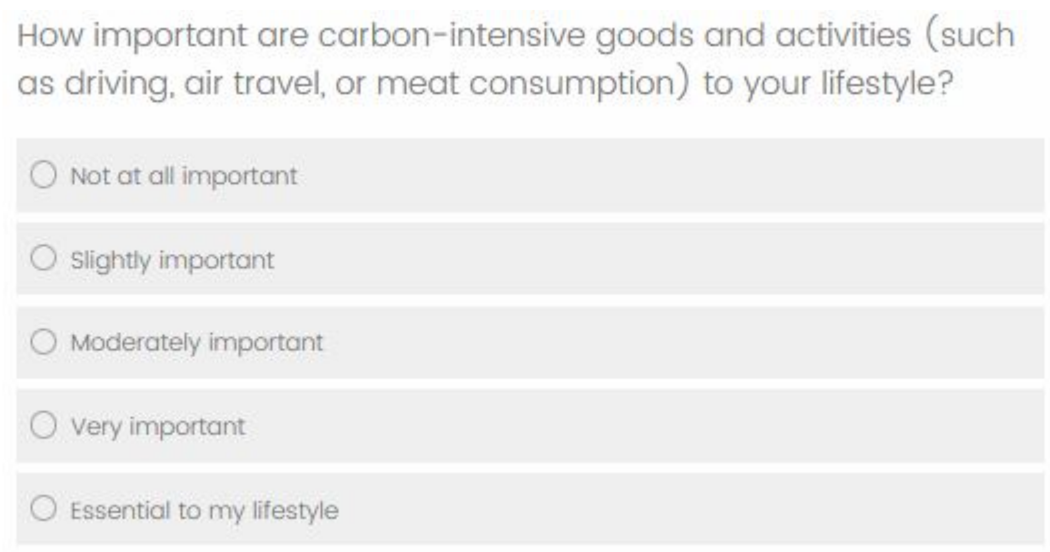
- Gasoline powered car
- Electric car
- I do not have a car
- Hybrid car

Figure 17: Car usage question (UK version).

Figure 18: Car usage question (U.S. version). The car usage question was adapted to reflect U.S. driving patterns.

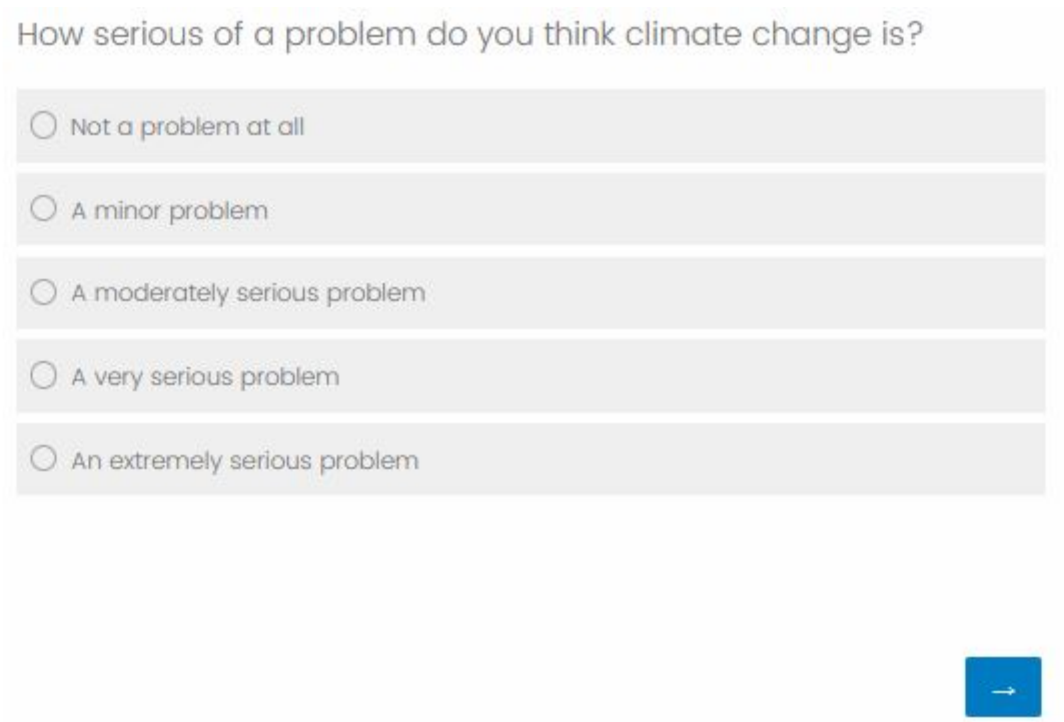
θ and δ

Figure 19: Elicitation of θ.


How important are carbon-intensive goods and activities (such as driving, air travel, or meat consumption) to your lifestyle?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Essential to my lifestyle

Figure 20: Elicitation of δ.


How serious of a problem do you think climate change is?

- Not a problem at all
- A minor problem
- A moderately serious problem
- A very serious problem
- An extremely serious problem

Carbon Tax Block

Figure 21: Carbon tax information screen.

Please read the following information carefully. You will be asked a few comprehension questions about it afterward.

A **carbon tax** is a policy that puts a tax on carbon emissions. It works by increasing the cost of goods and services that lead to more CO_2 emissions—such as gasoline, heating fuel, and electricity generated from fossil fuels. This tax is based on consumption: the more a person consumes goods that produce emissions, the more they pay in carbon taxes.

Suppose such a tax is implemented as follows: the money collected from the tax is not kept by the government. Instead, it is fully returned to the public as equal lump-sum payments to all people. That means everyone receives the same amount back, regardless of how much they paid in carbon taxes. People who emit less than average will get more back than they paid, while those who emit more than average will pay more than they get back.

Figure 22: Carbon tax comprehension check.

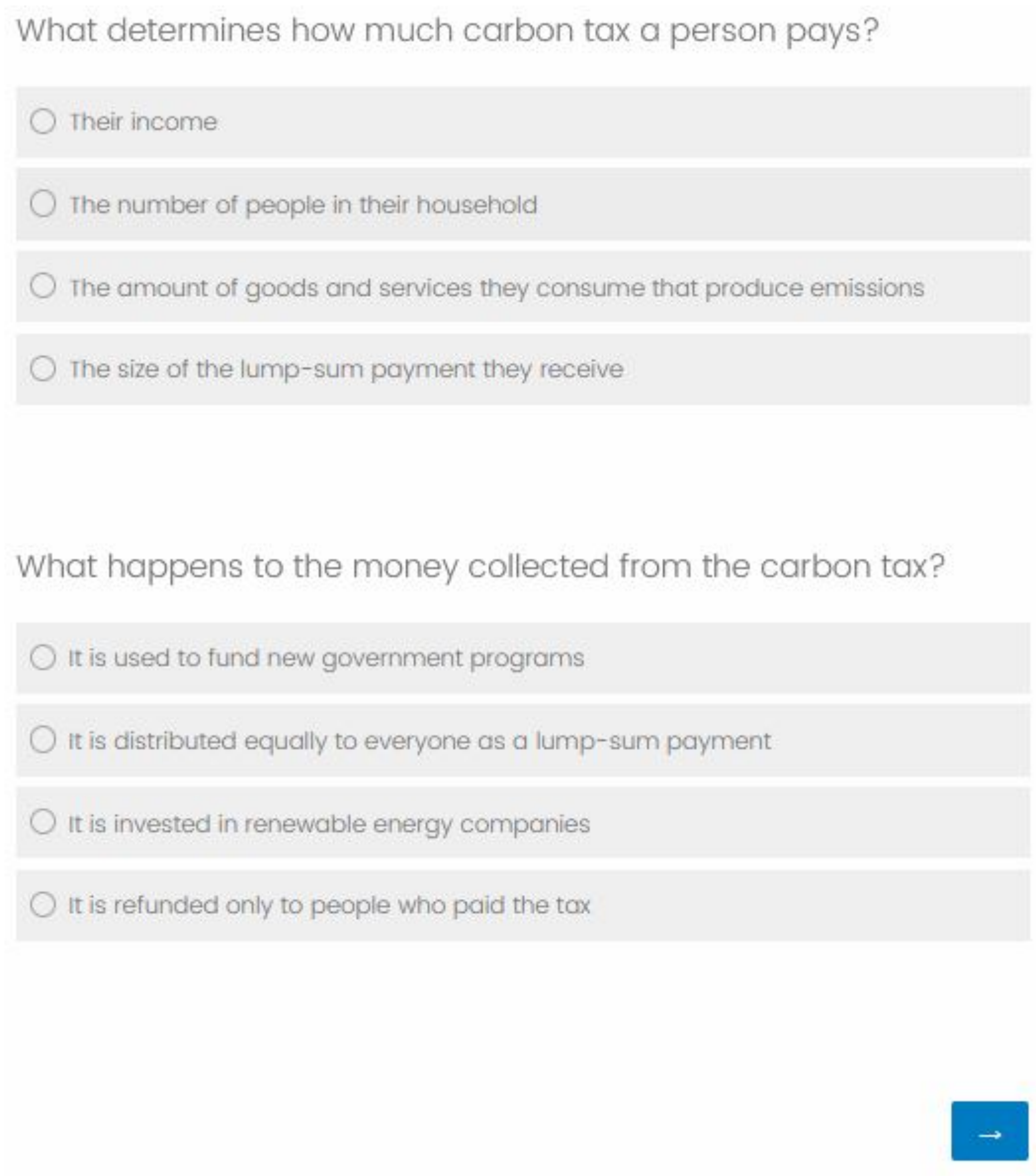
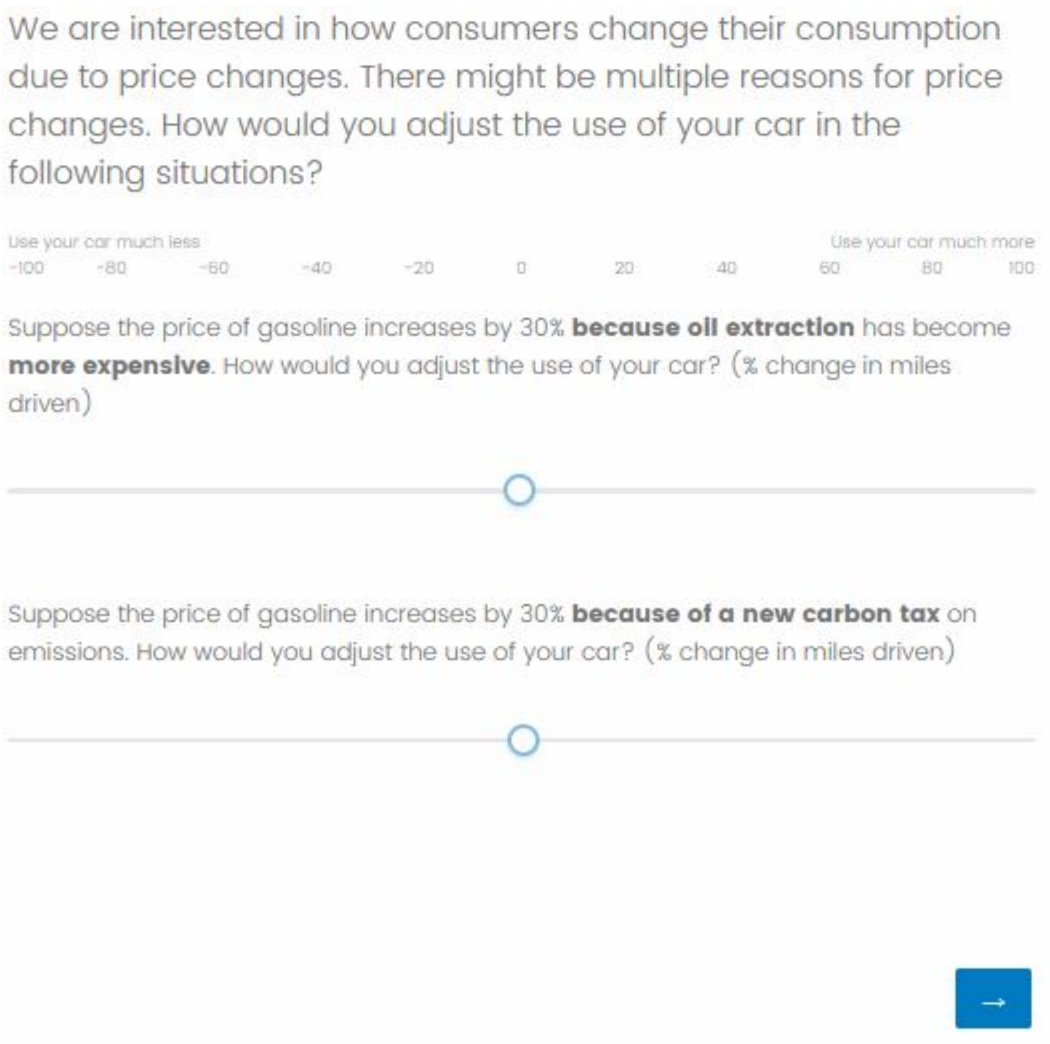


Figure 23: Price adjustment under carbon tax.



We are interested in how consumers change their consumption due to price changes. There might be multiple reasons for price changes. How would you adjust the use of your car in the following situations?

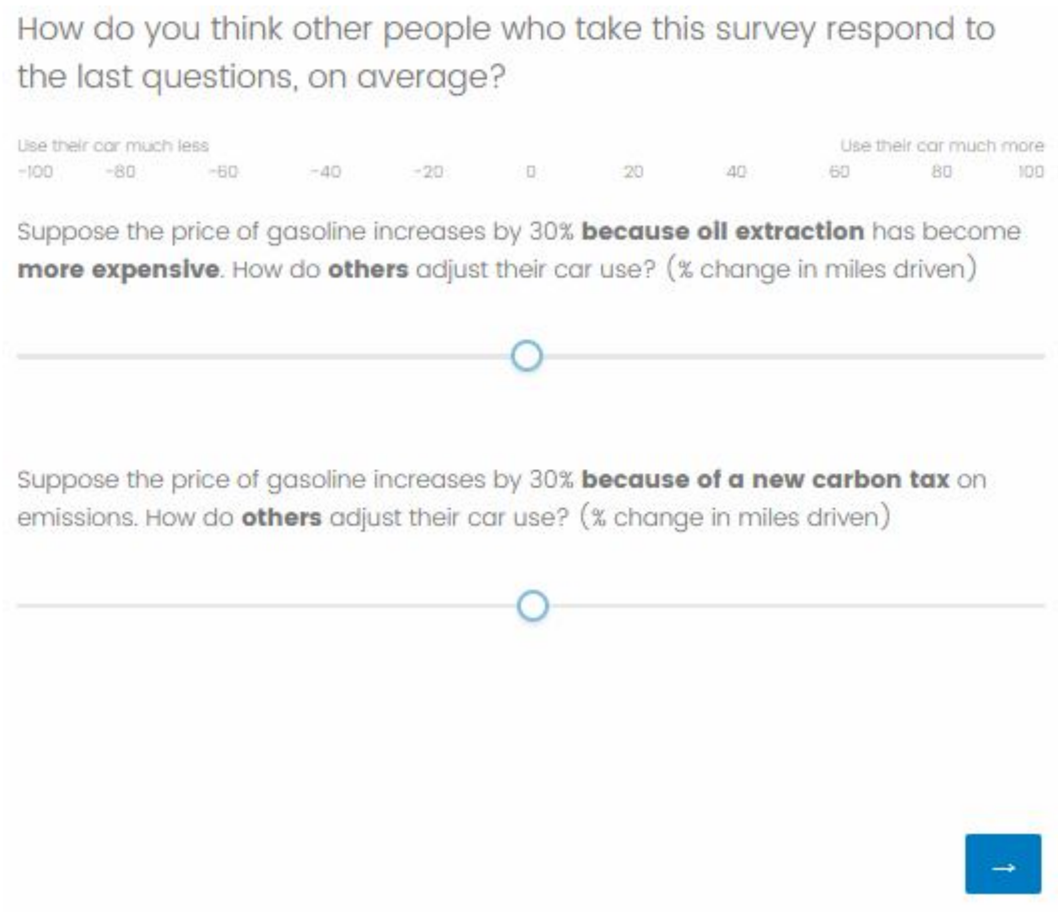
Use your car much less										Use your car much more
-100	-80	-60	-40	-20	0	20	40	60	80	100

Suppose the price of gasoline increases by 30% **because oil extraction** has become **more expensive**. How would you adjust the use of your car? (% change in miles driven)

Suppose the price of gasoline increases by 30% **because of a new carbon tax** on emissions. How would you adjust the use of your car? (% change in miles driven)

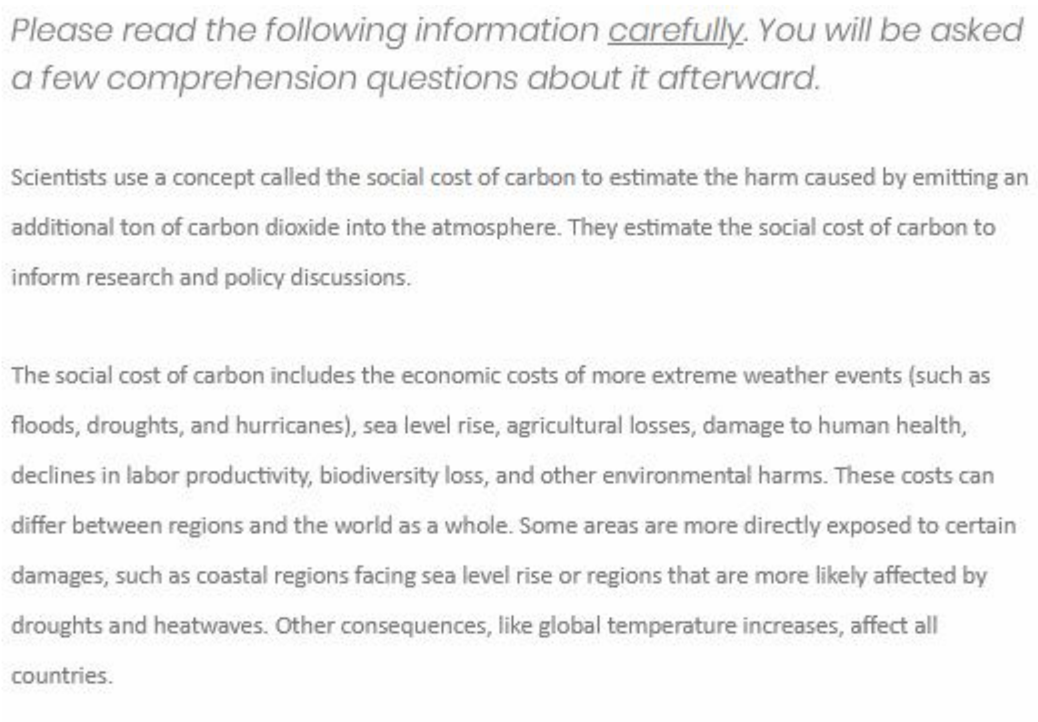
→

Figure 24: Price under carbon tax, hold-out version.



Social Cost of Carbon Block

Figure 25: Social cost of carbon information screen.

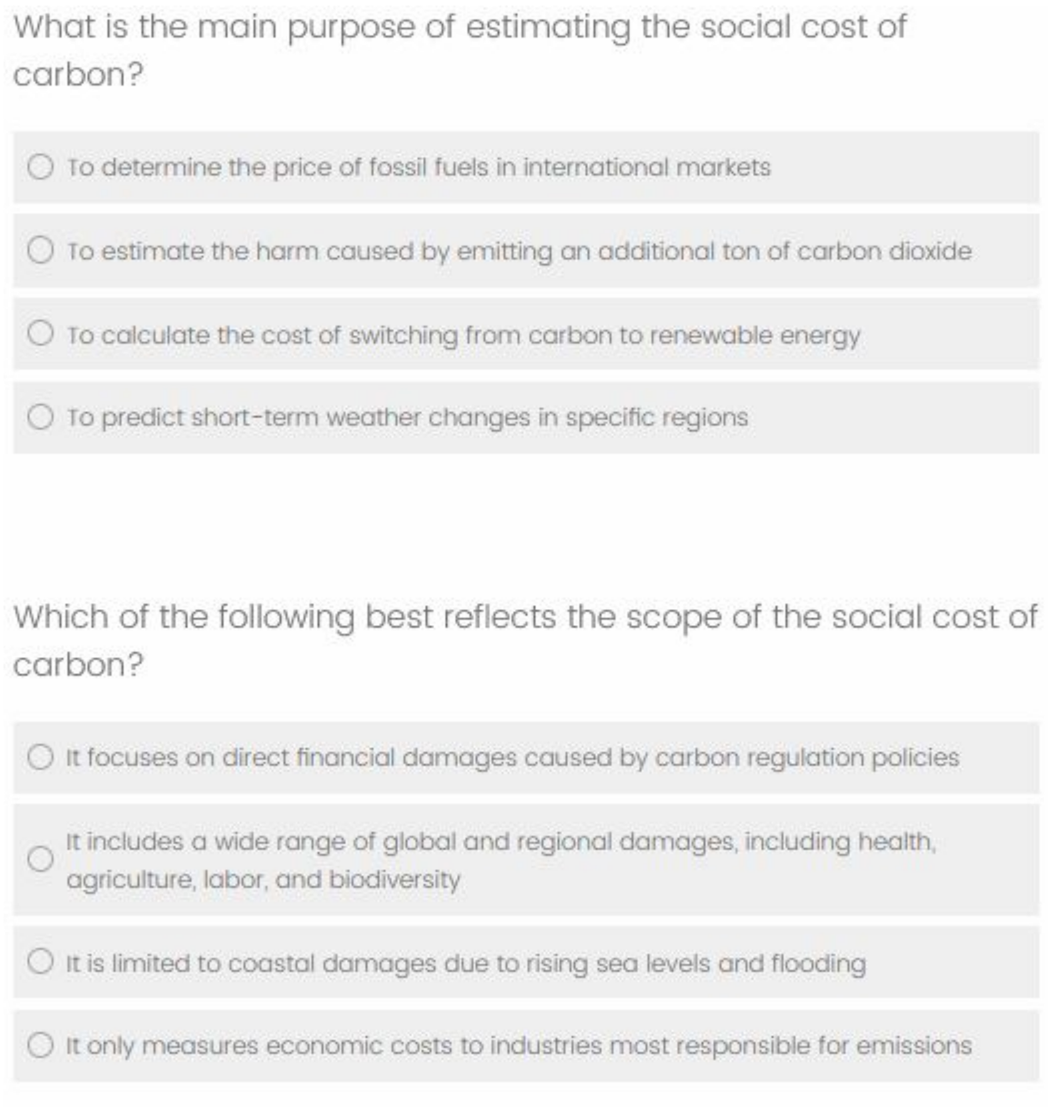


Please read the following information carefully. You will be asked a few comprehension questions about it afterward.

Scientists use a concept called the social cost of carbon to estimate the harm caused by emitting an additional ton of carbon dioxide into the atmosphere. They estimate the social cost of carbon to inform research and policy discussions.

The social cost of carbon includes the economic costs of more extreme weather events (such as floods, droughts, and hurricanes), sea level rise, agricultural losses, damage to human health, declines in labor productivity, biodiversity loss, and other environmental harms. These costs can differ between regions and the world as a whole. Some areas are more directly exposed to certain damages, such as coastal regions facing sea level rise or regions that are more likely affected by droughts and heatwaves. Other consequences, like global temperature increases, affect all countries.

Figure 26: Social cost of carbon comprehension check.



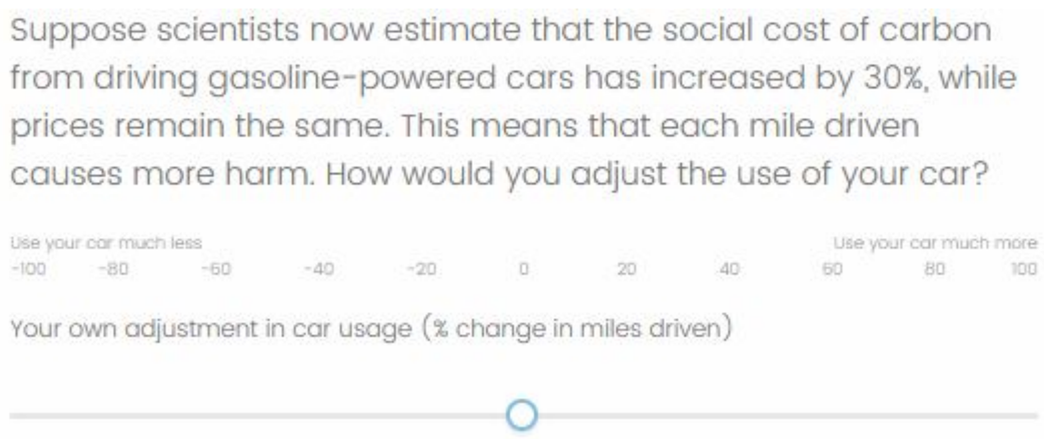
What is the main purpose of estimating the social cost of carbon?

- To determine the price of fossil fuels in international markets
- To estimate the harm caused by emitting an additional ton of carbon dioxide
- To calculate the cost of switching from carbon to renewable energy
- To predict short-term weather changes in specific regions

Which of the following best reflects the scope of the social cost of carbon?

- It focuses on direct financial damages caused by carbon regulation policies
- It includes a wide range of global and regional damages, including health, agriculture, labor, and biodiversity
- It is limited to coastal damages due to rising sea levels and flooding
- It only measures economic costs to industries most responsible for emissions

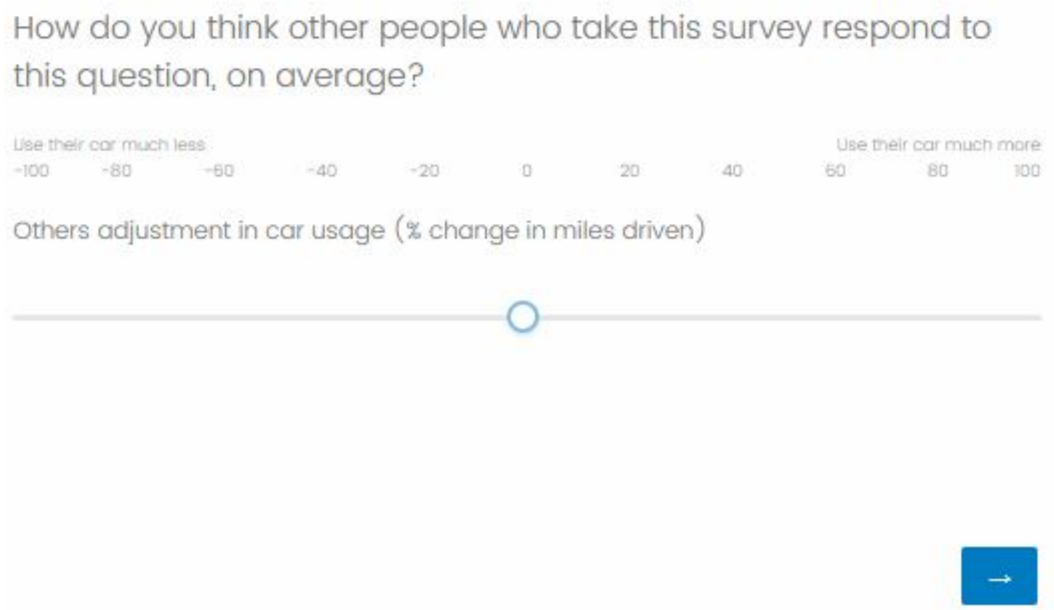
Figure 27: Social cost of carbon adjustment screen.



Suppose scientists now estimate that the social cost of carbon from driving gasoline-powered cars has increased by 30%, while prices remain the same. This means that each mile driven causes more harm. How would you adjust the use of your car?

Your own adjustment in car usage (% change in miles driven)

Figure 28: Social cost of carbon adjustment, hold-out version.



Policy Preferences: Taxation

Figure 29: Preferred tax level (UK baseline version).

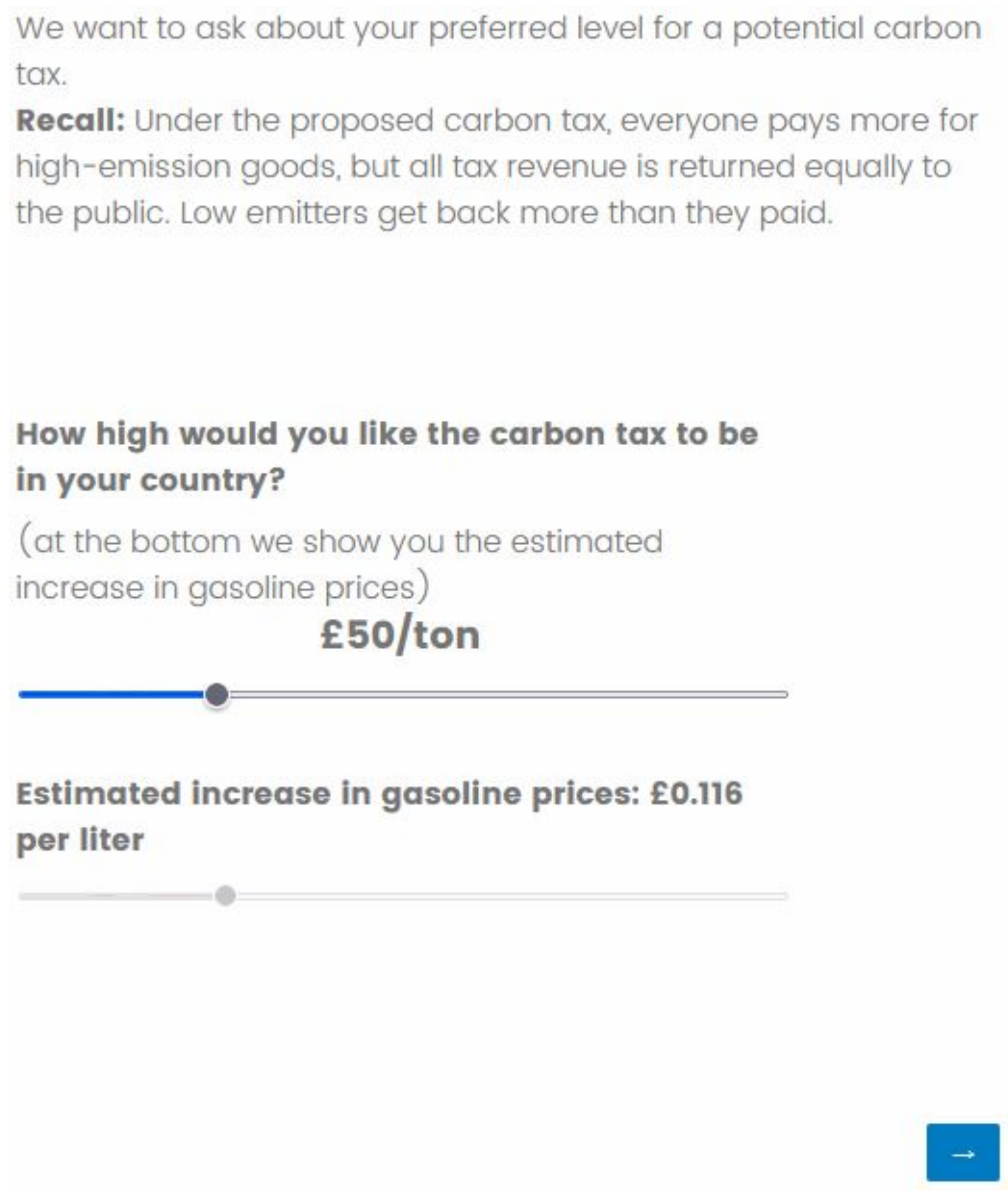


Figure 30: Preferred tax level, after the slider was moved.

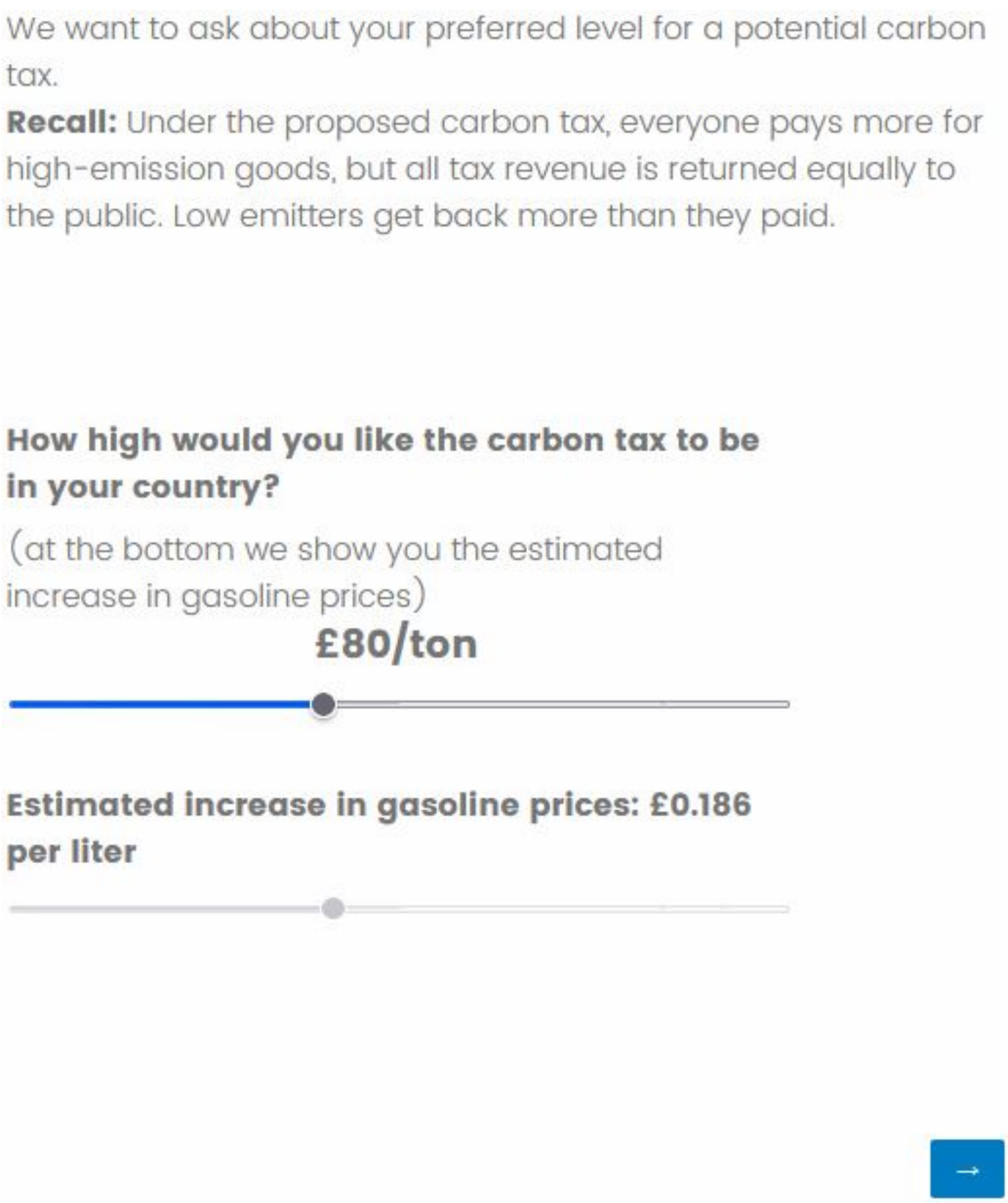
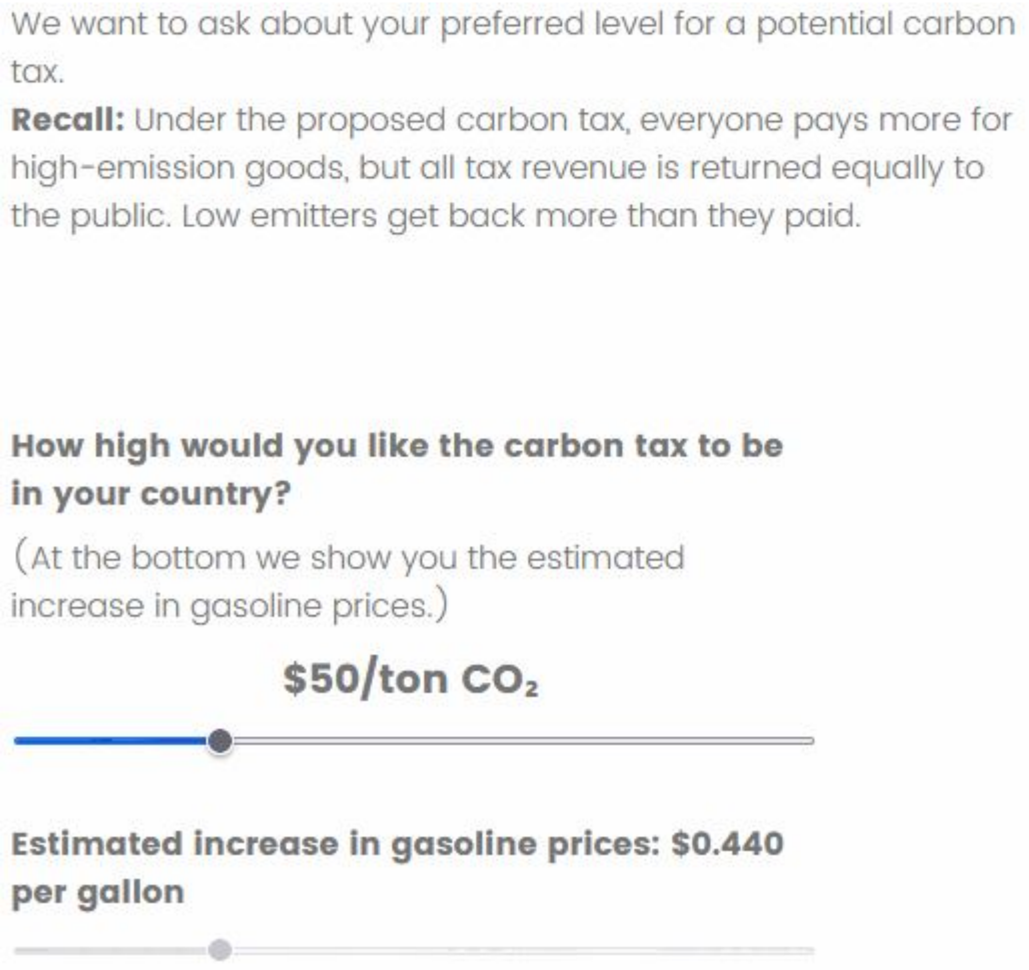


Figure 31: Preferred tax level (U.S. version). The preferred tax question was adapted to U.S. context (currency and reference points).



We want to ask about your preferred level for a potential carbon tax.
Recall: Under the proposed carbon tax, everyone pays more for high-emission goods, but all tax revenue is returned equally to the public. Low emitters get back more than they paid.

How high would you like the carbon tax to be in your country?

(At the bottom we show you the estimated increase in gasoline prices.)

$50/ton CO_2

Estimated increase in gasoline prices: $0.440 per gallon

Figure 32: Local tax based on SCC (UK version).


Now suppose that **only in the UK**, the **social cost of carbon** (the damages caused by climate change to people, the economy, and the environment **in the UK**) has **increased by 30%**.

You previously chose a carbon tax of £**50** per ton. How would you adjust **your preferred carbon tax rate in the UK**?

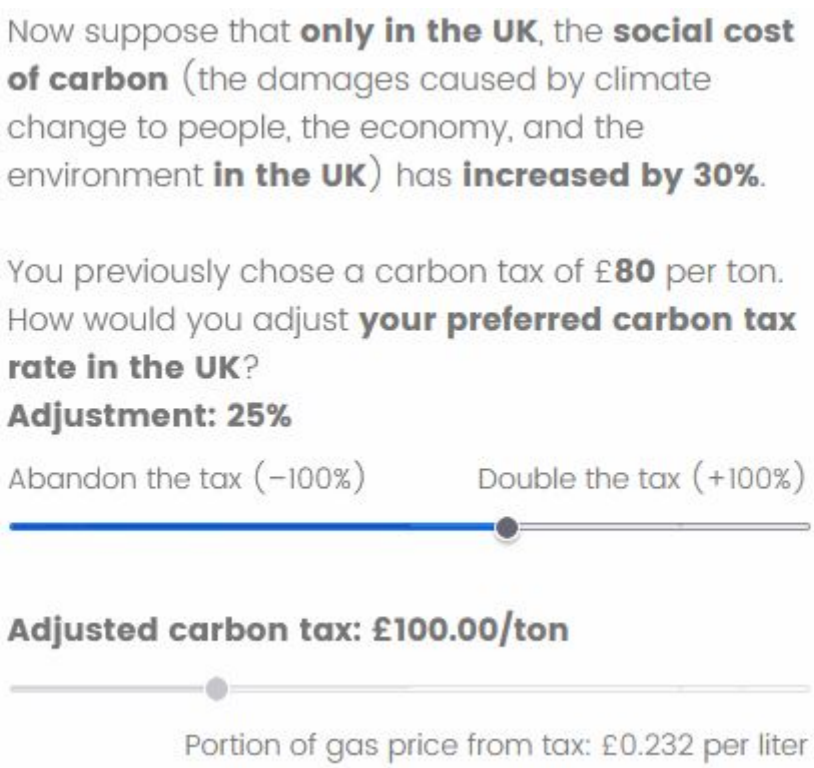
Adjustment: 0%

Abandon the tax (−100%) Double the tax (+100%)

Adjusted carbon tax: £50.00/ton

Portion of gas price from tax: £0.116 per liter

Figure 33: Local tax based on SCC, after the slider was moved.


Now suppose that **only in the UK**, the **social cost of carbon** (the damages caused by climate change to people, the economy, and the environment **in the UK**) has **increased by 30%**.

You previously chose a carbon tax of £**80** per ton. How would you adjust **your preferred carbon tax rate in the UK**?

Adjustment: 25%

Abandon the tax (−100%) Double the tax (+100%)

Adjusted carbon tax: £100.00/ton

Portion of gas price from tax: £0.232 per liter

Figure 34: Local tax based on SCC (U.S. version). The local SCC-based tax question was adapted to U.S. context.

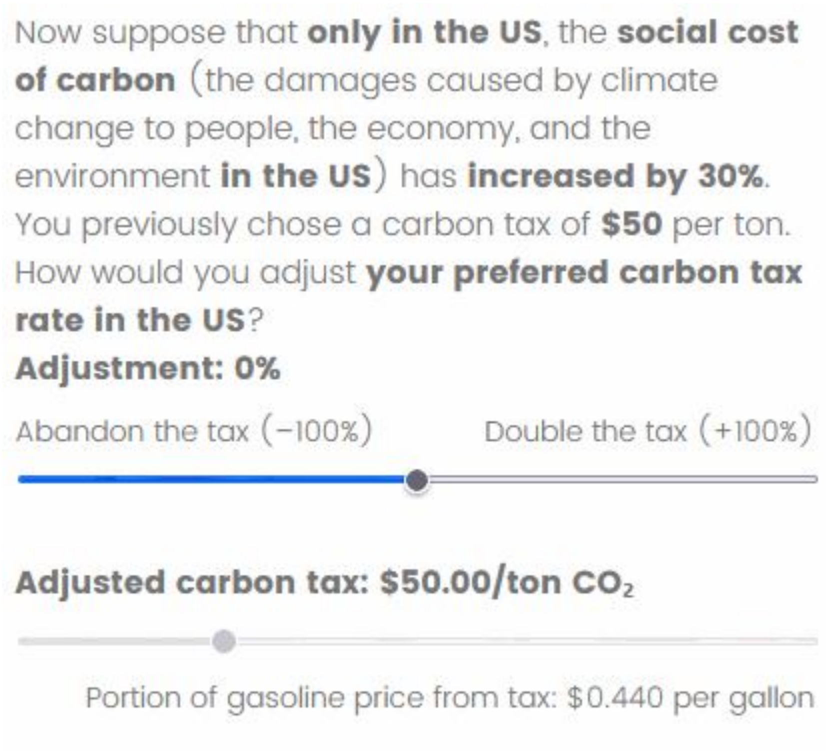


Figure 35: Local tax based on SCC, hold-out version.

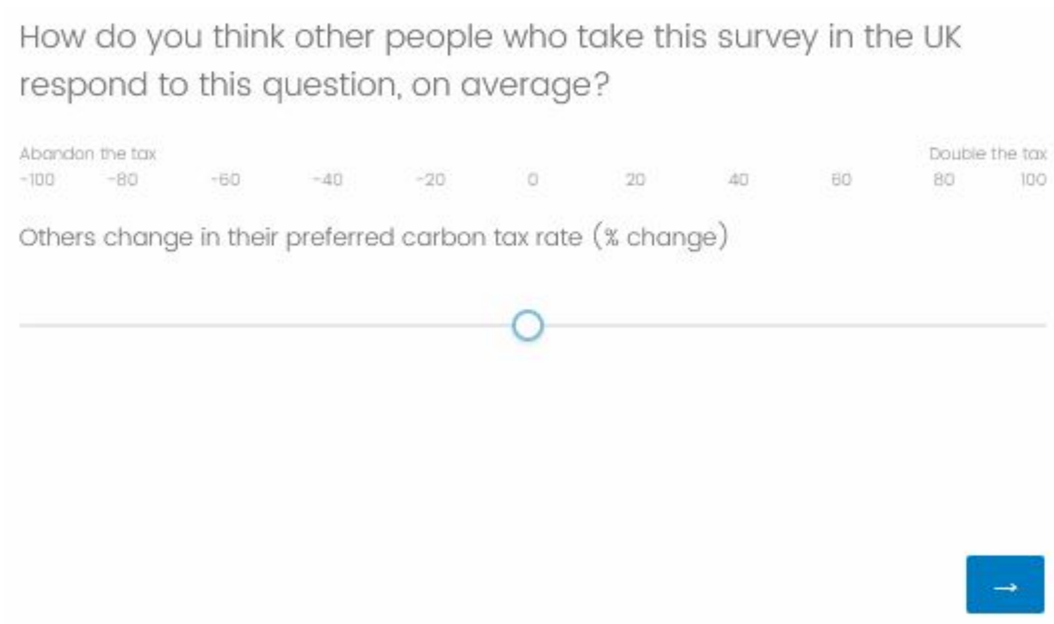


Figure 36: Global tax based on SCC (UK baseline version).

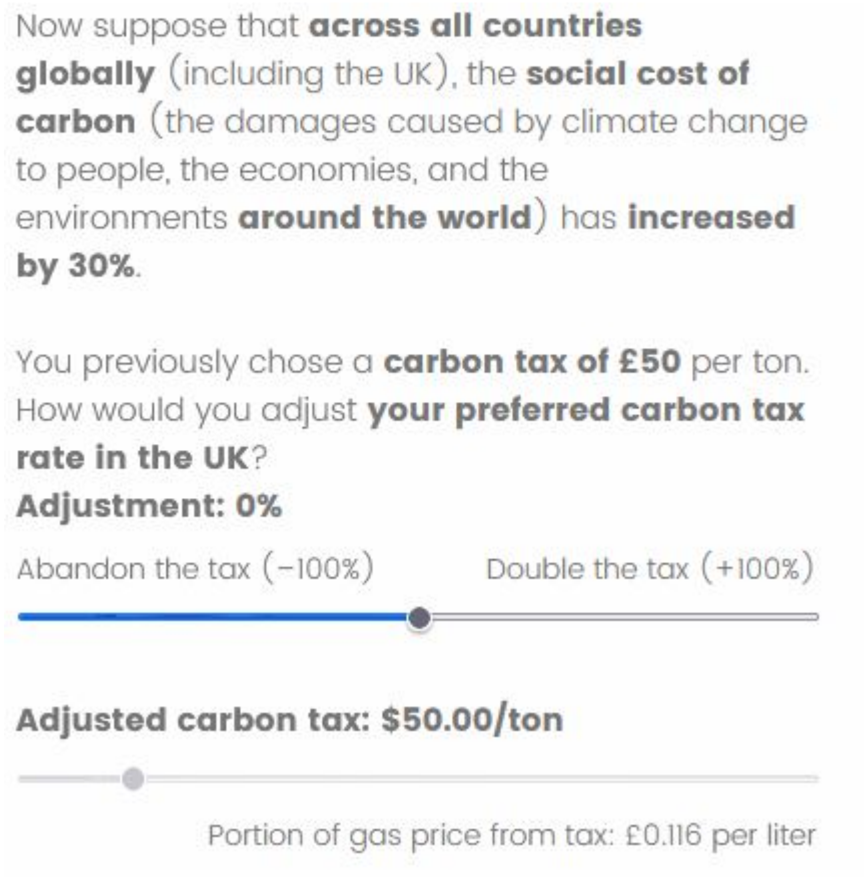


Figure 37: Global tax based on SCC, after the slider was moved.

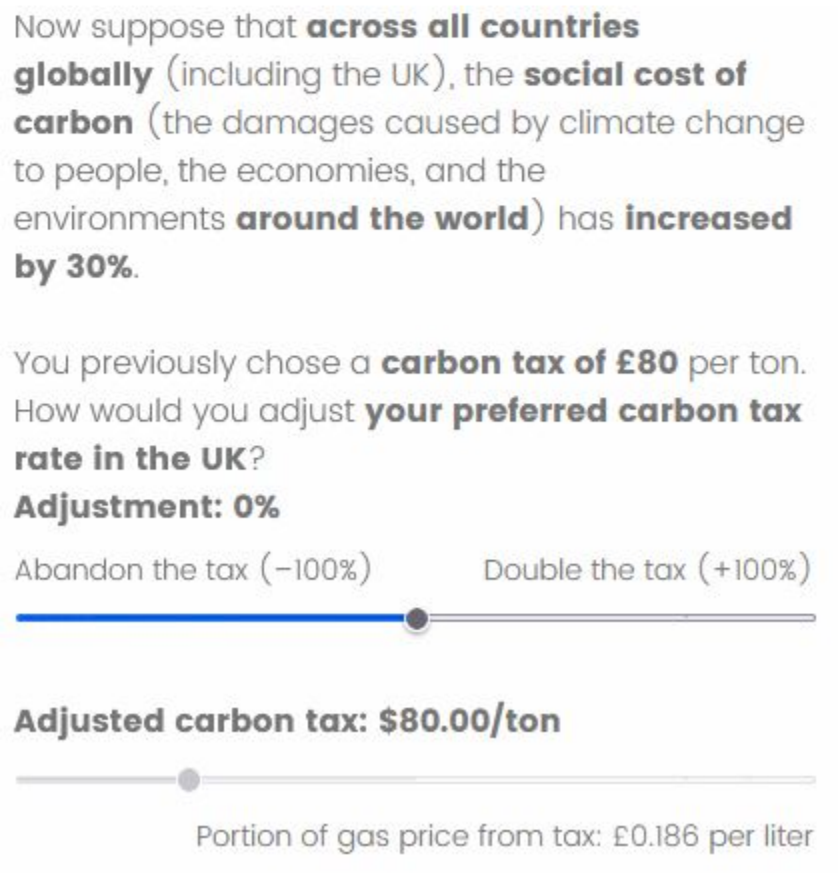


Figure 38: Global tax based on SCC (U.S. version). The global SCC-based tax question was adapted to U.S. context.

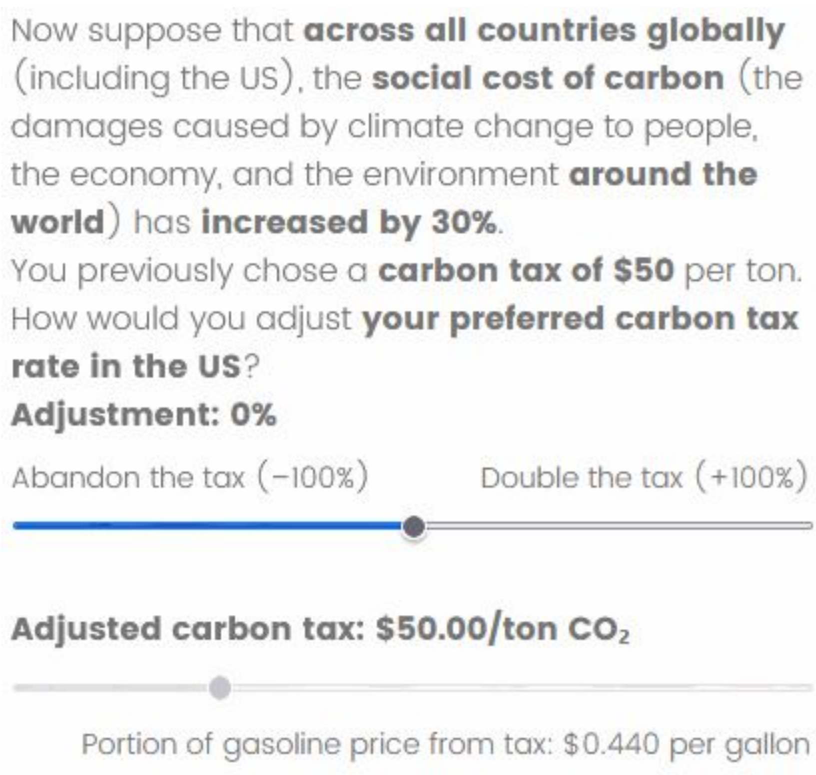
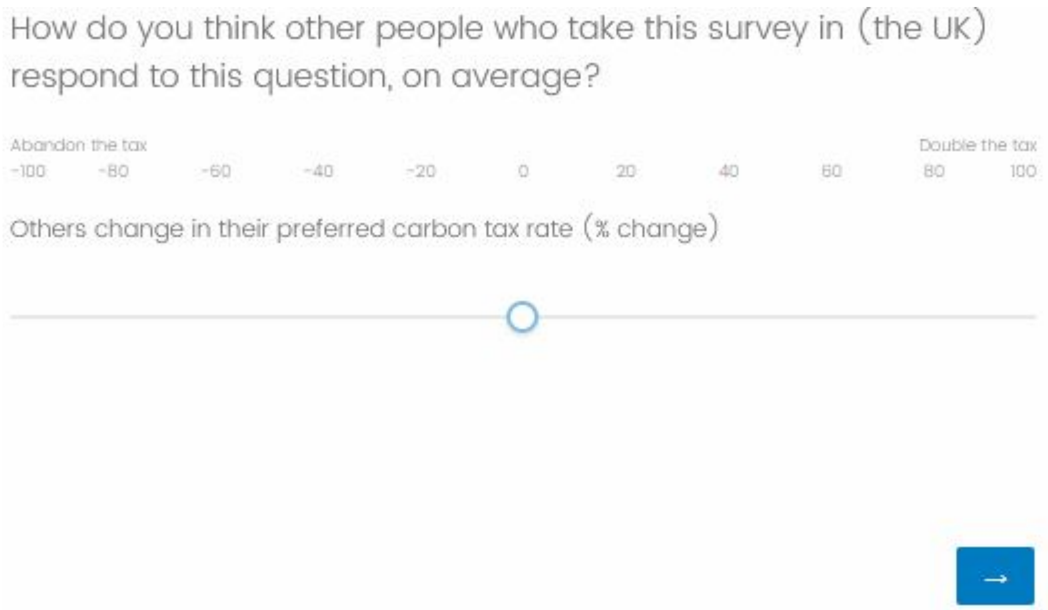
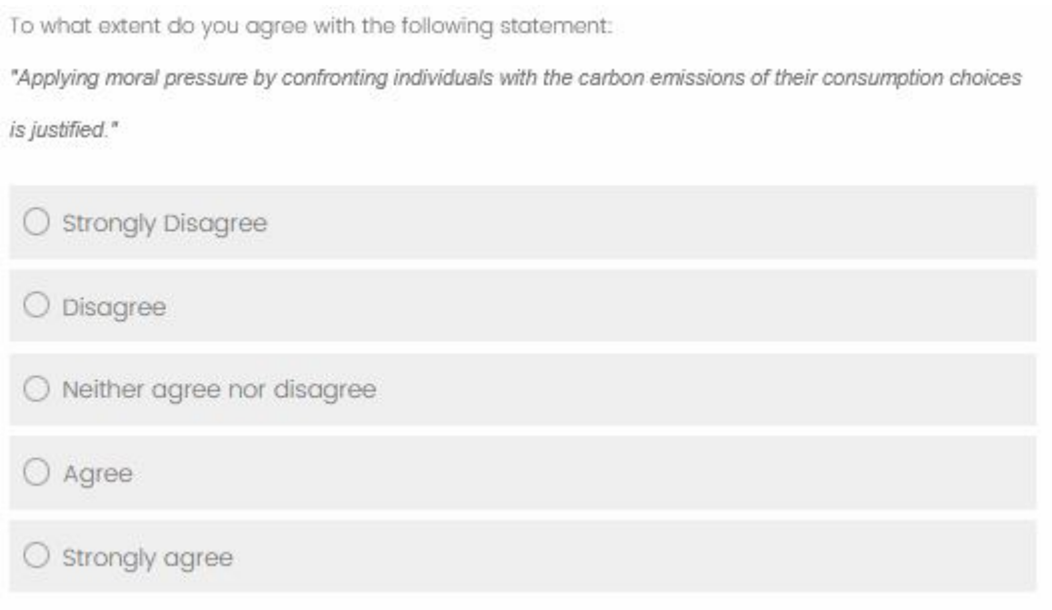


Figure 39: Global tax based on SCC, hold-out version.



Moral Pressure Block

Figure 40: Moral pressure, positive


To what extent do you agree with the following statement:

"Applying moral pressure by confronting individuals with the carbon emissions of their consumption choices is justified."

- Strongly Disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

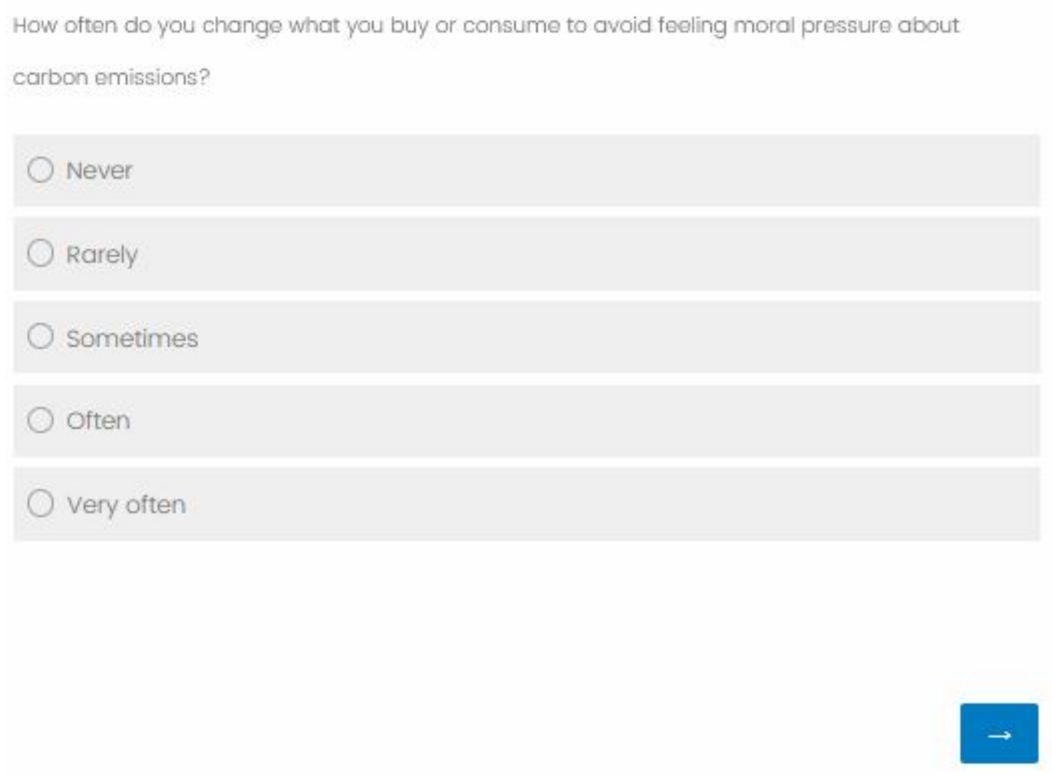
Figure 41: Moral pressure, negative


To what extent do you agree with the following statement:

"Moral pressure about individuals' carbon-intensive consumption is intrusive and should be avoided."

- Strongly Disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Figure 42: Response to Moral Pressure



How often do you change what you buy or consume to avoid feeling moral pressure about carbon emissions?

- Never
- Rarely
- Sometimes
- Often
- Very often

→

Behavioral Measures

Figure 43: Price versus Harm

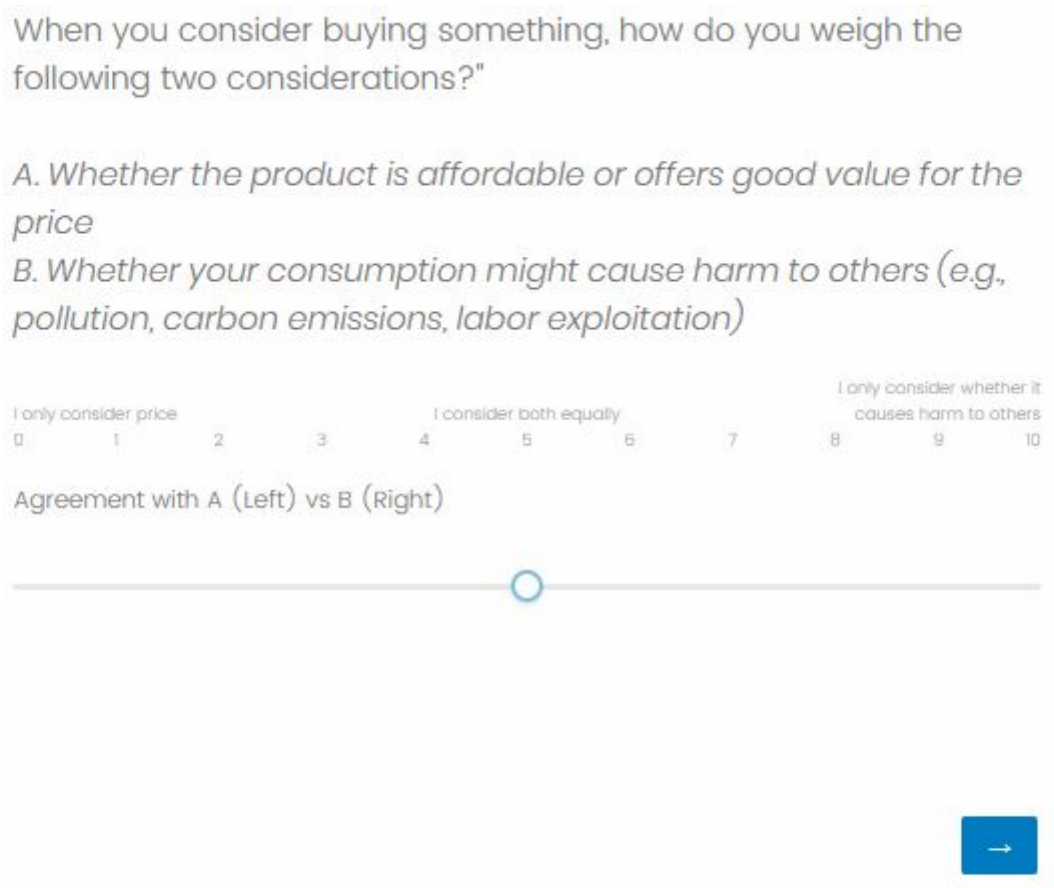
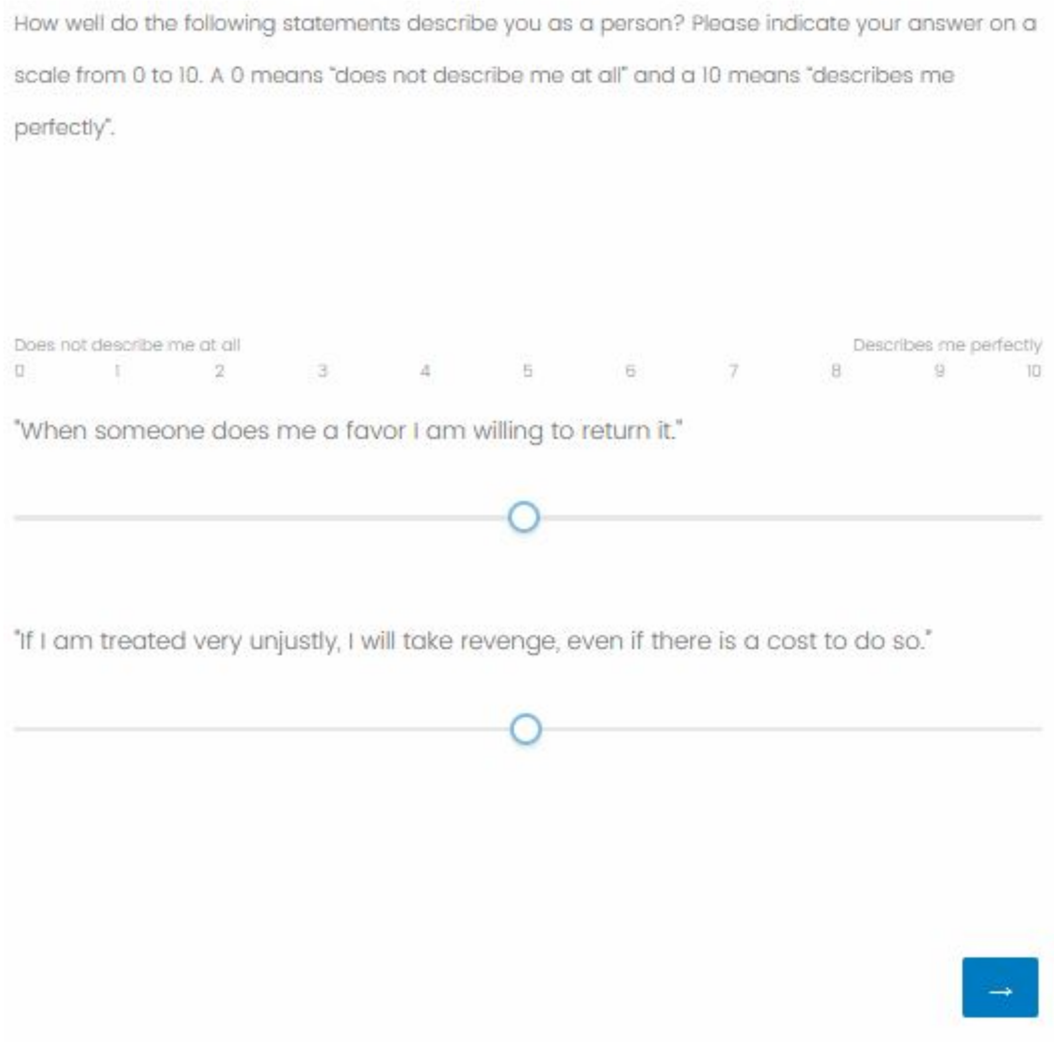


Figure 44: Reciprocity measure.


How well do the following statements describe you as a person? Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all" and a 10 means "describes me perfectly".

Does not describe me at all
Describes me perfectly
0 1 2 3 4 5 6 7 8 9 10

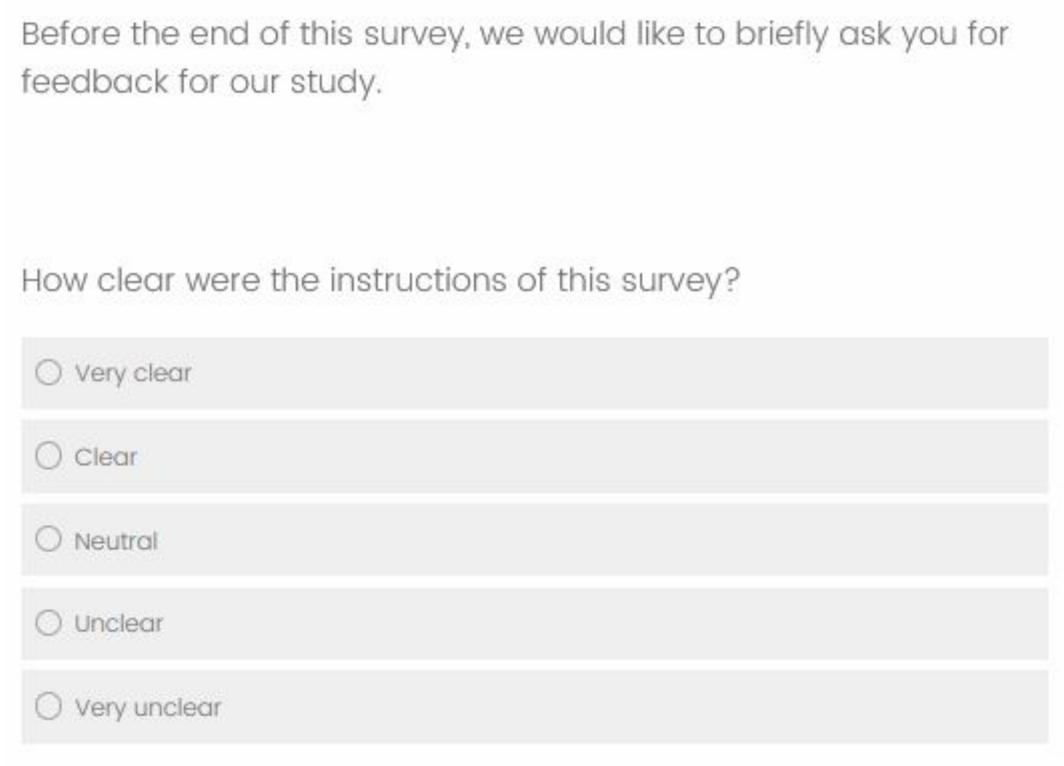
"When someone does me a favor I am willing to return it."

"If I am treated very unjustly, I will take revenge, even if there is a cost to do so."

→

Clarity

Figure 45: Clarity check.



Before the end of this survey, we would like to briefly ask you for feedback for our study.

How clear were the instructions of this survey?

- ◯ Very clear
- ◯ Clear
- ◯ Neutral
- ◯ Unclear
- ◯ Very unclear

Figure 46: Self-assessment screen.