The Side Effects of Green Soft Policies

Julien Picard

Department of Geography and Environment

London School of Economics

A thesis submitted to the Department of Geography and Environment
of the London School of Economics for the Degree of Doctor of Philosophy

London, 28/03/2024

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 23,800 words, excluding the bibliography.

Statement of co-authored work:

I confirm that Chapters 1, 2, and 5 are my own work.

I confirm that the data used for writing Chapter 3 was jointly collected with Dr. Sanchayan Banerjee. I contributed 95% of the work presented in this chapter.

I confirm that the data used for writing Chapter 4 was jointly collected with Dr. Anja Kobrich Leon and Dr. Janosch Schobin. I contributed 90% of the work presented in this chapter. A version of this work is currently being reviewed in an academic journal.

Acknowledgements

My first thanks go to my supervisors: Dr. Eugenie Dugoua, Dr. Marion Dumas and Prof. Susana Mourato. I am particularly indebted to Eugenie Dugoua, who supported me at every step of my five years at LSE. I am also very grateful to Marion Dumas. Together, you helped me think in a straightforward way, present ideas in a straightforward way, and write papers in a straightforward way. This was a lot of work.

I am also grateful to Susana Mourato, who sowed in my mind the idea that behavioural spillover effects were an interesting thing to look at. Your nudge at the beginning of the PhD significantly influenced the direction I took for the rest of it.

I thank my co-authors, Sanchayan Banerjee, Anja Kobrich Leon, Janosch Schobin, Ben Groom, and Ben Balmford. It is a pleasure to work with you. My special thanks go to Sanchayan Banerjee, who, on top of being a great colleague, is also a close friend.

I would also like to thank Anomitro Chatterjee, Elisabeth Gsottbauer, Matteo Galizzi, Stephen Jarvis, Simon Dietz, Sefi Roth, Charles Palmer, Ganga Shreedar, Aurelien Saussay, and Gregor Singer. Your advice has been extremely helpful.

I am also grateful to my colleagues at the Department of Geography and Environment and at

the Grantham Research Institute for their advice, moral support and friendship: Glenn Gostlow, Manuel Linsenmeier, Ghassane Benmir, Beatriz Jambrina Canseco, Pedro Llanos-Paredes, Line Relisieux, Romano Tarsia, Andrea Herrera Bórquez, Chiara Sotis, Lorenzo Sileci, Yang Zheng, Ina Drouven, Antonio Avila-Uribe, Louise Bernard, Capucine Riom, Emmanuel Awohouedji, Mook Bangalore, Ignaciao Aravena González, Tea Gamtkitsulashvili, Nikolaus Hastreiter, Vittoria Sottini, Sarah Elven, Jeff Pagel, Martin Brown Munene, Lukas Makovsky, Juan Alvarez-Vilanova, Violet Lasdun, Martina Pardy, Frida Timan, Melissa Weihmayer, Margarida Bandeira Morais.

I would also like to thank my parents and my sister. You have been incredibly supportive.

Last but not least, I could never thank my partner enough, Sarah Gharbi. We started our PhD at the same time. We coached each other, proofread each other, and supported each other through the best and the worst parts of it. My doctoral studies would not have been the same without you.

Abstract

We act pro-socially to make up for past wrongs, uphold our personal beliefs, get social approbation, or enjoy the warm-glow feeling of helping out. Pro-social soft policies tap into these motives to foster selfless deeds. Still, we know little about *how* soft policies change behaviour. In the introductory chapter of this dissertation, I endeavour to explain why understanding mechanisms is important. I also lay out my approach to studying them in this thesis. In Chapter 2, I create an economic model that rationalises "behavioural spillovers", i.e., the within-individual effect of doing a pro-social action on one's likelihood to do another. I show that pro-social policies weaken or amplify this spillover effect depending on the psychological mechanism through which they induce behaviour change. Thus, estimating such second-order effects can shed light on mechanisms. A key application of this theory lies in tackling global warming. In Chapter 3, I study if eating less meat — an individual action with high mitigation potential — induces us to do more for the environment. I also assess whether promoting vegetarian choices with social norm nudges amplifies or weakens this spillover effect. Using an online experiment (n=2775), I find that when the social norm succeeds in promoting vegetarianism, it is at the cost of crowding out this willingness to do more. This "crowding-out" effect suggests that social norm messaging induces people to act out of extrinsic motivations (e.g., to temper social pressure). Chapter 4 explores how two narratives used by politicians or environmental activists to promote environmental activism can foster or hinder further engagement. The first triggers guilt from not doing enough by stressing the negative consequences of inaction. The second triggers pride from doing the right thing by stressing the benefit of climate action. I test their effectiveness in a large survey experiment (n=10,670). None of these approaches work in promoting pro-environmental actions. Putting these results in perspective with Chapter 3, I draw some implications for the design of pro-environmental soft policies. I conclude this thesis by reflecting on my research practices in Chapter 5.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

When aggregated, our climate-friendly choices can yield significant mitigation gains. For instance, Van de Ven et al. (2018) estimate that behaviour change can amount to 14 to 25% of the European Union mitigation targets for 2050. Yet, our lifestyles are still far away from being environmentally sustainable. Making green choices the social norm is thus an important objective of policymakers. Economic theory provides a simple solution to induce such behavioural shifts. Negative externalities should be priced in the cost of goods through Pigouvian taxes or cap-and-trade markets. Once social costs are accounted for, "green" demand will naturally reach its social optimum.

However, such social optima are unlikely to be achieved through the sole use of market-based policies. Pigouvian taxes are unpopular and perceived as regressive (Douenne and Fabre, 2022). Their implementation might be met with a threshold beyond which any further increase will be politically infeasible. Cap-and-trade schemes suffer from their limitations, too. Any additional mitigation effort is compensated by a fall in carbon prices (Rosendahl, 2019). Theoretical and experimental evidence suggests this water-bed effect is counterproductive in a society with climate-conscious consumers (Herweg and Schmidt, 2022; Ockenfels et al., 2020), hampering the habit changes that are most needed. Finally, we are not always as sensitive to price signals as the "representative agent" of economic models (Grubb, 2014).

So, what about complementing market-based with soft policies? Individually, our actions to reduce greenhouse gas emissions might not significantly impact climate change. Yet, many of us continue to do our part. The motivations behind these actions are not always entirely pristine and disinterested (Sapolsky, 2018). We act pro-environmentally to boost our self-esteem, stay true to our beliefs, avoid disapproving glances, or get social approbation. Some behavioural policies tap into these motives to foster behaviour change (e.g., social comparisons, moral appeals, information provision). On their own, these soft policies will not solve climate change (Nisa et al., 2019).

Nonetheless, they can enhance the effectiveness of market-based instruments and bring us closer to the social optimum (Stern, 2020). But for that to happen, we need to understand how they work. Still, little is known about their mechanisms.

There are at least three reasons why this is important. First, a better grasp of the mechanisms of soft policies will improve their design, our understanding of the population segments on which they work, and how they interact with other traditional regulatory approaches. In other words, understanding mechanisms will enable us to use soft policies more efficiently. Second, behavioural policies are often criticised for being unethical (Bovens, 2009; Oliver, 2013). Therefore, generalising their use requires setting best practices. In this regard, encouraging behaviour change for reasons diverging from policymakers' motivations can be seen as manipulative. For example, leveraging social pressure to promote climate-friendly behaviours may lack transparency. Indeed, the reason *why* people change their behaviour in reaction to the policy (avoiding social pressure) is not the reason *why* the policymaker implemented the policy (mitigating emissions). Understanding why people change their behaviour after exposure to a policy might help decide on best practices. Third, understanding mechanisms is fundamental knowledge. In economic terms, whilst market-based approaches induce behaviour change through a shift in the budget constraint, soft policies do so through the utility function. Understanding how they work will help us better understand how people make choices.

How do we study soft policies' mechanisms? If one wants to know whether policies A and B work differently, comparing their effects on the behaviour they target is not informative. If both policies are effective, they should both have positive first-order effects. In this dissertation, I argue that studying policies' second-order effects can shed light on their mechanisms. By second-order effects, I consider the effect that soft policies may trigger on decisions that were not initially targeted.

In this regard, an expanding literature documents instances where soft policies yielded these unexpected side effects, so-called behavioural spillovers (Dolan and Galizzi, 2015). The starting point of this dissertation is to assume that if Policy A is so different from Policy B — i.e. if it induces people to change their behaviours for very different reasons — then their second-order effects will likely differ. For instance, if Policy A changes behaviour by raising people's awareness of climate change, there is more chance to observe positive effects on other non-targeted pro-environmental actions. Indeed, people may pay more attention to other aspects of their behaviour that they can change for the environment. This might not be the case if it fosters behaviour change through social pressure. Once the pressure disappears, one might feel tempted to slacken. There is also a practical interest in studying behavioural spillovers. We need more than a one-off change when promoting environmentally sustainable lifestyles. Knowing whether policies yield co-benefits will also improve their evaluation.

Experimenters must overcome two challenges to make inferences from second-order effects. The first challenge is *heterogeneity-related*. We cannot assume that we are all "wired" the same. One policy might induce behaviour change through different mechanisms for different people. This implies that the signs of the second-order effects of policies might differ from one person to another. Exploring this heterogeneity by merely interacting treatment dummies with social-demographic covariates is insufficient to understand what drives these differences. One must conceptualise the "latent parameters" underpinning heterogeneity and find proxies to measure them experimentally. The second challenge is *action-related*. The nature of the pro-environmental actions that policies promote might also shape second-order effects. Pro-environmental behaviours differ along a wide range of dimensions: their difficulty, their social desirability, their frequency, their novelty *et cetera*. Measuring the spillover effects of a policy on non-targeted actions by merely comparing a control

and a treatment group will not work. We need to disentangle the effect of the policy from the effect of the actions they promote.

My objective is to develop a theory and an identification strategy to study policies' second-order effects. In Chapter 2, I lay out the theoretical foundations for the rest of the dissertation. I develop a utility maximisation framework to map how different psychological mechanisms trigger different side effects on non-targeted decisions. I focus on pro-social decisions, i.e., actions yielding positive externalities on others. I model pro-social decisions as either intrinsically motivated (e.g., "I act pro-socially because it is who I am") or extrinsically motivated (e.g., "I have to act pro-socially if I don't want to be judged"). Soft pro-social policies either play on individuals' intrinsic or extrinsic motivations. This model reconciles several contradicting findings in the social psychology literature. It also provides micro-foundations to behavioural spillover effects, backfiring effects, and the heterogeneous effects of pro-social policies.

Importantly, the model highlights two channels through which policies alter other non-targeted pro-social decisions. The first channel is *indirect*. The policy induces a change in the targeted pro-social action, which, as a domino, affects the other non-targeted decision. When the targeted action is done out of intrinsic motivation (i.e., out of convictions), it is harder not to act pro-socially again. We do not like to be at odds with our convictions. When done out of extrinsic motivations (as a means to an end), the targeted action licences subsequent self-serving actions (e.g., "I already did my part"). The second channel is *direct*. Its sign captures the psychological mechanism through which policies induce behaviour change in the first place. When targeting intrinsic motivations, the policy reinforces consistency across pro-social deeds. When it targets extrinsic motivations, it weakens it. Disentangling the *indirect* from the *direct spillover* effect is key to addressing the *action-related* challenge.

In Chapter 3, I develop an experimental design to estimate these two channels causally. The design is structured in three parts. First, I randomly expose people to the policy of interest. Second, I offer them the opportunity to engage in the behaviour promoted by the policy. Third, respondents can make another non-related pro-social decision. The identification strategy relies on two sources of randomisation. The first is the policy: experimental subjects are randomly allocated between control and treatment groups. The second source of randomisation is a change in the choice architecture of the experiment. The objective is to unconsciously alter respondents' participation in the targeted behaviour — i.e. without altering their intrinsic or extrinsic motivations. Namely, deciding to act pro-socially is slightly more cumbersome for one group and slightly less for the other. I use this additional source of randomisation as an instrumental variable to get a causal effect of the *indirect spillover* effect. This enables me to disentangle it from the *direct spillover* effect.

With my coauthor, I apply this design in an online experiment (n=2775). We test if encouraging vegetarianism with a social norm nudge alters environmental donations. Cutting on meat is one of the most effective ways to reduce one's carbon footprint. Here, we are interested in whether stressing a rising trend of vegetarians increases people's likelihood of choosing a vegetarian dish in the experiment. By looking at the side effects of the prompt on green donations, we aim to answer three questions. First, does choosing vegetarian dishes causally affect respondents' willingness to do more for the environment, as proxied by the donation task? Second, does prompting vegetarian choices with the social norm nudge amplify or weaken this effect? Third, depending on whether the nudge amplified or weakened this effect, what can we say about the psychological mechanism through which it worked?

Reading the social norm message effectively increases intentions to choose vegetarian food. We also see a positive *indirect* effect: choosing vegetarian food increases donations. However, on av-

erage, the social norm nudge does not weaken nor reinforce this *indirect spillover* effect. In other words, we do not find evidence for a *direct spillover* effect. In an exploratory analysis, we address the *heterogeneity-related* challenge by using a machine learning algorithm trained on additional survey data. We predict respondents' inclination to follow the social norm. We categorise our sample into four groups: those that do not want to follow the norm (the *unwilling* group), those hesitating about following it (the *hesitant* group), those trying to follow the norm (the *trying* group), and those that are already following it (the *transitioned* group). We find that the *trying* group drives the effect of the nudge on vegetarian food choice intentions. We also find substantial heterogeneity behind the null *direct spillover* effect. Namely, we identify a negative and robust *direct effect* of the social norm nudge in the *trying* group. This negative *direct* effect outweighs the positive *indirect spillover* effect. These results indicate that when the social norm nudge successfully increases respondents' likelihood of choosing vegetarian food, it is at the cost of crowding out their willingness to do more. In other words, there is no free lunch.

One can expect three reactions to social norm messaging. The first is a belief-updating reaction, e.g., "Since many people are doing it, it must be good for the environment". In this case, there is no clear explanation for the negative direct effect we estimate. The second is an intrinsically motivated reaction: people feel emboldened by learning they are not the only ones doing their parts. In this case, the model of Chapter 2 predicts the social norm message to trigger a positive *direct effect*, reinforcing consistency across environmental decisions. This is not what we observed in the experiment. The last reaction is extrinsically motivated: the message induces people to act to get a contingent reward or avoid a contingent punishment. For instance, the message can make people feel they are not part of the group if they do not follow the norm. In this case, the model predicts a negative *direct effect*: individuals feel licensed to do more once the pressure is released

by doing the targeted action. Results from Chapter 3 suggest social norm nudges enhance extrinsic motivations.

In Chapter 4, I employ the same methodology but expand the study scope along three dimensions. The first dimension is to test more policies. I seek to compare two moral appeals: a "doom-and-gloom" and a "win-win" approach. The "doom-and-gloom" appeal stresses the costs of inaction against climate change. It is fine-tuned in pilot studies to trigger negative feelings, such as guilt from not doing enough. The "win-win" appeal stresses the benefits of taking action against climate change. It seeks to embolden people to do their bit for the environment. It is fine-tuned in pilot studies to trigger positive feelings, such as pride from being pro-environmental. Both arguments are used in narratives developed by politicians or environmental activists to foster climate action. I seek to test which arguments improve neutral information provision.

The second dimension is to measure actual behaviours. Conclusions from the experiment presented in Chapter 3 are limited because food choices are intentional. In Chapter 4, the action targeted by the policies is consequential. It is a real effort task meant to help the research team develop an algorithm to assess the carbon footprint of our food choices. As in Chapter 3, the non-targeted decision is consequential, too: we ask participants to sign a pro-environmental petition.

The third dimension is to vary the characteristics of the targeted and the non-targeted decisions. Namely, I seek to test whether the difficulty of the targeted decision, the real effort task, moderates the *indirect spillover* effect. I also seek to test whether making the petition about a health-related cause instead of an environment-related cause changes the sign and magnitude of the *direct* and the *indirect spillover* effects.

My priors were that making people act out of guilt amounts to playing on their extrinsic motivations. One does not behave pro-environmentally because one cares about the environment but because one wants to temper the discomfort from feeling guilty. I expected to observe the "doom-and-gloom" narrative to trigger negative *direct spillover* effects. I also anticipated this effect to be heterogeneous in people's propensity to feel guilty. On the other hand, making people act out of pride for doing the right thing would induce an intrinsically motivated reaction. I assumed that triggering pride would enhance one's pro-environmental identity, making it harder not to act pro-environmentally again. As such, I expected the "win-win" narrative to yield positive *direct spillovers*. I posited this effect to be heterogeneous depending on people's pro-environmental attitude.

With my coauthors, I address these questions in an online experiment with 10,670 German respondents. The results were unexpected. Providing neutral information does not increase the uptake of the real-effort task. Stressing costs or benefits does not improve neutral information either. Consequently, we do not observe any *direct spillover* effects of win-win or doom-and-gloom arguments on the environment-related petition. We explore heterogeneity using scales to measure guilt-proneness and pro-environmental and altruistic attitudes. These null results do not hide heterogeneity. Furthermore, there is no evidence of an *indirect spillover effect*: doing the real effort task does not increase respondents' likelihood to sign the petition. Varying the difficulty of the real effort task does not moderate this effect. Finally, we find suggestive evidence for an *direct spillover* effect of the doom-and-gloom treatment on the likelihood of signing the health-related petition. However, this result does not pass multiple hypothesis correction.

Null results can be as informative as statistically significant ones. When put in perspective with the experiment of Chapter 3, I see three potential explanations for why we do not find any effect

of moral appeals on the uptake of the real effort task. First, the most obvious explanation is that the real effort task is consequential, whilst food choices are intentional. Respondents have to bear an opportunity cost when doing the real effort task as they are paid the same regardless of the time spent doing the survey. Therefore, the interventions may not be strong enough to induce behaviour changes.

The second explanation regards the phrasing of the interventions tested. Contrary to social norm messaging, doom-and-gloom and win-win appeals are "top-down": we state what respondents should do. This "top-down" nature may not be as effective as emphasising what others do through social norm messaging. Subjects may weigh more the information retrieved from observing their peers than from a figure of authority.

Third, it could be that soft policies only spur pro-environmental decisions by playing on factors that are orthogonal to environmental impact, such as warm-glow, self-esteem or social recognition. In contrast to food choices, the real effort task was new to respondents. For instance, they did not know if it was social desirability or the social norms associated with it. In other words, there were no other "rewards" from doing it than merely knowing its environmental impact, so there were no "ropes" the moral appeals could pull.

Overall, this dissertation presents a theoretical and an empirical framework to study the side effects of soft policies. It also presents two large survey experiments applying these frameworks in different contexts. I endeavour to be as transparent as possible on the way data was collected and the thought process that led me to investigate these questions. When presenting these experiments, I provide links to the survey material, the pre and post-analysis plans for Chapter 3 and the registered report for Chapter 4. The R codes and the data sets are available upon request and will

be available online once the experiments are published in the following shared folder. In Chapter 5, I conclude this thesis by reflecting on my research practices.

Bibliography

Bovens, L. (2009). The ethics of nudge. In *Preference change*, pages 207–219. Springer.

Dolan, P. and Galizzi, M. M. (2015). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16.

Douenne, T. and Fabre, A. (2022). Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1):81–110.

Grubb, M. (2014). *Planetary economics: energy, climate change and the three domains of sustainable development*. Routledge.

Herweg, F. and Schmidt, K. M. (2022). How to regulate carbon emissions with climate-conscious consumers. *The Economic Journal*, 132(648):2992–3019.

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., and Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature communications*, 10(1):1–13.

Ockenfels, A., Werner, P., and Edenhofer, O. (2020). Pricing externalities and moral behaviour. *Nature Sustainability*, 3(10):872–877.

Oliver, A. (2013). From nudging to budging: using behavioural economics to inform public sector policy. *Journal of Social Policy*, 42(4):685–700.

Rosendahl, K. E. (2019). Eu ets and the waterbed effect. *Nature Climate Change*, 9(10):734–735.

Sapolsky, R. M. (2018). *Behave: The biology of humans at our best and worst*. Penguin.

Stern, P. C. (2020). A reexamination on how behavioral interventions can promote household action to limit climate change. *Nature communications*, 11(1):1–3.

Van de Ven, D.-J., González-Eguino, M., and Arto, I. (2018). The potential of behavioural change for climate change mitigation: a case study for the european union. *Mitigation and adaptation strategies for global change*, 23:853–886.

Chapter 2

A Model of Pro-Social Policies

I Introduction

We are not pure hedonists. On many occasions, we forego immediate pleasure for the benefit of others. Various motivations underpin such pro-social deeds. We act pro-socially because it is our nature, to get social approbation, to return a favour, or to obtain a tax rebate. Policies fostering pro-social actions tap into these various motivations through communication campaigns or reward schemes. However, a growing literature indicates that our past actions *and the motivations underpinning them* affect our propensity to act pro-socially again. As such, is there a risk that, by altering these motives, pro-social policies yield unintended consequences?

Economic models rationalising selfless decisions do not provide an answer to this question. For Andreoni (1990), pro-social actions are motivated by a warm-glow feeling. For Akerlof and Kranton (2000), individuals' pro-social identity, shaped by their social networks, induces pro-social deeds. These models explain why selfless actions are more common than what classic economic models predict. Yet, they do not distinguish different motivations for acting pro-socially. This gap is filled by Bénabou and Tirole (2006), showing how the fear of appearing greedy can render incentives to act pro-socially counterproductive. But here again, this static model does not capture the influence of past pro-social actions on people's likelihood to do another. Bénabou and Tirole (2011) explain these dynamics by the desire to strengthen one's pro-social identity. In their model, a weakly held identity induces consistency: we engage in another pro-social action to confirm a pro-social nature. On the other hand, an already strong identity allows people to indulge in self-serving actions after a first pro-social act: the first action proves our virtue and licenses us to do more. Bénabou and Tirole (2011)'s results rely on the assumption that individuals imperfectly remember their pro-social identity. This assumption is relaxed by Ulph et al. (2023), showing how compensatory and consistent behaviours can arise in a model where individuals manage their

"stock" of self-worth. However, these two approaches ignore the fact that pro-social actions can be done for different reasons, implying that different policies can yield different unintended consequences.

This paper fills this gap with a simple utility maximisation framework. I assume that individuals act pro-socially out of intrinsic motivations (because this is "who they are") or out of extrinsic motivations (as a means to an end). At each period, the strength of individuals' pro-social identity is affected by their past pro-social actions and the motives underpinning them. Acting pro-socially out of intrinsic motivation raises the cost of reverting to self-serving behaviours. For instance, helping someone in need or serving a cause due to conviction may encourage us to do more to stay true to our beliefs. Acting pro-socially out of extrinsic motivation reduces the cost of reverting to self-serving behaviours. For instance, helping someone in need out of social pressure can cause us to slacken afterwards as the initial action was not self-driven. In the model, policies foster pro-social actions by targeting intrinsic or extrinsic motivations.

I make three contributions to the literature modelling the effect of pro-social policies. First, this model allows me to rationalise the existence of "behavioural spillover" effects: the effect of a first pro-social action on our propensity to do another (Thøgersen, 1999). In doing so, I reconcile in the same framework competing theories in social psychology, either predicting that a first deed reinforces the need to do another or that it licenses subsequent selfish deeds. I review these theories in section II. In the model, the motivations underpinning the first pro-social act determine which effects prevail.

Second, I show that pro-social policies can yield partial or net backfiring effects on targeted and non-targeted pro-social decisions. When individuals have self-esteem, i.e., when they care about

how their decisions affect their pro-social identity, policies increasing extrinsic motivations are counter-productive. Indeed, acting pro-socially out of greed weakens individuals' pro-social identity. On the other hand, pro-social policies that strengthen intrinsic motivations are more effective as they heighten their pro-social identity. Policies also influence non-targeted pro-social decisions. Policies playing on intrinsic motivations strengthen consistency. On the other hand, policies playing on extrinsic motivations weaken the need to act pro-socially after a first pro-social deed.

Finally, this model also rationalises heterogeneity in the effect of pro-social policies. Individuals' past experiences drive heterogeneity. Past experiences are captured by individuals' endowed pro-social identity in the model. Pro-social policies playing on extrinsic motivations are more effective on individuals holding a weak pro-social identity. Conversely, targeting intrinsic motivations is more effective for individuals with a strong pro-social identity.

This paper is organised as follows. In section II, I review the main psychological mechanisms this model rationalises and derive stylised facts. In section III, I present the model and the main results and develop recommendations for experimental scientists to use the model's insights. Section IV concludes.

II Psychological Foundations and Stylised Facts

II.A Psychological Foundations

A growing literature presents evidence of *behavioural spillover* effects, whereby a first action influences another (Carrico et al., 2018; Maki et al., 2019).[1] For instance, Comin and Rode (2023) find causal evidence that adopting solar panels increases support for the Green Party in Ger-

[1]For literature reviews on behavioural spillovers, the reader should refer to Dolan and Galizzi (2015) and Truelove et al. (2014).

many. Conversely, Mazar and Zhong (2010) find causal evidence suggesting that buying green products increases self-serving and immoral behaviours. Several theories in psychology explain these spillover effects.

Cognitive dissonance theory posits that we experience discomfort from behaving at odds with our past actions, beliefs, and values (e.g., Festinger 1962; Elliot and Devine 1994). In other words, we prefer to behave consistently. Similarly, *social identity* theory, developed by Tajfel et al. (1979), predicts that we are more likely to act according to our social identity after a first action increases its salience. Both theories rationalise the existence of positive behavioural spillovers. They make similar predictions: we seek to stay consistent with an identity we identify with.

Conversely, *moral licensing* describes feeling freed from engaging in another pro-social act after a first one (e.g., Monin and Miller 2001; Effron et al. 2009). Moral licensing is often invoked to rationalise negative behavioural spillover effects. *Moral cleansing* captures the opposite, i.e., feeling morally obliged to act pro-socially after failing to do a first pro-social act (e.g., Lee and Schwarz 2010). In the same vein, *conscience accounting* describes people behaving immorally when knowing they will have an opportunity to act pro-socially later (e.g., Gneezy et al. 2014). These findings explain compensatory behaviours: we tend to ease off after a first pro-social act or redeem ourselves after a selfish one.

When are we more likely to exhibit consistent or compensatory behaviours? In the next section, I argue that the motivations underpinning our actions and the context in which we make them play an important role.

II.B Stylised Facts

The motives for acting pro-socially now are likely to influence subsequent pro-social deeds. For simplicity, I distinguish two main motives: intrinsic and extrinsic. Intrinsically motivated decisions are made as an "end to themselves". In contrast, extrinsically motivated decisions are made as a "means to an end". In their meta-analysis, Maki et al. (2019) show that policies targeting intrinsic motivations are more likely to induce positive spillover effects.

> STYLISED FACT 1: Policies playing on intrinsic motivations are more likely to induce consistent pro-social behaviours.

Promoting pro-social actions by appealing to our values and beliefs is more likely to induce intrinsically motivated decisions (e.g., I sort my waste because I am environmentally friendly). For instance, Evans et al. (2013) show that emphasising self-transcending reasons to perform a pro-environmental behaviour leads to positive spillover effects contrary to emphasising monetary gains.

In the same vein, casting decision-makers with a given identity is likely to trigger consistency. Baca-Motes et al. (2013) find that giving hotel guests pins "Friends of the Earth" after they accepted to reuse their towels leads to more efficient use of lighting. Lacasse (2016) shows that labelling people as "environmentalists" increases their support for environmental policies in their neighbourhoods. Gneezy et al. (2012) find that consistent behaviours are more likely when the first one signals a pro-environmental identity.

Emphasising rules of conduct can also induce consistent behaviours. For instance, Cornelissen

et al. (2013) find that making people think with a deontologist mindset leads them to give more to charities and cheat less after remembering a good deed than a control group.[2]

On the other hand, the meta-analyses of Deci et al. (1999) and Cameron and Pierce (1994) suggest that rewards that are contingent on acting pro-socially undermine people's motivations to do another pro-social deed once the reward is removed.

> STYLISED FACT 2: Policies playing on extrinsic motivations are more likely to induce compensatory pro-social behaviours.

Incentives can be of different natures. Financial incentives, for instance, can induce respondents to slacken after exercising a certain level of effort (e.g., Dolan and Galizzi 2014; Xu et al. 2018; Steinhorst and Matthies 2016). Dolan and Galizzi (2014) find that monetary rewards for exercising have to be high enough to increase subsequent unhealthy eating behaviours. This suggests that monetary incentives only induce negative spillovers above a certain threshold.

Incentives can also be immaterial. For instance, one can undertake a behaviour to get social approbation. Tiefenbeck et al. (2013)'s study suggests that giving people feedback on their water consumption by comparing it to that of their most efficient neighbours led people to decrease water use but also increase energy use. Similarly, Kristofferson et al. (2014) shows that publicising a first altruistic action leads to a decrease in subsequent altruistic deeds. Finally, the meta-analysis of Maki et al. (2019) suggests that interventions playing on guilt yield negative spillover effects.

[2]Deontologists judge an action as ethical if it does not contradict their values and principles. For instance, consider the hypothetical scenario of a trolley running at full speed, without brake, that threatens to kill five people. The only alternative is to hit a switch, putting the trolley on new tracks where it would only kill one person. Deontologists judge the status quo as the only acceptable solution, as hitting the switch would imply actively killing someone.

Different types of pro-social policies can, therefore, induce different effects on non-targeted decisions. In the next section, I model these side effects based on the stylised fact highlighted in this part.

III The Model

III.A Defining the Utility Function

I consider an individual who chooses between allocating time or effort to acting pro-socially or selfishly. Intrinsic and extrinsic motives explain pro-social decisions. Intrinsically motivated decisions are undertaken for their own sake (e.g., "*I act* altruistically because it is the right thing to do"). In contrast, extrinsically motivated decisions are made as a means to an end (e.g., "*I need* to show that I am altruist, so *I have* to do a good deed"). Stylised facts 1 and 2 imply that past pro-social actions influence current ones. Acting out of intrinsic motivation reinforces the need to act pro-socially again (e.g., "I don't like being at odds with my past commitments"). Acting out of extrinsic motivations reduces this need (e.g., "I have already done my bit").

To capture these dynamics, I consider a simple utility framework where individuals' decisions to act pro-socially depend on two factors: the context in which they make their decisions and their pro-social identity. The context influences individuals' intrinsic or extrinsic motivations to act pro-socially. The strength of individuals' pro-social identity alters the utility derived from pro-social and selfish deeds, as in Akerlof and Kranton (2000). The peculiarity of my model lies in the fact that identity is shaped by past choices and the motives underpinning them.

I consider a three-period framework. In period zero, individuals start with an "endowed" pro-social identity. This endowment can be seen as individuals' past experiences before an external

observer (i.e., the experimenter) scrutinises their choices. In period one, individuals decide how much time or effort to allocate to pro-social or pro-self activities. This decision is influenced by their endowed identity and the context, which determines their motivations to act pro-socially. In the next section, I show that policies that encourage selfless deeds modify the context to increase individuals' motivations. Decisions in period one subsequently alter how individuals perceive themselves: they "update" their identity based on what they did and why. In period two, the situation repeats. Individuals allocate their time or efforts to pro-social and pro-self activities, given their newly updated identity and the context of period two. I assume the following utility function for periods one and two:

$$V_1 \equiv v(x_1, y_1 | I_0, \eta_1, \kappa_1) \qquad V_2 \equiv v(x_2, y_2 | I_1, \eta_2, \kappa_2) \tag{III.1}$$

The choice variable x_t captures the time and effort allocated to pro-social activities in period $t \in \{1, 2\}$. Conversely, y_t corresponds to the time or effort allocated to pro-self activities. Parameters η_t and κ_t denote the strength of intrinsic and extrinsic motives for acting pro-socially. The context of the decision-making determines these two parameters. When pro-social policies change the context, they alter these parameters. Parameter I_0 is given. It corresponds to individuals' endowed pro-social identity (i.e., past experiences). On the other hand, I_1 is a function such that $I_1 : \{x_1, \eta_1, \kappa_1, I_0\} \mapsto \mathbb{R}$. I_0 and I_1 can be seen as sources of intrinsic motivation, which, contrary to η_1 or η_2, are determined by respondents' past experiences. I drop the subscripts in what follows when it does not affect understanding. I make the following assumptions:

ASSUMPTION 1: *Utility function V is twice continuously differentiable, increasing and concave in x and y: $\partial_x V > 0$. $\partial_y V > 0$, $\partial_{xx} V \leq 0$, and $\partial_{yy} V \leq 0$.*

Assumption 1 ensures the utility function is "well-behaved": individuals derive utility from acting pro-socially and selfishly.

ASSUMPTION 2: *The cross-derivative of utility function V between x and $\theta \in \{\eta, \kappa\}$ is positive. The cross-derivative between y and θ is negative: $\partial_{\theta x} V \geq 0$ and $\partial_{\theta y} V \leq 0$.*

Assumption 2 implies that higher intrinsic or extrinsic motives increase the marginal utility of acting pro-socially and decrease that of acting selfishly.

ASSUMPTION 3: *Function I_1 is increasing in η_1 and I_0, and decreasing in κ_1. It is strictly increasing in x_1 if and only if $\eta_1 + I_0 > \kappa_1$. The cross-derivatives of I_1 between x_1 and η_1 and between x_1 and I_0 are positive. The cross-derivative of I_1 between x_1 and κ_1 is negative: $\partial_{\eta_1} I_1 > 0$, $\partial_{I_0} I_1 > 0$, $\partial_{\kappa_1} I_1 < 0$, $\partial_{\eta_1 x_1} I_1 > 0$, $\partial_{I_0 x_1} I_1 > 0$, $\partial_{\kappa_1 x_1} I_1 < 0$, and $\partial_{x_1} I_1 > 0 \Leftrightarrow \eta_1 + I_0 > \kappa_1$.*

Assumptions 3 imply that remembering higher intrinsic motivations (η_1) or a stronger past pro-social identity (I_0) strengthen pro-social identity I_1. Conversely, remembering that one was extrinsically motivated to act pro-socially (κ_1) weakens pro-social identity I_1. The signal sent by acting pro-socially in period 1 also depends on the motivations driving the pro-social action. When intrinsic motivations dominate extrinsic motivations ($\eta_1 + I_0 > \kappa_1$), acting pro-socially reinforces pro-social identity I_1. On the other hand, if extrinsic motivations are stronger than intrinsic motivations ($\kappa_1 > \eta_1 + I_0$), then acting pro-socially sends a negative signal about one's pro-social identity. Assumptions on the cross-derivative capture the idea that stronger intrinsic (extrinsic) motives reinforce (weaken) the marginal return of acting pro-socially.

ASSUMPTION 4: *Utility function V is increasing and concave in I with positive (negative) cross-derivatives between I and x (y): $\partial_I V \geq 0$, $\partial_{II} V \leq 0$, $\partial_{Ix} V \geq 0$, and $\partial_{Iy} V \leq 0$.*

Assumptions 4 imply that individuals derive utility from feeling pro-social ($\partial_I V \geq 0$) and from staying consistent with this identity ($\partial_{Ix} V \geq 0$, $\partial_{Iy} V \leq 0$).

At each period $t = 1, 2$, individuals have a time or effort budget B_t that they allocate between performing pro-social activity x_t or performing self-serving activity y_t. The budget constraint is of the form:

$$B_t = x_t + y_t \tag{III.2}$$

I consider two cases. Decision-makers can be either shortsighted – they only consider their utility of the current period when making a choice – or they can be farsighted. In this case, they anticipate period two. In the shortsighted case, individuals maximise:

$$V_1^s \equiv v(x_1, B_1 - x_1 | I_0, \eta_1, \kappa_1) \tag{III.3}$$

In the second case, they maximise:

$$V_1^f \equiv v(x_1, B_1 - x_1 | I_0, \eta_1, \kappa_1) + \beta \cdot v(x_2, B_2 - x_2 | I_1, \eta_2, \kappa_2) \tag{III.4}$$

Without loss of generality, I set the discount factor β to 1. Finally, I also make the following assumption:

ASSUMPTION 5: *The hessians of maximisation problems* (III.3) *and* (III.4) *are negative definite.*

Assumption 5 ensures that acting pro-socially in period 1 is desirable. Farsighted individuals

account for how their actions in period 1 affect their pro-social identity I_1. They care about what their future selves will think of themselves. Here, one can draw a parallel between this future self and the "impartial observer" described by Adam Smith:

> "When I endeavour to examine my own conduct, [...] I divide myself, as it were, into two persons: and that I, the examiner and judge, represent a different character from that other I, the person whose conduct is examined and judged of" (Smith, 1853).

In more contemporaneous words, this property captures the fact that individuals care about their self-esteem. This model has the following properties:

PROPERTY 1. *CONSCIENCE ACCOUNTING: Individuals endowed with a low pro-social identity I_0, such that $\eta_1 + I_0 < \kappa_1$, reduce their pro-social effort of period 1 when knowing they will have another occasion to act pro-socially in period 2.*

All the proofs are presented in Appendix 2.A.A. Property 1 allows me to capture the conscience accounting effect described in section II. In the model, this effect only occurs for individuals with a low endowed pro-social identity. When this is the case, they perceive pro-social actions as substitutable.

PROPERTY 2. *CONSISTENT AND COMPENSATORY BEHAVIOURS: The amount of effort or time allocated to pro-social activities in period 2 is lower when pro-social activities of period 1 are extrinsically motivated ($\kappa_1 > \eta_1 + I_0$). It is higher when they are intrinsically motivated ($\eta_1 + I_0 > \kappa_1$).*

Property 2 captures two opposite phenomena described in section II. The first is a preference for consistency (e.g., Festinger 1962): a first selfless deed makes it harder for individuals to forego another one. The second is moral licensing (e.g., Monin and Miller 2001; Effron et al. 2009): the first pro-social act licenses future selfish acts. Property 2 states that the former effect is more likely to

occur when the initial act is intrinsically motivated and the latter when it is extrinsically motivated. Indeed, acting out of intrinsic motivations strengthens pro-social identity I_1, which increases the return of acting pro-socially again. Conversely, acting out of extrinsic motivation weakens pro-social identity I_1, which reduces the return of acting pro-socially in period 2. From Property 2 I state the following corollary:

COROLLARY 1. *MORAL CLEANSING: Failing to act pro-socially in period 1 increases the level of pro-social effort in period 2 for individuals endowed with a low pro-social identity (such that $\eta_1 + I_0 < \kappa_1$).*

Again, in the model, moral cleansing effects only occur for individuals endowed with a low pro-social identity.

III.B Pro-Social Policies

The context in which individuals make decisions influences individuals' motives for acting pro-socially. By changing the context, pro-social policies alter these motives. For instance, information campaigns making salient values and norms associated with a pro-social identity can increase individuals' intrinsic motivation to act pro-socially. Similarly, educative pieces of information inducing individuals to interiorise new values and norms could also increase their intrinsic motivation. On the other hand, situations characterised by peer pressure regarding a desired level of pro-social activity or associating material or immaterial rewards with pro-social deeds can increase extrinsic motivations.

One can, therefore, define a function mapping how a policy p, altering the context, affects pro-social motivations:[3] $\Phi : p \in \mathbb{R} \mapsto \{\eta, \kappa\} \in \mathbb{R}^2$. For simplicity, I assume a one-to-one mapping from

[3]One could even envision a functional form $\Phi(p, I)$ to account for how one's identity shaped one's understanding and interpretation of the context. These refinements are out of the scope of this paper.

policies to motivations. I also consider the simple case where policies either increase intrinsic or extrinsic motivations but cannot increase both at the same time.

Effects on Period One Choices

In this subsection, I study the effect of policies aiming to increase participation in pro-social activity of period one. I consider first the case of shortsighted individuals.

PROPOSITION 1. *Pro-social policies increase the amount of time or effort shortsighted individuals allocate to period one pro-social activities:*

$$\frac{\partial x_1}{\partial \theta_1} = \frac{\partial_{x_1\theta_1} V_1^s}{-\partial_{x_1x_1} V_1^s} > 0 \quad \forall \theta_1 \in \{\eta_1, \kappa_1\} \tag{III.5}$$

In the simple shortsighted case, both types of pro-social policies strictly increase pro-social efforts in period one. This is not the case when considering farsighted individuals:

PROPOSITION 2. *The direct effect of pro-social policies is the sum of three effects:*

$$\frac{\partial x_1}{\partial \theta_1} = \frac{1}{\Delta}\Big(\underbrace{\Phi_2 \cdot \partial_{\theta_1 x_1} V_1}_{\textit{Main effect}} + \underbrace{\Phi_2 \cdot \partial_{\theta_1 x_1} V_2}_{\textit{Self-esteem effect}} + \underbrace{\varphi \cdot \partial_{\theta_1 x_2} V_2}_{\textit{Trade-off effect}} \Big) \quad \forall \theta_1 \in \{\eta_1, \kappa_1\} \tag{III.6}$$

Where:

$$\partial_{\theta_1 x_1} V_2 \equiv \partial_{II} V_2 \cdot \partial_{x_1} I_1 \cdot \partial_{\theta_1} I_1 + \partial_I V_2 \cdot \partial_{\theta_1 x_1} I_1$$

$$\partial_{\theta_1 x_2} V_2 \equiv (\partial_{I x_2} V_2 - \partial_{I y_2} V_2) \cdot \partial_{\theta_1} I_1$$

And $\Phi_t \equiv -\partial_{x_t x_t} V_1^f > 0$, $\varphi \equiv \partial_{x_1 x_2} V_1^f$, *and* $\Delta \equiv \Phi_1 \Phi_2 - (\varphi)^2 > 0$ *by assumption 5.*

The *main effect*, as labelled in equation (III.6), is equivalent to the effect of pro-social policies in the shortsighted case. Yet, when reacting to policies, farsighted individuals account for what their choices say about themselves (*self-esteem* effect) and trade-off acting pro-socially now or later (*trade-*

off effect). In what follows, for simplicity, I consider the case where $\partial_{II}V_2 = 0$ (see Appendix 2.A.B for a discussion of the cases where $\partial_{II}V_2 < 0$).

PROPOSITION 3. *Assuming $\partial_{II}V_2 = 0$ and holding the trade-off effect constant, increasing intrinsic motivations is more effective than increasing extrinsic motivations.*

Acting pro-socially for the sake of a reward (i.e. when κ_1 is high) signals a greedy identity rather than a pro-social one. Rewarding pro-social actions can, therefore, have the counter-productive effect of reducing their uptake. This relates to taboo trade-off aversion: our reluctance to render a selfless act transactional (Fiske and Tetlock, 1997). For instance, payments for blood donations are often perceived negatively by donors as they crowd out the altruistic nature of giving blood (Chell et al., 2018). On the other hand, interventions playing on individuals' intrinsic motivations increase the return of signalling a pro-social identity to one's future self.

The *trade-off* effect captures the fact that farsighted individuals weigh acting pro-socially in period one when they know they will have another occasion to act pro-socially in period two. The sign of this effect depends on the extent to which *individuals are consistent across their choices* and how policies alter the *return of acting pro-socially in period two*. The expression that captures the extent to which individuals are consistent is:

$$\varphi = \partial_{x_1 x_2} V_1^f \Leftrightarrow \varphi = (\partial_{x_2 I} V_2 - \partial_{y_2 I} V_2) \cdot \partial_{x_1} I_1$$

When individuals are endowed with a high pro-social identity I_0 ($\eta_1 + I_0 > \kappa_1$), then, by assumption 3, φ is positive: pro-social deeds of period one and two are perceived as complementary. In other words, individuals prefer to stay consistent across their choices.

Conversely, when individuals have a low endowed pro-social identity I_0 ($\kappa_1 > \eta_1 + I_0$), then φ is

negative and pro-social deeds of periods one and two are perceived as substitutable. This implies that a first pro-social deed reduces the need to do another. The expression capturing the effect of pro-social policies on the return of a selfless act in period two is:

$$\partial_{\theta_1 x_2} V_2 = (\partial_{I x_2} V_2 - \partial_{I y_2} V_2) \cdot \partial_{\theta_1} I_1$$

Increasing intrinsic motivations reinforces individuals' pro-social identity, raising the return of acting pro-socially in period two. This is the opposite when policies increase extrinsic motivations. From there, I can derive the following proposition:

PROPOSITION 4. *Holding the self-esteem effect constant, increasing intrinsic motivations is more effective on individuals endowed with a high pro-social identity I_0. Increasing extrinsic motivations is more effective on individuals endowed with a low pro-social identity I_0.*

Proposition 4 rationalises several empirical evidence. For instance, Landry et al. (2010) find that offering a reward when soliciting past donors to give again is less effective than requests without contingent rewards. Conversely, rewards effectively attract donations from people who have never contributed. In the health domain, financial incentives increase healthy behaviours among the least sporty people and crowd out those of the sportiest (Charness and Gneezy, 2009; Gonzalez et al., 2023).[4] Propositions 3 and 4 imply that increasing extrinsic motivations generates negative *self-esteem* and *trade-off* effects for individuals endowed with a high pro-social identity I_0. Policies backfire when these effects outweigh the *main* effect.

[4] Although health-related behaviours cannot be considered pro-social, they retain some of their characteristics: (1) consumers refrain from indulging in immediate hedonic pleasure to benefit someone else, their future selves; (2) being healthy is socially desirable and can be constitutive of someone's identities; (3) motivations for being healthy can be intrinsic (e.g., "I run because I like it") or extrinsic (e.g., "I run for my summer body").

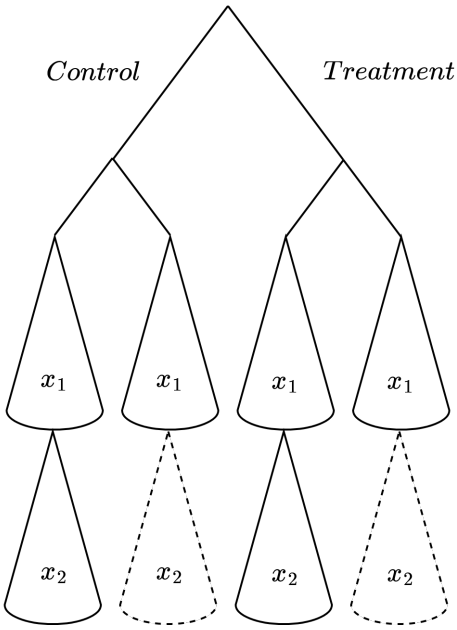


Figure 2.1: Randomisation tree to estimate the trade-off effect

Note: Randomisation tree of an experimental design that allows to estimate the trade-off effect. Dashed lines indicate that the non-targeted decision x_2 is not communicated to participants when making the targeted decision x_1.

PROPOSITION 5. *Policies increasing extrinsic motivations yield partial or net backfiring effects for individuals endowed with a high pro-social identity I_0.*

Taken together, the *self-signalling* and the *trade-off* effects explain why some policies sometimes backfire.[5] However, empirically disentangling these two effects is not trivial. The extent to which we care about our image can hardly be manipulated in an experimental setting. We can obtain some insights on the extent to which people's *self-esteem* is affected by pro-social policies by relying on self-reported measures with scales developed in the social psychology literature (e.g., Jordan et al. 2015).

[5]Another reason often invoked is reactance (e.g., Rains 2013). Reactance describes the feeling of a loss of autonomy triggered by policies, which can lead individuals to act in opposition to the policies' objectives. A way this model could capture this phenomenon is through the mapping function $\Phi : p \in \mathbb{R} \mapsto \{\eta, \kappa\}$: the policy induces a decrease in motivation for acting pro-socially.

The *trade-off* effect stems from the fact that individuals anticipate they will have another occasion to act pro-socially. This effect can be experimentally estimated using a design similar to Gneezy et al. (2014). Experimental subjects are randomly made aware of another occasion to perform a pro-social decision when doing the first, as in the randomisation tree in Figure 2.1.

In the next part of this section, I show that policies can also alter non-targeted decisions of period 2.

Effects on Period Two Choices

In this section, I study the effect of pro-social policies of period one on non-targeted pro-social decisions of period two. Irrespective of whether individuals are farsighted or shortsighted, I consider the simple case of an individual maximising period two utility, taking choices of period one as given. The model is solved by forward induction.

PROPOSITION 6. *Pro-social policies alter period two choices through their effect on period one choices (indirect spillover effect) and by altering the sign of this indirect spillover effect (direct spillover effect).*

$$\underbrace{\frac{dx_2}{d\theta_1}}_{\textit{Net spillover effect}} = \underbrace{\frac{\partial x_1}{\partial \theta_1} \times \frac{\partial x_2}{\partial x_1}}_{\textit{Indirect spillover}} + \underbrace{\frac{\partial x_2}{\partial \theta_1}}_{\textit{Direct spillover}} = \Gamma \cdot \Bigg[\underbrace{\partial_{x_1} I_1 \cdot \frac{\partial x_1}{\partial \theta_1}}_{\textit{Indirect spillover}} + \underbrace{\partial_{\theta_1} I_1}_{\textit{Direct spillover}} \Bigg] \tag{III.7}$$

Where $\Gamma \equiv \frac{\partial_{x_2 I} V_2 - \partial_{y_2 I} V_2}{-\partial_{x_2 x_2} V_2^s}$.

The *indirect spillover* effect captures the effect of undertaking the targeted pro-social decision on the non-targeted one ($\frac{\partial x_2}{\partial x_1}$), scaled by the main effect of the policy ($\frac{\partial x_1}{\partial \theta_1}$). When $\frac{\partial x_2}{\partial x_1}$ is positive, acting pro-socially in period one increases the appeal of acting pro-socially in period two. Conversely, when it is negative, acting pro-socially reduces the need to act pro-socially in period two. These dynamics are similar to those described by property 2. The *direct spillover* effect captures the ef-

fect of policies on the marginal utility of acting pro-socially in period two through their effect on individuals' pro-social identity.

PROPOSITION 7. *Increasing intrinsic motivations raises the return of acting pro-socially in period two. Conversely, increasing extrinsic motivations decreases the return of acting pro-socially in period two.*

Experimenters cannot directly observe the psychological mechanisms of behaviour change triggered by pro-social policies. Nevertheless, we can infer these mechanisms by studying the effects of policies on non-targeted decisions. Equation (III.7) describes the channels through which policies θ_1 spill over to non-targeted decisions x_2.

The sign of the *direct spillover* hints at potential psychological mechanisms. In the model, inducing individuals to act pro-socially as an "end to itself" yields positive *direct spillovers* by strengthening their pro-social identity. Conversely, negative *direct spillovers* arise when promoting pro-social actions as a "means to an end", weakening individuals' pro-social identity.

On the other hand, the sign of the *indirect spillovers* provides information on whether the two behaviours studied are perceived as complements or substitutes, provided the main effect of the policy on the targeted action is positive. Here, $\frac{\partial x_2}{\partial x_1}$ corresponds to the behavioural spillover effect described in the social psychology literature (Thøgersen, 1999). Estimating this effect is crucial to determine which behaviours constitute entry points towards adopting other pro-social behaviours.

How to experimentally estimate these effects? In a setting where the policy is randomised, behavioural spillovers and direct effects can be retrieved by regressing the non-targeted decision x_2 on the targeted one x_1 and on policy exposure θ_1. However, unobserved confounding variables likely bias the effect of x_1 on x_2. To obtain causal estimates, a solution is to embed an instrumental

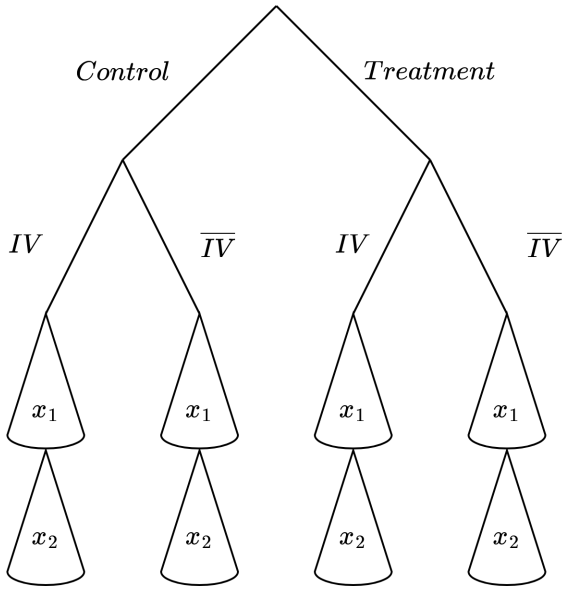


Figure 2.2: Randomisation tree to estimate spillover effects

Note: Randomisation tree of an experimental design allowing to disentangle behavioural spillovers from the direct effect of policies on non-targeted decisions.

variable for the targeted decision in the experimental design (see randomisation tree in Figure 2.2). Chapter 2 discusses this estimation strategy in detail and its associated assumptions.

IV Conclusion

I model pro-social actions as either intrinsically or extrinsically motivated. When pro-social deeds are undertaken out of intrinsic motivation, i.e., for their own sake, this strengthens individuals' pro-social identity. This makes them more likely to act pro-socially again. On the other hand, acting pro-socially out of extrinsic motivations, i.e., as a means to an end, has the opposite effect. It weakens individuals' pro-social identity, lowering engagement in subsequent pro-social actions. Linking pro-social actions in this way reconciles several mechanisms identified in the social psychology literature. It also generates interesting insights into the effect of policies promoting selfless actions.

First, I show that pro-social policies can affect non-targeted behaviours through two channels. They indirectly spill over non-targeted pro-social decisions through their influence on the targeted decision. They also directly affect them through their effect on individuals' motivations. Second, when individuals care about their self-esteem, i.e., what their choices say about themselves, pro-social policies targeting extrinsic motivations are less effective. Third, heterogeneity in the effect of pro-social policies arises when individuals anticipate another occasion to act pro-socially. Being able to trade off acting pro-socially now or later implies that interventions targeting intrinsic motivations are more effective on intrinsically motivated individuals. Conversely, interventions targeting extrinsic motivations are more effective on extrinsically motivated individuals.

This model opens new avenues in understanding how identity alters the effectiveness of pro-social policies. In its present version, the model ignores social network influences. Potential future extensions could account for this influence. Indeed, acting out of extrinsic motivations in a social context where extrinsically motivated pro-social actions are commonplace will likely induce different dynamics from the opposite case.

Bibliography

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The quarterly journal of economics*, 115(3):715–753.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477.

Baca-Motes, K., Brown, A., Gneezy, A., Keenan, E. A., and Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, 39(5):1070–1084.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Cameron, J. and Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational research*, 64(3):363–423.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018). Putting your money where your mouth is: an experimental test of pro-environmental spillover from reducing meat consumption to monetary donations. *Environment and Behavior*, 50(7):723–748.

Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.

Chell, K., Davison, T. E., Masser, B., and Jensen, K. (2018). A systematic review of incentives in blood donation. *Transfusion*, 58(1):242–254.

Comin, D. A. and Rode, J. (2023). Do green users become green voters? Technical report, National Bureau of Economic Research.

Cornelissen, G., Bashshur, M. R., Rode, J., and Le Menestrel, M. (2013). Rules or consequences? the role of ethical mind-sets in moral dynamics. *Psychological Science*, 24(4):482–488.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627.

Dolan, P. and Galizzi, M. M. (2014). Because i'm worth it: a lab-field experiment on the spillover effects of incentives in health. Centre for Economic Performance, London School of Economics and Political Science.

Dolan, P. and Galizzi, M. M. (2015). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16.

Effron, D. A., Cameron, J. S., and Monin, B. (2009). Endorsing obama licenses favoring whites. *Journal of experimental social psychology*, 45(3):590–593.

Elliot, A. J. and Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of personality and social psychology*, 67(3):382.

Evans, L., Maio, G. R., Corner, A., Hodgetts, C. J., Ahmed, S., and Hahn, U. (2013). Self-interest and pro-environmental behaviour. *Nature Climate Change*, 3(2):122–125.

Festinger, L. (1962). Cognitive dissonance. *Scientific American*, 207(4):93–106.

Fiske, A. P. and Tetlock, P. E. (1997). Taboo trade-offs: reactions to transactions that transgress the spheres of justice. *Political psychology*, 18(2):255–297.

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., and Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, 58(1):179–187.

Gneezy, U., Imas, A., and Madarász, K. (2014). Conscience accounting: Emotion dynamics and social behavior. *Management Science*, 60(11):2645–2658.

Gonzalez, N. I. V., Kee, J. Y., Palma, M. A., and Pruitt, J. R. (2023). The relationship between monetary incentives, social status, and physical activity. *Journal of Behavioral and Experimental Economics*, page 102155.

Jordan, J., Leliveld, M. C., and Tenbrunsel, A. E. (2015). The moral self-image scale: Measuring and understanding the malleability of the moral self. *Frontiers in Psychology*, 6:1878.

Kristofferson, K., White, K., and Peloza, J. (2014). The nature of slacktivism: How the social observ-

ability of an initial act of token support affects subsequent prosocial action. *Journal of Consumer Research*, 40(6):1149–1166.

Lacasse, K. (2016). Don't be satisfied, identify! strengthening positive spillover by connecting pro-environmental behaviors to an "environmentalist" label. *Journal of Environmental Psychology*, 48:149–158.

Landry, C. E., Lange, A., List, J. A., Price, M. K., and Rupp, N. G. (2010). Is a donor in hand better than two in the bush? evidence from a natural field experiment. *American Economic Review*, 100(3):958–983.

Lee, S. W. and Schwarz, N. (2010). Dirty hands and dirty mouths: Embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. *Psychological science*, 21(10):1423–1425.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315.

Mazar, N. and Zhong, C.-B. (2010). Do green products make us better people? *Psychological science*, 21(4):494–498.

Monin, B. and Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology*, 81(1):33.

Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human communication research*, 39(1):47–73.

Smith, A. (1853). *The theory of moral sentiments*. HG Bohn.

Steinhorst, J. and Matthies, E. (2016). Monetary or environmental appeals for saving electricity?–potentials for spillover on low carbon policy acceptability. *Energy Policy*, 93:335–344.

Tajfel, H., Turner, J. C., Austin, W. G., and Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16.

Thøgersen, J. (1999). Spillover processes in the development of a sustainable consumption pattern. *Journal of economic psychology*, 20(1):53–81.

Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013). For better or for worse? empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy*, 57:160–171.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138.

Ulph, A., Panzone, L., and Hilton, D. (2023). Do rational people sometimes act irrationally? a dynamic self-regulation model of sustainable consumer behavior. *Economic Modelling*, 126:106384.

Xu, L., Zhang, X., and Ling, M. (2018). Spillover effects of household waste separation policy on electricity consumption: evidence from hangzhou, china. *Resources, Conservation and Recycling*, 129:219–231.

2.A Appendix

2.A.A Proofs

Proof. (Property 1) The first order conditions of maximisation programme (III.4) are:

$$\begin{cases} \partial_{x_1} v(x_1, B_1 - x_1|\cdot) - \partial_{y_1} v(x_1, B_1 - x_1|\cdot) + \partial_I v(x_2, B_2 - x_2|I_1, \eta_2, \kappa_2)\partial_{x_1} I_1 = 0 \\ \partial_{x_2} v(x_2, B_2 - x_2|\cdot) - \partial_{y_2} v(x_2, B_2 - x_2|\cdot) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \partial_{x_1} v(x_1, B_1 - x_1|\cdot) - \partial_{y_1} v(x_1, B_1 - x_1|\cdot) + \partial_I v(x_2, B_2 - x_2|I_1, \eta_2, \kappa_2)\partial_{x_1} I_1 = 0 \\ x_2 = f(x_1, \eta_2, \kappa_2, \eta_1, \kappa_1, I_0) \end{cases}$$

So the level of pro-social activity x_1 that is the solution of the maximisation programme solves this equation:

$$\partial_{x_1} v(x_1, B_1 - x_1|\cdot) - \partial_{y_1} v(x_1, B_1 - x_1|\cdot) = \\ - \partial_I v(f(\cdot), B_2 - f(\cdot)|I_1, \eta_2, \kappa_2)\partial_{x_1} I_1$$

Which can be rewritten as follows:

$$\partial_{x_1} V_1^s(x_1, \cdot) = -\partial_I V_2^s(f(x_1, \cdot), \cdot)\partial_{x_1} I_1(x_1, \cdot)$$

When the right-hand side of this equation equals zero, this expression corresponds to the first-order condition of the maximisation programme (III.3). By assumption 5, we know that $\partial_{x_1} V_1^s(x_1, \cdot)$ is decreasing in x_1. By assumption 4, we know that $\partial_I V_2^s(f(x_1, \cdot), \cdot) > 0$. Finally, assumption 3 states that when individuals are intrinsically motivated ($\eta_1 + I_0 > \kappa_1$), we have $\partial_{x_1} I_1(x_1, \cdot) > 0$. This implies that the level of pro-social effort chosen in the farsighted case is always greater or equal to that of the shortsighted case. Conversely, when individuals are extrinsically motivated ($\kappa_1 > \eta_1 + I_0$), we have $\partial_{x_1} I_1(x_1, \cdot) < 0$. This implies that the level of pro-social

effort chosen in the farsighted case is always lower or equal to that of the shortsighted case. Conscience accounting follows from this. ■

Proof. (Property 2) In period 2, individuals solve the following equation given the choice they made in period 1:

$$\partial_{x_2} V_2^s(x_2, I_1, \eta_2, \kappa_2) = 0$$

Implicit differentiation yields:

$$\begin{aligned} &\partial_{x_2 x_2} V_2^s(x_2, I_1, \eta_2, \kappa_2) \cdot \frac{\partial x_2}{\partial x_1} + \partial_{x_2 I} V_2^s(x_2, I_1, \eta_2, \kappa_2) \partial_{x_1} I_1 = 0 \\ \Leftrightarrow &\frac{\partial x_2}{\partial x_1} = \frac{\partial_{x_2 I} V_2^s(x_2, I_1, \eta_2, \kappa_2) \partial_{x_1} I_1}{-\partial_{x_2 x_2} V_2^s(x_2, I_1, \eta_2, \kappa_2)} \end{aligned}$$

By assumption 3, this expression is positive when individuals are intrinsically motivated ($\eta_1 + I_0 > \kappa_1$) and negative otherwise. ■

Proof. (Proposition 1) When decision-makers are shortsighted, they solve maximisation program (III.3) whose first order condition is:

$$\partial_{x_1} v(x_1, B_1 - x_1 | I_0, \eta_1, \kappa_1) - \partial_{y_1} v(x_1, B_1 - x_1 | I_0, \eta_1, \kappa_1) = 0 \tag{2.A.1}$$

Which can be rewritten as follows for exposition purposes:

$$\partial_x V_1 - \partial_y V_1 = 0 \tag{2.A.2}$$

Expression (III.5) is obtained by applying the implicit function theorem to equation (2.A.2). First I differentiate (2.A.2) with respect to $\theta_1 \in \{\eta_1, \kappa_1\}$:

$$\Big(\partial_{xx}V_1 - 2\partial_{xy}V_1 + \partial_{yy}V_1\Big)\frac{\partial x_1}{\partial \theta} + \partial_{\theta x}V_1 - \partial_{\theta y}V_1 = 0$$

Then I isolate $\frac{\partial x_1^s}{\partial \theta_1}$:

$$\frac{\partial x_1^s}{\partial \theta_1} = \frac{\partial_{\theta x}V_1 - \partial_{\theta y}V_1}{-(\partial_{xx}V_1 - 2\partial_{xy}V_1 + \partial_{yy}V_1)}$$

Assumptions 2 and 5 imply that the denominator and the numerator are positive. ■

Proof. (Proposition 2) Decision makers solve a system of two equations corresponding to the first order conditions of the maximisation program (III.4), such as:

$$\begin{cases} \partial_{x_1}V_1^f = 0 \\ \partial_{x_2}V_1^f = 0 \end{cases} \Leftrightarrow \begin{cases} \partial_{x_1}V_1 - \partial_{y_1}V_1 + \partial_I V_2 \cdot \partial_{x_1}I_1 = 0 \\ \partial_{x_2}V_2 - \partial_{y_2}V_2 = 0 \end{cases} \tag{2.A.3}$$

Denote by $\nabla_{\theta_1}\boldsymbol{x}$ and $\nabla_{\theta_1\boldsymbol{x}}V_1^f$ the vectors $(\frac{\partial x_1}{\partial \theta_1}, \frac{\partial x_2}{\partial \theta_1})$ and $(\partial_{\theta_1 x_1}V_1^f, \partial_{\theta_1 x_2}V_1^f)$. The Hessian of this problem is:

$$\mathcal{H}_{Uf} = \begin{pmatrix} -\Phi_1 & \varphi \\ \varphi & -\Phi_2 \end{pmatrix}$$

Where $\Phi_t \equiv -\partial_{x_t x_t}V_1^f > 0$, and $\varphi \equiv \partial_{x_1 x_2}V_1^f$. Expression (III.6) is obtained by applying the implicit function theorem to the system of equations (2.A.3). The implicit function theorem implies that:

$$\nabla_{\theta_1}\boldsymbol{x} = \mathcal{H}_{Uf}^{-1} \cdot \nabla_{\theta_1\boldsymbol{x}}V_1^f$$

Where:

$$\mathcal{H}_{Uf}^{-1} = \frac{1}{\Delta}\begin{pmatrix} \Phi_2 & \varphi \\ \varphi & \Phi_1 \end{pmatrix}$$

By assumption 5, we have $\Delta \equiv \det(\mathcal{H}_{Uf}) = \Phi_1\Phi_2 - (\varphi)^2 > 0$. ■

Proof. (Proposition 3) Assuming that $\partial_{II}V_2 = 0$ implies that $\partial_{\theta_1 x_2}V_2 = \partial_I V_2 \cdot \partial_{\theta_1 x_1} I_1$. By assumption 3, this expression is negative when $\theta_1 = \kappa_1$ and positive when $\theta_1 = \eta_1$. ■

Proof. (Proposition 4) Policies increasing intrinsic motivations increase the marginal utility of acting pro-socially in period two, which increases (decreases) the return of acting pro-socially in period one when the two decisions are complementary (substitutable). Policies increasing extrinsic motivations reduce the marginal utility of acting pro-socially in period two, which decreases (increases) the return of acting pro-socially in period one when the two decisions are complementary (substitutable). ■

Proof. (Proposition 5) Proposition 5 is a direct implication of propositions 3 and 4. ■

Proof. (Proposition 6) When deciding on their level of consumption in period two, decision-makers solve the following:

$$\partial_{x_2}V_2 - \partial_{y_2}V_2 = 0$$

Denote the solution of this equation by $x_2(x_1(\theta_1), \theta_1)$. To derive expression (III.7), I differentiate this equation by $\theta_1 \in \{\eta_1, \kappa_1\}$ and apply the implicit function theorem. Differentiating by θ_1 yields:

$$(\partial_{x_2x_2}V_2 - 2\partial_{x_2y_2}V_2 + \partial_{y_2y_2}V_2) \cdot \frac{dx_2(x_1(\theta_1), \theta_1)}{d\theta_1} + (\partial_{x_2I_1}V_2 - \partial_{y_2I_1}V_2) \cdot (\partial_{x_1}I_1 \frac{\partial x_1}{\partial \theta_1} + \partial_{\theta_1}I_1)$$

Rearranging these terms yields:

$$\frac{dx_2(x_1(\theta_1),\theta_1)}{d\theta_1} = \Gamma \cdot \left[\partial_{x_1} I_1 \frac{\partial x_1}{\partial \theta_1} + \partial_{\theta_1} I_1\right]$$

Where:

$$\Gamma \equiv \frac{\partial_{x_2 I_1} V_2 - \partial_{y_2 I_1} V_2}{-\partial_{x_2 x_2} V_2^s}$$

$$\partial_{x_2 x_2} V_2^s \equiv \partial_{x_2 x_2} V_2 - 2\partial_{x_2 y_2} V_2 + \partial_{y_2 y_2} V_2$$

$$\frac{dx_2(x_1(\theta_1),\theta_1)}{d\theta_1} = \frac{\partial x_1}{\partial \theta_1} \times \frac{\partial x_2}{\partial x_1} + \frac{\partial x_2}{\partial \theta_1}$$

■

Proof. (Proposition 7) By assumption 3, when $\theta_1 = \eta_1$, then the direct effect is positive. It is negative when $\theta_1 = \kappa_1$. ■

2.A.B Special cases

In this section, I relax the assumption that $\partial_{II} V_2 = 0$ to consider the cases where $\partial_{II} V_2 < 0$.

When $\eta_1 + I_0 > \kappa_1$ and $\theta_1 = \eta_1$: When individuals are intrinsically motivated ($\eta_1 + I_0 > \kappa_1$), and the policy plays on intrinsic motivations, then individuals face a trade-off. Either they increase their pro-social effort of period 1 as they derive higher self-esteem ($\partial_I V_2 \cdot \partial_{\eta_1 x_1} I_1 > 0$), or they indulge in self-serving activities and maintain a constant level of self-esteem ($\partial_{II} V_2 \cdot \partial_{\eta_1} I_1 \cdot \partial_{x_1} I_1 < 0$).

When $\eta_1 + I_0 > \kappa_1$ and $\theta_1 = \kappa_1$: When individuals are intrinsically motivated ($\eta_1 + I_0 > \kappa_1$), and the policy plays on extrinsic motivations, then the negative effect of acting pro-socially on

self-esteem ($\partial_I V_2 \cdot \partial_{\kappa_1 x_1} I_1 < 0$) is attenuated ($\partial_{II} V_2 \cdot \partial_{\kappa_1} I_1 \cdot \partial_{x_1} I_1 > 0$). Intuitively, even though the policy decreases the return of acting pro-socially on self-esteem ($\partial_{\kappa_1 x_1} I_1 < 0$), this return is still positive ($\partial_{x_1} I_1 > 0$).

When $\eta_1 + I_0 < \kappa_1$ and $\theta_1 = \eta_1$: When individuals are extrinsically motivated ($\eta_1 + I_0 < \kappa_1$), and the policy plays on intrinsic motivations, then this reinforces the return of acting pro-socially on self-esteem ($\partial_I V_2 \cdot \partial_{\eta_1 x_1} I_1 > 0$ and $\partial_{II} V_2 \cdot \partial_{\eta_1} I_1 \cdot \partial_{x_1} I_1 > 0$). Intuitively, even though individuals still act out of extrinsic motivations and, therefore, harm their self-esteem, the policy decreases the negative return of pro-social deeds on self-esteem, making it less harmful.

When $\eta_1 + I_0 < \kappa_1$ and $\theta_1 = \kappa_1$: When individuals are extrinsically motivated ($\eta_1 + I_0 < \kappa_1$), and the policy plays on extrinsic motivations, then this reinforces the negative effect of acting pro-socially on self-esteem ($\partial_I V_2 \cdot \partial_{\kappa_1 x_1} I_1 < 0$ and $\partial_{II} V_2 \cdot \partial_{\kappa_1} I_1 \cdot \partial_{x_1} I_1 < 0$). Intuitively, acting out of extrinsic motivations harms their self-esteem. By strengthening extrinsic motivations, the policy makes it even more harmful to act pro-socially.

Chapter 3

Estimating the Side Effects of Social Norm Nudges

I Introduction

Climate change is one of the most critical challenges of the 21st century. Devastating economic consequences are looming without significant lifestyle changes in industrialised countries (Shukla et al., 2022). Research suggests that an initial pro-environmental action influences our propensity to do more. This "behavioural spillover", as coined by Thøgersen (1999), can take different forms. For instance, Comin and Rode (2023) find that installing solar panels increases people's likelihood to vote for green parties. Conversely, Mazar and Zhong (2010) find that people become less altruistic after buying green products. Thus, policymakers should not only focus on promoting actions yielding large decreases in carbon emissions, but they should also promote actions that inspire people to do more for the environment. But do positive behavioural spillovers persist when policies cause the initial pro-environmental action?

In this chapter, we develop an empirical strategy to answer this question. We then focus on a social norm nudge promoting vegetarianism in an online randomised control trial (n=2775). Meat consumption is an important source of greenhouse gas emissions and should be reduced (Green et al., 2015; Riahi et al., 2022; Bonnet et al., 2020). However, we do not know whether changing our diet makes us want to do more for the environment. Social norm nudges are simple messages. They give information on what others do, approve or disapprove (Bicchieri, 2016). These messages are effective in shifting behaviours[1]. In the environmental domain, they have been used to foster re-

[1]See Rhodes et al. (2020) and Melnyk et al. (2010) for meta-analyses on the effectiveness of social norm messaging in general. For meta-analyses and reviews of the effectiveness of social norm messaging applied to the environmental domain, see Farrow et al. (2017); Abrahamse and Steg (2013); Andor and Fels (2018); Cialdini and Jacobson (2021).

cycling,[2] promote sustainable diets,[3] improve water and electricity consumption,[4] and even foster towel reuse in hotels.[5] Yet, little is known about these messages' side effects on non-targeted pro-environmental decisions. In our experiment, our social norm message emphasises an increasing trend of vegetarianism. We randomly show respondents the message before letting them choose their preferred meal on a restaurant menu. At the end of the experiment, respondents can donate to a pro-environmental charity of their choice. We use this task to proxy their willingness to do more for the environment.

In the model developed in Chapter 2, I show that nudges spill over non-targeted pro-environmental behaviours through two channels. The first channel is indirect. Nudges foster the initial pro-environmental decision, triggering behavioural spillovers. Positive behavioural spillovers arise when the initial decision is intrinsically motivated (e.g. because it is "something we care about"). Acting out of convictions reinforces the need to do more. Negative behavioural spillovers arise when the initial decision is extrinsically motivated (e.g., as "a means to an end"). Acting for a reward frees people from doing more once the reward is obtained. The second channel is direct and either amplifies or weakens behavioural spillovers. Its sign depends on whether nudges play on intrinsic or extrinsic motivations to foster the initial pro-environmental action. The sign of the direct spillover effect indicates the mechanisms through which nudges operate. A positive direct spillover effect implies that nudges enhance intrinsic motivations. A negative direct spillover effect implies that nudges enhance extrinsic motivations.

[2]See for instance Andersson and von Borgstede (2010); Bratt (1999); Fornara et al. (2011); Nigbur et al. (2010).

[3]See for instance Sparkman and Walton (2017); Sparkman et al. (2020); Salmivaara and Lankoski (2019); Testa et al. (2018); Stea and Pickering (2019); Wenzig and Gruchmann (2018); Richter et al. (2018).

[4]See for instance Allcott (2011); Costa and Kahn (2013); Carrico and Riemer (2011); Nolan et al. (2008); Handgraaf et al. (2013); Ferraro et al. (2011); Lapinski et al. (2007).

[5]See for instance Reese et al. (2014); Goldstein et al. (2008); Schultz et al. (2008).

Disentangling these two channels is crucial to understanding how nudges alter non-targeted decisions. Nevertheless, getting a causal estimate of behavioural spillovers is difficult. In the experiment, we embed an instrumental variable in the design. Namely, beyond allocating participants into control (no message) and treatment groups (receiving the social norm message), we vary the salience of vegetarian items on the restaurant menus. This alters the likelihood of choosing a vegetarian dish without directly affecting donations. This allows us to estimate the causal effect of choosing a vegetarian meal on donations.

Respondents' inclination to follow the norm may differ from one person to another. As such, respondents can perceive the social norm nudge differently. Hence, the effects of the nudge on food choices and donations may be heterogeneous. In another treatment arm (n=2782), respondents revealed their inclination to follow the norm. We use this extra survey data to investigate this heterogeneity. As part of an exploratory analysis, we train a gradient tree boosting classifier on this additional dataset to predict this inclination based on respondents' social-demographic characteristics, attitudinal information and self-reported beliefs (Friedman, 2001). We use this algorithm to classify respondents from the main experiment into different profiles. This allows us to get a conditional treatment effect of the nudge for each profile. Unlike mediation analysis, our heterogeneity analysis does not rely on direct measurements. This sidesteps the challenges of pre-treatment questions that can hint at the study's objectives. Furthermore, unlike other machine learning techniques, as detailed by Künzel et al. (2019), the source of heterogeneity is explicit. In our case, heterogeneity stems from people's readiness to conform to the norm.

Our results show that the social norm nudge is effective. The message increases the likelihood of choosing a vegetarian item on average. Respondents predicted to be trying to follow the norm drive this effect. However, they do not significantly decrease the carbon footprint of their food

choices. Conversely, respondents predicted to be hesitant about following the norm did not choose more vegetarian food but made less carbon-intensive food choices when nudged. The nudge does not affect the choices of respondents who are predicted to be unwilling to conform and those who are predicted to be already conforming. Bryan et al. (2021) recommends addressing heterogeneity when evaluating behavioural policies. Our study confirms the importance of doing so. Our results provide insights into the social-demographic profiles prone to change after seeing a social norm message. To our knowledge, we are the first to conduct such an investigation.

We also find evidence of a positive behavioural spillover effect on average. Namely, respondents choosing vegetarian food are more likely to give to pro-environmental charities. However, the social norm nudge decreases donations of those predicted to be trying to conform through a negative direct spillover effect. The negative direct spillover effect dominates the positive behavioural spillover effect. The model of Chapter 2 suggests that the nudge pushes this group to act out of extrinsic motivation (e.g., through social pressure). This, in turn, reduces their engagement in the donation task. Our results suggest that choosing to eat less meat encourages people to do more for the environment. However, *there is no free lunch*. When the social norm nudge succeeds in fostering vegetarian choices, it also crowds out this positive behavioural spillover by triggering extrinsically motivated vegetarian decisions.

We contribute to a burgeoning literature studying the side effects of policies. The meta-analyses of Maki et al. (2019) and Geiger et al. (2021) find only weak evidence for behavioural spillovers. However, methodological discrepancies make studies hard to compare.[6] This could explain this

[6]Some studies compare respondents exposed to a policy with those allocated to a control group (Carrico et al., 2018; Liu et al., 2021; Wolstenholme et al., 2020; Van Rookhuijzen et al., 2021; Jessoe et al., 2021; Goetz et al., 2022). This method does not distinguish policies' direct effects from behavioural spillovers. Other studies randomly offer participants the targeted behaviours to estimate spillover effects (Alt and Gallier, 2022; Clot et al., 2022; Margetts and Kashima, 2017). This design

scarcity of compelling evidence. Our estimation strategy aligns with Bonev (2023)'s recommendations to estimate behavioural spillovers. To our knowledge, only Alacevich et al. (2021), Comin and Rode (2023), and Alt et al. (2023) have used an instrumental variable to estimate behavioural spillovers. Our paper is most closely related to Alt et al. (2023). In a concurrent study, the authors assessed how different prompts altered participation in a non-targeted task. We differ from them by using an empirical strategy grounded in theory. This enables us to infer the mechanisms of nudges from the signs of their direct spillover effects. We also look at pro-environmental decisions whilst Alt et al. (2023) used abstract real-effort tasks. Thus, our study provides richer insights into the trade-offs policymakers may face between nudging pro-environmental behaviours and crowding out others.

The remaining of this article is articulated as follows. In Section II, we present our empirical strategy. Section III presents the experiment and the data. Section IV details the statistical models we use. The results are presented and discussed in Section V. Section VI explores the heterogeneity of the effects of the social norm nudge. Section VII concludes.

II Empirical Strategy

Intuitions of mechanisms: The effect of nudges on non-targeted decisions can be decomposed into two channels, as depicted in Figure 3.1. The red arrow captures behavioural spillovers, i.e., the effect of doing the targeted pro-environmental behaviour on other pro-environmental actions. The motivation underpinning the targeted action determines the sign of this effect. When the first action is intrinsically motivated — i.e., done out of convictions — then this makes it harder not

supposes that choosing (not) to do the targeted behaviour is the same as (not) being proposed to do it. This assumption is, however, debatable.

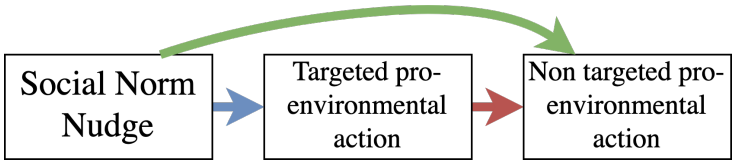


Figure 3.1: Effect of policies on non-targeted decisions.

Note: The blue arrow is the main effect of the policy on the targeted decision. The red arrow captures the effect of doing the targeted decision on one's likelihood of doing the non-targeted decision (behavioural spillover). The green arrow captures the direct spillover effect of the policy on the non-targeted decision.

to act pro-environmentally again: we do not like to be at odds with our convictions. In this case, behavioural spillovers are positive. When we act pro-environmentally out of extrinsic motivations — e.g., to get a reward or avoid a sanction — there is no need to do more once the reward is obtained or the sanction is avoided. In this case, behavioural spillovers are negative. Policies can play on intrinsic or extrinsic motivations to foster the targeted action. In doing so, they trigger a direct spillover effect on non-targeted decisions (green arrow). When policies enhance intrinsic (extrinsic) motivations, they reinforce (weaken) the willingness to be consistent.

The signs of behavioural spillovers can tell us if the environmental action is an "entry point" for other green actions. On the other hand, direct spillover effects tell us if policies weaken or reinforce behavioural spillovers. However, estimating behavioural and direct spillovers is not trivial. Two complications arise. First, getting a causal estimate of behavioural spillovers is difficult. Unobserved variables can affect several pro-environmental actions simultaneously (e.g., values and beliefs). Second, a policy can enhance intrinsic motivations for some people and extrinsic motivations for others. Thus, direct spillovers can differ from one person to another. This section develops an empirical framework to address these two issues.

Addressing omitted variable biases First, we assume a population of N individuals indexed by i. Individuals are randomly exposed to a policy fostering a given pro-environmental deed. Denote by $\boldsymbol{x}_1$ the $N \times 1$ vector capturing individuals' decision to do the targeted pro-environmental action. Denote by $\boldsymbol{\theta}_1$ the $N \times 1$ vector capturing their treatment status. The following linear models estimate the effects of the policy on $\boldsymbol{x}_1$ and a non-targeted pro-environmental decision $\boldsymbol{x}_2$.:

$$x_{1i} = \alpha^{ME} + \beta^{ME}\theta_{1i} + \varepsilon_i^{ME} \tag{II.1}$$

$$x_{2i} = \alpha^{SE} + \beta^{SE}\theta_{1i} + \varepsilon_i^{SE} \tag{II.2}$$

Here, $\hat{\beta}^{ME}$ is the estimate of the effect of the policy on the targeted decision, $\boldsymbol{x}_1$. We refer to it as the main effect of the policy. $\hat{\beta}^{SE}$ is the estimate of the effect of the policy on the non-targeted decision, $\boldsymbol{x}_2$. We refer to it as the net spillover effect of the policy. These estimates are unbiased if the stable unit treatment value assumption holds and the error terms ε_i^{ME} and ε_i^{SE} are such that $cov(\boldsymbol{\varepsilon}^{ME}, \boldsymbol{\theta}_1) = cov(\boldsymbol{\varepsilon}^{SE}, \boldsymbol{\theta}_1) = 0$. This equality holds when the policy is randomised. As shown in Chapter 2, we can decompose the net spillover effect of policies as follows:

$$\underbrace{\frac{\Delta x_2}{\Delta \theta_1}}_{\text{Net spillover}} = \underbrace{\frac{\partial x_1}{\partial \theta_1}}_{\text{Main effect}} \times \underbrace{\frac{\partial x_2}{\partial x_1}}_{\text{Behavioural spillover}} + \underbrace{\frac{\partial x_2}{\partial \theta_1}}_{\text{Direct spillover}} \tag{II.3}$$

In what follows, we make the following assumption:

ASSUMPTION 1. *The magnitude and the sign of the behavioural spillover effect do not depend on the policy.*

Assumption 1 reflects the insights provided by the model. A naive approach to dissociate behavioural from direct spillover effects consists of fitting the following linear model:

$$x_{2i} = \tilde{\alpha} + \tilde{\beta}^{BS} x_{1i} + \tilde{\beta}^{C} \theta_{1i} + \tilde{\varepsilon}_i \tag{II.4}$$

$\hat{\tilde{\beta}}^{BS}$ is a naive estimate of the behavioural spillover. $\hat{\tilde{\beta}}^{C}$ is the naive estimate of the direct spillover. These estimates are biased when unobserved variables simultaneously affect $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, implying $cov(x_{1i}, \tilde{\varepsilon}_i) \neq 0$. This omitted variable bias can be solved with an instrumental variable. A good instrumental variable alters $\boldsymbol{x}_1$ without changing people's intrinsic or extrinsic motivations to do $\boldsymbol{x}_1$. This is equivalent to randomly allocating people to a pure *choice-architecture* nudge, i.e., a variation in the choice environment unconsciously altering individuals' likelihood to do $\boldsymbol{x}_1$. Denote by $\boldsymbol{c}_1$ the $N \times 1$ vector capturing people's allocation to this *choice architecture* nudge. We can then get unbiased estimates of behavioural and direct spillovers with two-stage least squares:

$$\begin{aligned} &\text{Stage 1: } x_{1i} = \alpha + \beta_1 c_{1i} + \beta_2 \theta_{1i} + \varepsilon_i \\ &\text{Stage 2: } x_{2i} = \alpha' + \beta^{BS} \hat{x}_{1i} + \beta^{C} \theta_{1i} + \varepsilon'_i \end{aligned} \tag{II.5}$$

Where $\hat{x}_{1i}$ are the predicted values for the first stage. Our instrumental variable should be relevant ($cov(\boldsymbol{c}_1, \boldsymbol{x}_1) \neq 0$), exogenous ($cov(\boldsymbol{c}_1, \boldsymbol{\varepsilon}') = 0$) and homogeneous ($x_{1i}(\bar{c}_1) \geq x_{1i}(\underline{c}_1)\ \forall i \in [1, ..., N]$ and $\bar{c}_1 > \underline{c}_1$). Estimates of behavioural and direct spillovers are unbiased when this is the case. Furthermore, one can derive the following proposition:

PROPOSITION 8. *Estimates of models* (II.2) *and* (II.5) *are such that:*

$$\underbrace{\hat{\beta}^{SE}}_{\textit{Net spillover}} = \underbrace{\hat{\beta}^{ME}}_{\textit{Main effect}} \times \underbrace{\hat{\beta}^{BS}}_{\textit{Behavioural spillover}} + \underbrace{\hat{\beta}^{C}}_{\textit{Direct spillover}} \tag{II.6}$$

See Appendix 3.A.A for the proof. Proposition 8 shows that we can interpret estimates of model (II.2) and (II.5) in the same way as equation (II.3).

Addressing heterogeneity Different people may react differently to a policy. We propose to explore this heterogeneity by defining different types. We define types according to characteristics influencing people's reactions to a policy. We then collect two data sets: a *main* sample and a *training* sample. In the *main* sample, we randomise the policy $\boldsymbol{\theta}_1$ and a choice architecture nudge $\boldsymbol{c}_1$. In the *training* sample, we elicit the types of new respondents. We use the *training* data to train an algorithm to predict these types. We then predict the types of respondents in the *main* sample with the algorithm.

Let us index by $j \in [1, .., N']$ the N' observations in the *training* sample where each observation's type y_j is known. Denote by $\boldsymbol{W}$ and $\boldsymbol{W}'$ the $N \times M$ and $N' \times M$ matrices of covariates of the *main* and the *training* samples. In three steps, we estimate the conditional average treatment effects of policy θ_1. First, estimate the function $y_i = f(W_i')$ such that:

$$\hat{f} \in \arg\min_{f} L(y_i, f(W_i')) \tag{II.7}$$

Where $L(\cdot)$ is a loss function. Then, predict the types of observations in the *main* sample:

$$\hat{y}_i = \hat{f}(W_i) \tag{II.8}$$

Finally, the treatment effects for each type are estimated.

The remainder of the chapter presents an application of this empirical framework to the case of a social norm nudge promoting vegetarianism.

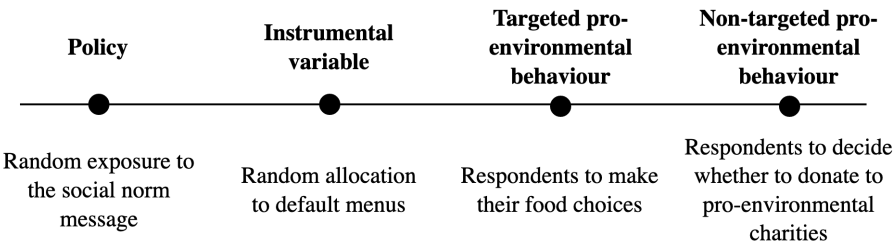


Figure 3.1: Timeline of the experiment

III Experimental Design and Data Collection

We study the side effects of a social norm nudge promoting vegetarian diets. We test if choosing vegetarian food increases environmental donations, our proxy for respondents' willingness to do more for the environment. We then assess whether the social norm nudge amplifies or weakens this behavioural spillover through a direct spillover effect.

We designed the survey experiment on Qualtrics and recruited respondents via Prolific. The experiment lasted approximately 10 minutes. We paid respondents according to Prolific's standard payment rate, £5 per hour. Upon finishing the survey, respondents have a 1/100 chance to win a £20 voucher. In total, we recruited a sample of 5,557 English respondents. They were divided between a *main* sample (n=2,775) and a *training sample* (n=2,782).

Respondents in the *main* sample took part in the main experiment. Its timeline is presented in Figure 3.1.[7] We use the *training* sample to look at the heterogeneity in our treatment effects as part of an exploratory analysis (see subsection VI).

[7]The survey questionnaire can be found here. We pre-registered the experimental design, power analysis, empirical strategy and instrumental variable strategy on Open Science Framework (here). The pre-analysis plan describes a broader project where three strands of research are investigated: 1) the effect of familiar food choices on one's inclination to choose vegetarian food; 2) the effect of reflection on the effectiveness of social norm nudges (now published, Banerjee and Picard 2023); 3) the present study. When reporting our results, we correct for the pre-registered hypotheses. Deviations from the pre-analysis plan are documented and justified here.

Policy: The policy of interest in this experiment is a social norm nudge. More precisely, we consider the following dynamic social norm message[8]:

> *A study published in The Lancet Planetary Health found that the share of British people who stopped eating meat has increased by more than 50% from 2008 to 2019. More and more people are choosing plant-based dishes that are kinder to the planet and in turn, are becoming climate-friendly.*

Its formulation is like the one used by Blondin et al. (2022). The authors find that this message effectively increases vegetarian food choice intentions. We randomly divided respondents into a treatment group where they see the message before making food choices (n=1391) or a control group (n=1384).

Instrumental variable: As explained in Section II, estimating behavioural spillovers requires embedding an instrumental variable in the design. To do this, we vary the salience of vegetarian options when respondents make food choices. Respondents first see a subset of food items presented as the chef's selection (see Figures 3.A.4 and 3.A.3 in Appendix 3.A.D). Half the respondents see a selection containing mostly meat-based items (n=1383). The other half see a selection containing mostly vegetarian options (n=1392). Respondents can choose an item from this selection or opt out and access the main menu containing all the items. We expect that respondents are more likely to choose a vegetarian item when vegetarian items are salient. Table 3.1 presents the sample size of each subgroup formed by the interaction between allocation to the nudge and the selections.

Targeted pro-environmental behaviour: We reproduce an online food order environment where participants choose a dish from a restaurant menu. The targeted pro-environmental decision is

[8]We construct it using the study of Stewart et al. (2021) analysing UK meat consumption trends using data from the National Diet and Nutrition Survey.

Table 3.1: Sample sizes of treatment groups

		Policy	
		Control	Treatment
Instrumental variable	Plant-intensive	690	693
	Meat-intensive	694	698

whether participants choose vegetarian food. We designed 24 versions of the main menu, varying the items' ordering and appearance. In all menus, we label food items with pictures of footprints ranging from green to red. An explanation indicates that green footprints mean "completely climate-friendly" and red footprints mean "not climate-friendly at all" (see Figure 3.A.2 in Appendix 3.A.D). As such, all participants have the same information on the environmental consequences of their choices. Table 3.A.16 in Appendix 3.A.D presents the characteristics of the dishes in the menus.

Non-targeted pro-environmental behaviour: At the end of the survey, we ask participants if they want to donate an amount between £0 and £10 to a pro-environmental charity.[9] This task is our non-targeted pro-environmental behaviour. We use donations to proxy respondents' willingness to do extra pro-environmental behaviours. Donations are consequential: we deduct them from the £20 voucher.

Sample characteristics We collected data from March 1st to April 24th of 2022. We pre-screened participants to select only native English speakers. We also excluded vegetarian and vegan participants. Attrition is low: 4.1% of respondents did not finish the survey. We excluded them. Table 3.2 shows descriptive statistics per treatment group. The median respondent is 35 years old,

[9]Respondents are offered to give to the following charities: World Wide Fund (WWF), Friends of the Earth, Carbon Fund, Campaign against Climate Change, The Vegetarian Society, The Vegan Society, Extinction Rebellion, Woodland Trust. Alternatively, they can select "other" and write the name of their chosen charity.

Table 3.2: Descriptive statistics

	Control group (n=1384)	Social norm group (n=1391)	p-value
Age			0.139
Mean	38.6 years old	37.9 years old	
Median	36 years old	35 years old	
Income			0.920
$<$ £10,000	18.6%	17.9%	
£10,000 - £15,999	11.5%	12.5%	
£16,000 - £19,999	11.3%	10.8%	
£20,000 - £29,999	27.2%	28.1%	
£30,000 - £39,999	16.2%	14.9%	
£40,000 - £49,999	8.5%	8.4%	
£50,000 - £69,999	4.5%	4.6%	
£70,000 - £89,999	1.5%	1.9%	
£90,000 - £119,999	0.6%	0.5%	
£120,000 - £149,999	0.2%	0.2%	
More than £150,000	0.0%	0.2%	
Gender			0.450
Female	48.3%	51.0%	
Male	50.7%	48.2%	
Other	1.0%	0.7%	
Education			0.961
No education	0.1%	0.1%	
Primary education	0.2%	0.1%	
Lower secondary education	2.5%	2.6%	
Upper secondary education	22.6%	21.9%	
Post-secondary non-tertiary education	15.6%	15.0%	
Short-cycle tertiary education	5.5%	6.6%	
Bachelor or equivalent	40.2%	39.4%	
Master or equivalent	11.9%	12.9%	
Doctoral or equivalent	1.5%	1.4%	

Note: Descriptive statistics per treatment group. We use a Wilcoxon test to determine the age difference between the treatment and control groups. We use a Chi-square test for gender differences. We use trend tests to determine the differences in education and income between the two groups.

earns between £20,000 and £30,000 per year and has a Bachelor's degree. There is a good gender balance, with 49.9% of females, 49.2% of males and 0.9% of respondents considering themselves genderfluid or agender. Comparisons using the UK census data and the survey of personal income suggest that our sample is younger, slightly poorer and more educated than the UK population (see Figure 3.A.6 in Appendix 3.A.D). Randomisation was successful. No significant differences exist across the treatment groups regarding age, gender, income, and education. About 98.28% of the *main* sample has passed an attention check placed at the beginning of the survey.[10] From these 98.28%, 99.75% passed a focus check we placed after the pre-treatment questionnaire.[11] Furthermore, 81.69% of the participants passed a manipulation check between the food choice and the donation task.[12] This suggests that respondents were attentive when taking the survey.

[10]They have to answer the following question on a 5-Likert scale, ranging from "not at all interested" to "extremely interested": "*People are very busy these days, and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both 'extremely interested' and 'very interested'*."

[11]Participants have to answer the following question: "*Most modern theories of decision making recognise that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below. Based on the text you read above, what colour have you been asked to select?*" They can select as many colours as they want from six colours. If they fail it, we show them the following message: "*The last question was here to check if you are being attentive. You did not answer it correctly. We are really interested in what you genuinely prefer. We kindly request you to read the questions more attentively.*"

[12]This attention check was the following:
Before being shown the restaurant menu, you were shown a message. What was the message about? [a) *People changing diets to become climate-friendly*, b) *People changing their diets to lose weight*, c) *People changing their diets to respect animals' well-being*, d)*I was not shown any specific message*, e) *I do not remember any specific message displayed*]

IV Statistical Models

To estimate the net spillover effect of the social norm nudge on donations, we fit a linear model analogous to specifications (II.2). We use ordinary least-squares estimation (OLS):

$$Donation_i = \alpha^{SE} + \beta^{SE} Norm_i + \varepsilon_i^{SE} \tag{IV.1}$$

$Donation_i$ is a dummy equal to 1 when respondents choose to give and 0 otherwise. We also consider a continuous variable from 0 to 10 for the amount given as another outcome variable. $Norm_i$ is the dummy capturing respondents' allocation to the social norm message.[13]

As we showed in Chapter 2, the effect of the nudge on donations is composed of a behavioural spillover and a direct spillover effect. A naive approach to disentangle these two effects consists of fitting an OLS model analogous to specification (II.4):

$$Donation_i = \tilde{\alpha} + \tilde{\beta}^{BS} FoodChoice_i + \tilde{\beta}^{C} Norm_i + \tilde{\varepsilon}_i \tag{IV.2}$$

$FoodChoice_i$ is a dummy equal to 1 if respondents choose a vegetarian item, 0 otherwise. We also consider the continuous variable capturing the carbon footprint of participants' food choices. Coefficients $\tilde{\beta}^{BS}$ and $\tilde{\beta}^{C}$ capture the behavioural spillover effect and the direct spillover effect, respectively. To tackle potential omitted variable biases, we instrument respondents' food choices

[13] Estimate $\hat{\beta}^{SE}$ corresponds to an intention-to-treat effect. In Appendix 3.A.C, we assess the complier average causal effect by regressing $Donation_i$ on a dummy equal to 1 when participants are shown the social norm message and correctly remember it in the manipulation check, and 0 otherwise. We instrument this dummy by respondents' random allocation to the social norm message.

by $Menu_i$, with the dummy equal to 1 if vegetarian items are salient 0 otherwise. We use a specification analogous to model (II.5), estimated with two-stage least squares (2SLS):[14]

$$\begin{aligned} 1^{st} \text{ stage: } & FoodChoice_i = \alpha + \beta_1 Menu_i + \beta_2 Norm_i + \varepsilon_i \\ 2^{nd} \text{ stage: } & Donation_i = \alpha' + \beta^{BS} \widehat{FoodChoice}_i + \beta^{C} Norm_i + \varepsilon'_i \end{aligned} \tag{IV.3}$$

Using OLS, we estimate the main effect of the nudge on food choices by fitting the first stage of model (IV.3). We use probability linear models whenever the explanatory and outcome variables are binary. We relax the linearity assumption in robustness checks.[15] We also add lasso-selected controls to increase the precision of our estimates (see Appendix 3.A.C, Belloni et al. 2014). We report standard p-values corrected for the false discovery rate (Benjamini and Hochberg, 1995), and p-values computed by re-randomising treatment allocation *à la* Young (2019). We use the latter approach as an extra robustness check to ensure leverage does not drive statistical significance. Finally, we also report adjusted confidence intervals for coefficient β^{BS} using Lee et al. (2022)'s procedure.

V Results

Main effect: Table 3.1 presents the effect of the social norm nudge on food choices.[16] The nudge increases intentions to choose vegetarian food by 6.7 percentage points and reduces the carbon

[14] In Appendix 3.A.C, we test Assumption 1 by interacting food choices with respondents' exposure to the social norm nudge.

[15] We use probit models for specification (IV.1) and when checking for the robustness of the main effect of the social norm on food choices. As an alternative to 2SLS estimation for specification (IV.3), we apply Rivers and Vuong (1988)'s two-step approach and a maximum likelihood estimation approach, as in Evans and Schwab (1995).

[16] Analyses were conducted on R using the package *estimatr* (Blair et al., 2022).

Table 3.1: ATE of the social norm message

Outcome	Chose vegetarian food (binary)	Food choice in kgCO2-eq
Specification	First stage	
Baseline	0.135***	23.400***
	(0.012)	(0.871)
Social norm	0.067***	−2.751**
	(0.016)	(0.928)
	q<0.01	q<0.01
Vegetarian salient	0.115***	−7.875***
	(0.016)	(0.928)
	q<0.01	q<0.01
Num.Obs.	2775	2775
R2	0.025	0.028

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents the effect of the social norm message and the effect of making vegetarian items salient on the likelihood of choosing a vegetarian food item (first column) and on the carbon footprint of food choices (second column). Coefficients are estimated using OLS. Robust standard errors are displayed in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last (q).

footprint of food choices by 11.8%. Results are robust to non-linear probit specifications (see Table 3.A.3 in Appendix 3.A.C).

Side effects: Table 3.2 displays the spillover effects of the social norm nudge on the binary decision to donate (Panel A) and the amount donated (Panel B). The first column contains the results of specification (IV.1) where we regress donations on exposure to the social norm nudge. This coefficient corresponds to the net spillover effect of the nudge. In both panels, these side effects are not significantly different from zero.

The second column displays the results obtained from fitting specification (IV.2). It corresponds to the naive approach for disentangling the direct from the behavioural spillover effects. The third

column displays the results obtained from the two-stage least square regression (IV.3), where we instrument food choices. The effect of the social norm nudge on donations when controlling for food choices is not significantly different from zero, whether or not we instrument food choices. Thus, we do not find evidence of direct spillover effects.

We find suggestive evidence of a positive behavioural spillover effect. The correlation between food choices and donations is statistically significant (column two, specification (IV.2)). When instrumenting food choices, we find that choosing a vegetarian dish increases the likelihood of giving by 36 percentage points. There is no statistically significant effect on the amount donated after p-value correction. We do not observe a statistically significant difference between the instrumented and non-instrumented coefficients. The signs and magnitudes of our estimates are robust when adding controls and when using non-linear specifications (see Tables 3.A.4 and 3.A.5 in Appendix 3.A.C). P-values of randomisation tests indicate that outliers do not drive statistical significance.

Strength of the IV: Our instrumental variable should be relevant, exogenous and homogeneous. Regarding relevance, results in Table 3.1 show a strong and highly significant effect of making vegetarian items salient on the likelihood of choosing vegetarian food. The F statistic of the IV is 53.400 in the binary case and 71.998 when looking at the carbon footprint of food choices. This F-statistic is robust to adding controls (59.432 in the binary case, 64.140 with carbon footprint). This suggests that our instrument is strong (Bound et al., 1995; Staiger and Stock, 1997).

Regarding exogeneity, our instrumental variable is like a default nudge. Respondents must opt out of the selection we show them "by default" to access the full menu. Recent empirical evidence suggests that default nudges affect people's decisions unconsciously (Gärtner, 2018; Van Gestel et al., 2020; Ortmann et al., 2023). This confirms priors in the literature (e.g., see Hansen and

Table 3.2: Total side effects, behavioural spillovers and direct effects

			Panel A		
			Decision to donate (binary)		
Baseline	0.477***	0.443***	0.408***	0.525***	0.578***
	(0.013)	(0.014)	(0.035)	(0.015)	(0.049)
Social norm	0.008	−0.004	−0.016	0.002	−0.006
	(0.019)	(0.019)	(0.022)	(0.019)	(0.020)
	q=0.661	q=0.849	q=0.473	q=0.941	q=0.773
Food choice		0.178***	0.357*	−0.002***	−0.005*
		(0.022)	(0.166)	(0.000)	(0.002)
			[0.004; 0.709]		[-0.010; -0.002]
			q<0.01		q<0.01
R2	0.000	0.022		0.015	
			Panel B		
			Amount donated (in £)		
Baseline	3.309***	3.023***	2.870***	3.695***	3.956***
	(0.108)	(0.111)	(0.272)	(0.124)	(0.389)
Social norm	−0.009	−0.109	−0.163	−0.063	−0.100
	(0.151)	(0.150)	(0.175)	(0.151)	(0.161)
	q=0.952	q=0.473	q=0.338	q=0.680	q=0.528
Food choice		1.490***	2.286	−0.020***	−0.033
		(0.187)	(1.309)	(0.003)	(0.019)
			[-0.495; 5.066]		[-0.072; 0.006]
			q<0.01		q<0.01
R2	0.000	0.024		0.015	
Food choice		Binary	Binary	kgCO2-eq	kgCO2-eq
Specification	OLS	OLS	2SLS	OLS	2SLS
Num.Obs.	2775	2775	2775	2775	2775

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of food choices and the effect of the social norm nudge on the decision to donate (Panel A) and on the amount donated (Panel B). The first column shows the net spillover effect of the social norm nudge on donations. The other columns show estimates of behavioural and direct spillover effects. The second and the fourth columns show OLS estimates of the social norm nudge and food choices on donations. The third and the fifth columns display the 2SLS estimates with food choices instrumented. Robust standard errors are displayed in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). The brackets display confidence intervals adjusted with Lee et al. (2022)'s procedure. P-values of randomisation tests with 5,000 re-sampling are displayed last (q).

Jespersen 2013; Thaler and Sunstein 2009). Thus, the instrumental variable is unlikely to affect donations other than through food choices.

Finally, the homogeneity assumption is violated in the presence of defiers. In our experiment, defiers choose meat-based options when vegetarian items are salient and vice-versa. The behaviour of systematically opposing what is suggested is likely to be orthogonal to pro-environmental beliefs and attitudes. As such, it is unlikely that the effect of choosing vegetarian food on donations for defiers differs from that of choosing vegetarian food on donations for compliers. Angrist et al. (1996) show that biases from violating the homogeneity assumption are small in this case. Angrist et al. (1996) also show that the bias is small when the number of defiers is small. We cannot measure the number of defiers. Nevertheless, we observe that 44% of respondents chose a meat-based item when vegetarian items are salient and vice-versa. This subsample also contains never-takers (always choosing meat) and always-takers (always choosing vegetarian). It seems, therefore, unlikely that the number of defiers is large.

Discussion: Our results suggest that choosing vegetarian food increases people's willingness to do more for the environment, as proxied by our donation task. It is, however, important to note that we only estimate a local average treatment effect. When the profile of compliers differs too much from the rest of the sample, this can affect the external validity of our results. We apply Marbach and Hangartner (2020)'s procedure to compare the profile of compliers with the rest of the sample. We find that, compared to the average of the sample, compliers agree more with the idea that acting against climate change is a moral duty, order food online less frequently and agree less with the idea that British food should be meat-based (see Figures 3.A.7 in Appendix 3.A.D).

Another caveat regards the hypothetical nature of food choices. To mitigate potential biases, we

ask two questions inspired by the literature on willingness-to-pay estimation (Andor et al., 2017; Ready et al., 2010; Champ et al., 2009; Mohammed, 2012). Namely, participants can revise their choices before continuing the survey.[17] Then, we asked them if they would go to a restaurant offering similar food items. Answers are reported on a 5-Likert scale, ranging from "strongly agree" to "strongly disagree". Revising one's choice suggests low confidence in one's preferences, increasing the risk of an intention-behaviour gap. Similarly, not wanting to go to a restaurant offering similar food items would suggest that participants would not make this choice in real life. Only 1.62% of respondents revised their choices, and only 15.56% of them somewhat disagreed or strongly disagreed with going to a restaurant serving the same menus.

Furthermore, we ask two more questions to measure whether respondents' choices truly reflected their preferences. First, we asked them if they felt they had to sacrifice something they liked to choose one of the climate-friendly options.[18] We observe a positive and significant correlation between choosing a meat-based dish and the feeling of having to make a sacrifice to choose vegetarian food. This indicates that meat-based choices reflected a genuine preference for meat. Second, at the end of the experiment, we asked respondents to engage in a thought experiment. Namely, we told them the restaurant could not provide the food they ordered. Instead, they could opt for one of the three options left, all vegetarians. We then asked them how much money they were willing to accept for having to choose one of the vegetarian items.[19] Here again, we observe a positive

[17]The exact wording of the question was: *If we contact the restaurant now to place this order for you, will you be happy for us to proceed?* [a) *Yes, please place this order for me*, b) *No, I would like to change my choice*]

[18]Participants reported on a 5-Likert scale ranging from *"strongly agree"* to *"strongly disagree"* the extent to which they agree with the two following statement: *"Think of your food choice and how you arrived at it. Did you feel that choosing a climate-friendly food item meant you had to sacrifice something you liked?"*

[19]Participants were shown the following message: *"Imagine the restaurant is running out of ingredients. They cannot offer you [participants' food choice]. Instead, it proposes to replace it with one of the following items: [Option 1 - £10; Option 2 - £10; Option 3 - £10]. The restaurant will offer you a refund*

correlation between choosing a meat-based dish and the amount reported. This suggests that the meat choosers were inconvenienced by being forced to choose vegetarian food. We also observe that, among those who chose a vegetarian item in the first place, the amount asked is significantly different from zero. If an experimenter's demand effect purely drove vegetarian choices, people would be indifferent to switching to another vegetarian option. This last result suggests this is not the case (see Table 3.A.18 in Appendix 3.A.D for the regression results).

Still, we cannot exclude that an experimenter's demand effect inflates the effect of the social norm message. Furthermore, the fact that food choices are intentional could induce participants who chose vegetarian food to donate because they could not realise their intentions. Nevertheless, choosing a vegetarian item correlates positively with the feeling of having exerted an effort for the environment, which seems to contradict this interpretation (see Table 3.A.6 in Appendix 3.A.C). Besides, our results align with evidence from field experiments finding positive behavioural spillovers between pro-environmental actions (Alacevich et al., 2021; Comin and Rode, 2023).

Finally, our empirical strategy relied on Assumption 1 (see Section II). We fail to reject this assumption. The interaction between food choices and respondents' exposure to the policy is not significantly different from zero (see Table 3.A.7 in Appendix 3.A.C).

As highlighted in Section II, average treatment effects can hide heterogeneity. We explore the heterogeneity of our causal effects in subsection VI.

based on the price difference. It will also offer you an additional discount for the inconvenience caused. What is the minimum amount of discount that you will be willing to accept to stay and choose one of these items?"

VI Heterogeneity Analysis

How people perceive the social norm nudge might depend on how much they are willing to follow the norm. For instance, telling respondents that more and more people are quitting meat can lead ditherers to change their behaviours, induce convinced meat-eaters to reaffirm their preferences and be ignored by vegetarians with no room for improvement. In other words, the same social norm nudge likely influences different psychological processes for different people. We investigate this heterogeneity by classifying people into different profiles as part of an exploratory analysis.

Training procedure: In a separate survey, we showed 2,782 additional respondents the social norm message and then asked the following question:[20]

Are you trying to change your diet to become more climate-friendly as well?

a) *No, I am not trying now, and I do not intend to try in future*

b) *No, I am not trying now, but I might consider changing my diet to be more-climate-friendly in future*

c) *Yes, I am trying to change my diet now to become more climate-friendly*

d) *Yes, I have already changed my diet to be more climate-friendly*

We assume that asking this question after the social norm message reveals respondents' inclination to follow the norm. It allows us to identify four types: the *transitioned type* is already conforming with the social norm; the *trying type* is inclined to conform; the *hesitant type* considers doing so in the future, and the *unwilling type* does not want to conform. We train a gradient tree-boosting ma-

[20]This question is part of another treatment arm designed for another research project testing if inducing people to think about their choices increases the effectiveness of social norm nudges. See Banerjee and Picard (2023) for more details.

chine learning classifier (GBM) to predict respondents' answers based on attitudinal measures and social-demographic characteristics. Then, we use the algorithm to predict the types of respondents in the *main* sample.[21] As with Random Forest, GBM fits multiple decision trees. Here, each additional decision tree is fitted on the errors made by the previous one (Friedman, 2001). We explain the algorithm in detail in Appendix 3.A.B. To test the robustness of our predictions, we train five other classification algorithms: random forest, a multinomial regression model, an ordered logit model, linear discriminant analysis, and quadratic discriminant analysis.[22] We estimate the average performance of GBM using nested 10×10 folds cross-validation. Overall, GBM performs twice as well as chance. Appendix 3.A.B details the procedure to estimate performances and the predictive power of each predictor. The four classes predicted by GBM are very similar to their counterparts in the training set (see density plots 3.A.8, 3.A.9 and 3.A.10 in Appendix 3.A.D).

Profile of predicted types: Table 3.A.19 in Appendix 3.A.C displays how each type differs from the average for each covariate. Respondents predicted to be *unwilling* to change their diet to follow the norm agree less with the idea that acting against climate change is a moral duty and agree more with the idea that climate change is exaggerated compared to the average. They also know less about the environmental impact of food. *Unwilling* respondents are older, less educated, less likely to live in London and more likely to be male and conservative than the average. Respondents in

[21] Despite having excluded vegan and vegetarian participants, 12,6% of respondents chose the last answer. We see three explanations for this apparent contradiction. First, the screening was based on social demographic information gathered by Prolific, our data provider. As such, people may have changed their diets between when they answered the Prolific questionnaire and when they took our survey. Second, answers can also capture intentions rather than behaviours. Third, the phrasing of this answer could have been perceived as vague enough to allow non-vegetarian participants to select it without contradicting their actual behaviour.

[22] The reader can refer to Gareth et al. (2013) for more information on how these algorithms work.

this group tend to agree more with the idea that typical British food should be meat-based. They report a stronger preference for meat-based food and are less likely to follow a specific diet.

Respondents predicted to be *hesitant* about following the norm live in an area where the unemployment rate is slightly higher and the number of students slightly lower. Their area of residence is also less likely to be rural than the average. These respondents agree less with the idea that acting against climate change is a moral duty. They know less about the environmental impact of food and are less confident in their knowledge of the environmental impact of food. They are younger, less educated, slightly more likely to be female, poorer, and more likely to live in the same area than their area of birth. They also agree more with the idea that British food should be meat-based. They report a stronger preference for meat-based food and order food online more frequently than the average.

Respondents predicted to be *trying* to follow the norm live in an area where the unemployment rate is slightly lower, and the number of students is slightly higher than the average. They agree more with the idea that acting against climate change is a moral duty and agree less that climate change is exaggerated. They know more about the environmental impact of food and are more confident in their knowledge. Respondents in this group are older, more educated, more likely to have moved out of their area of birth, more likely to live in London, richer and less conservative than the average. They also report a lower preference for meat-based food. They are less likely to follow a specific diet and order food online less frequently than the average.

Finally, respondents predicted to have *transitioned* to vegetarian diets are slightly more likely to live in a rural area with a lower share of unemployment. They agree more with the idea that acting against climate change is a moral duty and agree less with the idea that climate change is

exaggerated. They have a better knowledge of the environmental impact of food and are more confident in their knowledge. Respondents in this group are more educated, more likely to be female, to have moved out of their area of birth, and less conservative than the average. They agree less that British food should be meat-based. They also report a lower preference for meat-based food and are more likely to follow a specific diet. They also order food online less frequently than the average of the sample.

In what follows, we estimate the main effect of the social norm message and its crowding-out/in effect for each predicted profile.

Identification strategy: First, we estimate the effect of the social norm nudge on food choices for each predicted type. We use the *unwilling* type as our reference group and fit the following nested probability linear model:[23]

$$FoodChoice_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + u_i \tag{VI.1}$$

$$\Omega_- = \{\text{hesitant, trying, transitioned}\}$$

$$\Omega = \{\text{unwilling, hesitant, trying, transitioned}\}$$

And:

$$\mathbf{1}_k = \begin{cases} 1, \text{ if individual } i \text{ type } k \\ 0, \text{ otherwise} \end{cases}$$

[23]Such a specification is equivalent to fitting four separate linear models for each predicted profile.

Coefficient β_k is the average effect of the social norm nudge conditional on being predicted to be of type k. To estimate the effect of the nudge on donation for each predicted type, we fit the following model:

$$Donation_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + u_i \tag{VI.2}$$

Here again, β_k is the average net spillover effect of the social norm nudge conditional on being predicted to be of type k. To investigate heterogeneity in the direct spillover effect, we fit the following model:

$$Donation_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + \beta_2 \widehat{FoodChoice}_i + \varepsilon_i \tag{VI.3}$$

Here, the coefficient β_k is the direct spillover effect of the social norm message conditional of being predicted to be of type k. $\widehat{FoodChoice}_i$ captures instrumented food choices. To check robustness, we fit these models with the predictions of five other algorithms. Furthermore, we re-estimate our GBM algorithm by over-sampling the *unwilling* and *transitioned* categories that contain fewer observations. We also re-estimate our GBM model by adding income and political beliefs to the set of predictors. We previously excluded these variables as they contain too many missing values. We compute re-randomised p-values to ensure leverage does not drive statistical significance (Young, 2019).

Results: As shown in Figures 3.1, being predicted to follow the norm positively correlates with the likelihood of choosing a vegetarian dish. It is also negatively correlated with the carbon footprint of food choices. Table 3.1 displays the results obtained by fitting equation (VI.1). The social norm nudge only increases the likelihood of choosing vegetarian food for the predicted *trying* ($+10.5$ percentage points). The nudge only reduces the emissions of the predicted *hesitant* (-18

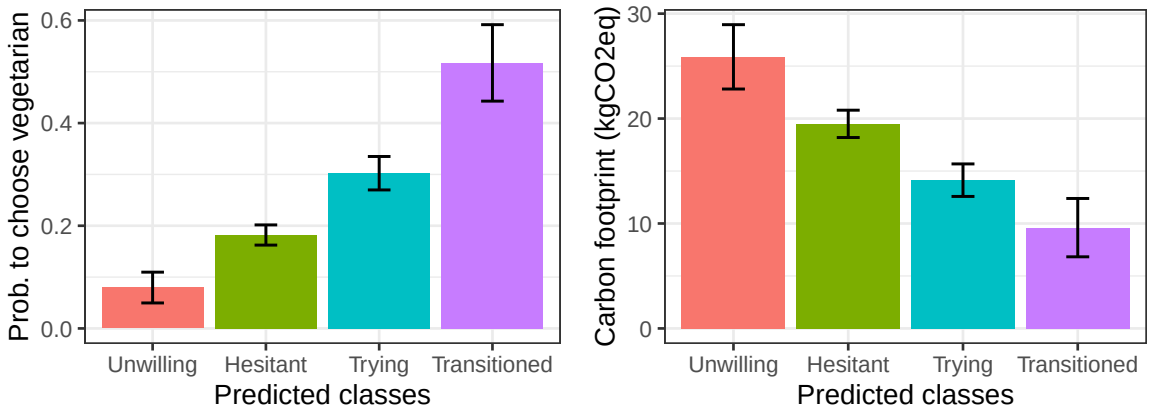


Figure 3.1: Food choices of each predicted type

pp). These results are robust (see Tables 3.A.8 and 3.A.9 in Appendix 3.A.C). Coefficients are of the same sign across all the algorithms and globally of the same order of magnitude. P-values of re-randomisation tests confirm that leverage does not drive statistical significance.

Table 3.2 shows the results of regression (VI.2) in the first two columns and regression (VI.3) in the last four columns. Although not significant after p-value correction, the net spillover effect of the social norm nudge on the amount donated is negative for the *trying type*. When controlling for instrumented food choices, we find that the nudge crowds out the amount they donate by about £0.829. It also crowds out the likelihood of donating by 9.3 percentage points. This negative direct spillover effect is globally robust (see Tables 3.A.12, 3.A.13, 3.A.14 and 3.A.15 in Appendix 3.A.C). Again, p-values of re-randomisation tests confirm that leverage is not driving statistical significance. We also observe suggestive evidence of a positive direct spillover effect among the predicted *unwilling*. However, this effect is not significant after correcting for multiple hypothesis testing.

Discussion: We find that the nudge is effective for the *hesitant* type and the *trying* type. For the *hesitant* type, the nudge decreases the carbon footprint of food choices but does not increase

Table 3.1: Main effect of the social norm message conditional on respondents' types

Specification	Nested OLS model	
Outcome	Chose vegetarian food	Food choice in kgCO2-eq
Unwilling (baseline)	0.091***	24.412***
	(0.023)	(2.122)
Hesitant	0.071***	−2.960
	(0.026)	(2.341)
Trying	0.159***	−9.205***
	(0.031)	(2.401)
Transitioned	0.337***	−13.954***
	(0.064)	(3.086)
Social norm × Unwilling	−0.025	3.076
	(0.030)	(3.138)
	q=0.407	q=0.326
Social norm × Hesitant	0.039	−3.851**
	(0.020)	(1.328)
	q=0.051	q<0.01
Social norm × Trying	0.108***	−2.261
	(0.033)	(1.577)
	q<0.01	q=0.152
Social norm × Transitioned	0.148	−1.426
	(0.077)	(2.899)
	q=0.056	q=0.621
Num.Obs.	2730	2730
R2	0.068	0.032

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of the social norm nudge on the likelihood of choosing vegetarian food (first column) and on the carbon footprint of food choices (second column) for each predicted type. For instance, coefficients labelled "Social norm × Trying" capture the average effect of the nudge on the predicted *trying* (the difference between control units and treatment units in this subsample). Coefficients labelled *Trying* capture the difference between the control units in the *trying* sample with the control units in the *unwilling* sample, our baseline. Robust standard errors are displayed in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.2: direct effects conditional on predicted types.

Specification	Nested OLS model		Nested 2SLS model			
Outcome	Amount (in £)	Decision (binary)	Amount (in £)	Decision (binary)	Amount (in £)	Decision (binary)
Food choice			Chose vegetarian food		Food choice in kgCO2-eq	
Unwilling (baseline)	1.348***	0.207***	1.168***	0.180***	2.060***	0.314***
	(0.234)	(0.032)	(0.259)	(0.035)	(0.509)	(0.065)
Food choice			1.963	0.295*	−0.029	−0.004*
			(1.225)	(0.154)	(0.018)	(0.002)
			q=0.003	q=0.001	q=0.002	q=0.001
Hesitant	1.552***	0.233***	1.413***	0.212***	1.466***	0.220***
	(0.273)	(0.037)	(0.285)	(0.039)	(0.282)	(0.039)
Trying	3.289***	0.419***	2.977***	0.372***	3.021***	0.379***
	(0.316)	(0.040)	(0.371)	(0.047)	(0.361)	(0.046)
Transitioned	3.538***	0.436***	2.876***	0.336***	3.131***	0.374***
	(0.556)	(0.066)	(0.723)	(0.087)	(0.617)	(0.073)
Social norm × Unwilling	0.699	0.073	0.748	0.080	0.789	0.086
	(0.377)	(0.049)	(0.380)	(0.049)	(0.390)	(0.051)
	q=0.064	q=0.134	q=0.049	q=0.107	q=0.040	q=0.085
Social norm × Hesitant	0.135	0.027	0.059	0.016	0.023	0.010
	(0.197)	(0.026)	(0.204)	(0.027)	(0.212)	(0.027)
	q=0.487	q=0.299	q=0.778	q=0.544	q=0.915	q=0.706
Social norm × Trying	−0.664*	−0.070	−0.877**	−0.102**	−0.730*	−0.080*
	(0.304)	(0.036)	(0.331)	(0.039)	(0.307)	(0.036)
	q=0.032	q=0.052	q<0.01	q=0.011	q=0.020	q=0.028
Social norm × Transitioned	−0.366	0.001	−0.658	−0.042	−0.408	−0.005
	(0.652)	(0.075)	(0.691)	(0.081)	(0.644)	(0.073)
	q=0.554	q=0.992	q=0.336	q=0.595	q=0.520	q=0.946
Num.Obs.	2730	2730	2730	2730	2730	2730
R2	0.051	0.051				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table displays the effect of the social norm message on the decision to donate (columns 2, 4 and 6) and on the amount donated (columns 1, 3, and 5) for each predicted type. The first two column shows the net spillover effect of the social norm nudge on donations. The direct spillover effect of the social norm message is then estimated in the other columns, controlling for instrumented food choices. For instance, coefficients labelled "Social norm × Trying" capture the effect of the social norm on the predicted *trying*. Coefficients labelled *Trying* capture the difference between the control units in the *trying* sample with the control units in the *unwilling* sample, our baseline. Robust standard errors are displayed in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last (q).

the uptake of vegetarian options. This apparent paradox might be caused by the predicted *hesitants* switching from carbon-intensive meat options to less intensive meat options. Conversely, the nudge increases the uptake of vegetarian food but does no significantly decrease carbon emissions for the predicted *trying*. This might be because participants classed as *trying* switch from less intensive meat options to vegetarian options. This implies no statistically significant decrease in carbon emissions. Furthermore, it seems that the nudge does not affect the respondents predicted to be *unwilling*. Although the absence of evidence is not evidence of the absence, this null result supports a common assumption in the literature that nudges are ineffective for those unwilling to change (Thaler and Sunstein, 2009). Similarly, the nudge does not significantly alter the choices of the predicted *transitioned* type. The *transitioned* respondents have the highest share of controlled units choosing vegetarian food. As such, it may be that *transitioned* respondents have no room for improvement. This heterogeneity suggests that experimenter demand is unlikely to drive our results provided that respondents' desire to please the experimenter is independent from their predicted types. Indeed, not all types behave in the direction expected by this bias.

We find robust evidence that the social norm message triggers a negative direct spillover effect for the predicted *trying* type. The model in Chapter 2 suggests the *trying* respondents may have treated the social norm message as an extrinsic pressure to choose vegetarian food. This would have induced them to slacken once the extrinsic pressure vanishes. The theoretical framework of Truelove et al. (2014) provides a similar interpretation. For the authors, policies can induce people to act to repair a morally threatened identity. This induces moral licensing once the identity is repaired. Interestingly, the social norm nudge does not produce a similar crowding-out effect for the predicted *hesitants*. Respondents classed as *trying* are more aware of the environmental impact of diets. This can make them more prone to guilt when exposed to our message. We also

find suggestive evidence of a positive direct spillover effect for the predicted *unwilling*. Although not statistically significant, this would explain why the average direct effect is close to zero. The fact that the predicted *unwilling* did not alter their food choices but chose to donate more suggests that this subsample may have engaged in moral cleansing (Sachdeva et al., 2009). Moral cleansing describes pro-social acts undertaken to repair deprecated moral self-worth. However, this interpretation should be considered with caution, given the fragility of this result.

VII Conclusion

In Chapter 2, I model the side effects of soft policies as the sum of two effects. The first effect, referred to as a *behavioural spillover*, emerges when a policy successfully fosters a targeted action. Doing the targeted action encourages or discourages further pro-environmental decisions. Thus, behavioural spillovers capture the effect of doing a first green action on our willingness to do more. I label the second a *direct spillover* effect. It captures the policy's impact on this willingness to engage further. Its sign depends on the nature of the policy used. In our experiment, we dissociate the behavioural spillover from the direct spillover effects. Furthermore, we explore heterogeneity in the effects of the social norm message by identifying profiles expected to respond differently to the nudge.

Our results are consistent with other studies that use an instrumental variable to estimate behavioural spillovers between pro-environmental decisions. Comin and Rode (2023) find that installing solar panels increase support for pro-environmental policies. Alacevich et al. (2021) find that sorting waste leads households to decrease the amount of waste they generate. We find that intentions to choose vegetarian food foster pro-environmental donations. As such, on top of yielding large re-

ductions in greenhouse gas emissions (Green et al., 2015; Riahi et al., 2022), cutting on meat seems to increase people's willingness to do more.

In this regard, using social norm messaging to promote vegetarianism is an effective strategy. However, the effect of this nudge is heterogeneous. We find the social norm nudge to work for people who are predicted to try to change their diets to follow the norm and those who hesitate about doing so. However, we only observe a decrease in the carbon footprint of food choices for the predicted *hesitants*. Besides, the message crowds out the predicted *tryings*' donations. This negative spillover effect outweighs the positive behavioural spillover effect. We do not observe a similar crowding-out effect on the respondents who were predicted to be hesitant. This suggests that policymakers seeking to use social norm nudges to reduce the environmental impact of food choices should target this population segment.

When it comes to increasing the uptake of vegetarian choices, our experimental findings indicate no "free lunch". When the social norm message effectively fosters vegetarian food choices, it is at the cost of crowding out further engagement. This result calls for more empirical evidence on whether other policies yield similar effects.

Informed consent and ethics approval: All participants participated in the study with their informed consent. This study aligned with the London School of Economics and Political Science research ethics guidelines. The study was approved vide reference 38224.

Funding Statement: This study was funded by the Royal Geographic Society (RGS–IBG) Frederick Soddy Postgraduate Award vide reference FSPA 05.21.

Code and Data Availability: The code and data for the analysis are available upon request.

Bibliography

Abrahamse, W. and Steg, L. (2013). Social influence approaches to encourage resource conservation: A meta-analysis. *Global environmental change*, 23(6):1773–1785.

Alacevich, C., Bonev, P., and Söderberg, M. (2021). Pro-environmental interventions and behavioral spillovers: Evidence from organic waste sorting in sweden. *Journal of Environmental Economics and Management*, 108:102470.

Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics*, 95(9-10):1082–1095.

Alt, M., Bruns, H., and DellaValle, N. (2023). The more the better?-synergies of prosocial interventions and effects on behavioral spillovers. *Synergies of prosocial interventions and effects on behavioral spillovers (June 27, 2023)*.

Alt, M. and Gallier, C. (2022). Incentives and intertemporal behavioral spillovers: A two-period experiment on charitable giving. *Journal of Economic Behavior & Organization*, 200:959–972.

Andersson, M. and von Borgstede, C. (2010). Differentiation of determinants of low-cost and high-cost recycling. *Journal of Environmental Psychology*, 30(4):402–408.

Andor, M. A. and Fels, K. M. (2018). Behavioral economics and energy conservation–a systematic review of non-price interventions and their causal effects. *Ecological economics*, 148:178–210.

Andor, M. A., Frondel, M., and Vance, C. (2017). Mitigating hypothetical bias: Evidence on the effects of correctives from a large field study. *Environmental and Resource Economics*, 68(3):777–796.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Banerjee, S. and Picard, J. (2023). Thinking through norms can make them more effective. experimental evidence on reflective climate policies in the uk. *Journal of Behavioral and Experimental Economics*, page 102024.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). *estimatr: Fast Estimators for Design-Based Inference*. https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr.

Blondin, S., Attwood, S., Vennard, D., and Mayneris, V. (2022). Environmental messages promote plant-based food choices: An online restaurant menu study. *World Resources Institute*.

Bonev, P. (2023). Behavioral spillovers. PsyArXiv.

Bonnet, C., Bouamra-Mechemache, Z., Réquillart, V., and Treich, N. (2020). Regulating meat consumption to improve health, the environment and animal welfare. *Food Policy*, 97:101847.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation

when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

Bratt, C. (1999). The impact of norms and assumed consequences on recycling behavior. *Environment and behavior*, 31(5):630–656.

Bryan, C. J., Tipton, E., and Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018). Putting your money where your mouth is: an experimental test of pro-environmental spillover from reducing meat consumption to monetary donations. *Environment and Behavior*, 50(7):723–748.

Carrico, A. R. and Riemer, M. (2011). Motivating energy conservation in the workplace: An evaluation of the use of group-level feedback and peer education. *Journal of environmental psychology*, 31(1):1–13.

Champ, P. A., Moore, R., and Bishop, R. C. (2009). A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review*, 38(2):166–180.

Cialdini, R. B. and Jacobson, R. P. (2021). Influences of social norms on climate change-related behaviors. *Current Opinion in Behavioral Sciences*, 42:1–8.

Clot, S., Della Giusta, M., and Jewell, S. (2022). Once good, always good? testing nudge's spillovers on pro environmental behavior. *Environment and Behavior*, 54(3):655–669.

Comin, D. A. and Rode, J. (2023). Do green users become green voters? Technical report, National Bureau of Economic Research.

Costa, D. L. and Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology:

Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3):680–702.

Evans, W. N. and Schwab, R. M. (1995). Finishing high school and starting college: Do catholic schools make a difference? *The Quarterly Journal of Economics*, 110(4):941–974.

Farrow, K., Grolleau, G., and Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, 140:1–13.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, 101(3):318–22.

Fornara, F., Carrus, G., Passafaro, P., and Bonnes, M. (2011). Distinguishing the sources of normative influence on proenvironmental behaviors: The role of local norms in household waste recycling. *Group Processes & Intergroup Relations*, 14(5):623–635.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gärtner, M. (2018). The prosociality of intuitive decisions depends on the status quo. *Journal of Behavioral and Experimental Economics*, 74:127–138.

Geiger, S. J., Brick, C., Nalborczyk, L., Bosshard, A., and Jostmann, N. B. (2021). More green than gray? toward a sustainable overview of environmental spillover effects: A bayesian meta-analysis. *Journal of Environmental Psychology*, 78:101694.

Goetz, A., Mayr, H., and Schubert, R. (2022). One thing leads to another: Evidence on the scope and persistence of behavioral spillovers. *Available at SSRN 4479949*.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482.

Green, R., Milner, J., Dangour, A. D., Haines, A., Chalabi, Z., Markandya, A., Spadaro, J., and Wilkinson, P. (2015). The potential to reduce greenhouse gas emissions in the uk through healthy and realistic dietary change. *Climatic Change*, 129(1):253–265.

Handgraaf, M. J., De Jeude, M. A. V. L., and Appelt, K. C. (2013). Public praise vs. private pay: Effects of rewards on energy conservation in the workplace. *Ecological Economics*, 86:86–92.

Hansen, P. G. and Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3–28.

Jessoe, K., Lade, G. E., Loge, F., and Spang, E. (2021). Spillovers from behavioral interventions: Experimental evidence from water and energy use. *Journal of the Association of Environmental and Resource Economists*, 8(2):315–346.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Lapinski, M. K., Rimal, R. N., DeVries, R., and Lee, E. L. (2007). The role of group orientation and descriptive norms on water conservation attitudes and behaviors. *Health communication*, 22(2):133–142.

Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. (2022). Valid t-ratio inference for iv. *American Economic Review*, 112(10):3260–3290.

Liu, Y., Kua, H., and Lu, Y. (2021). Spillover effects from energy conservation goal-setting: A field intervention study. *Resources, Conservation and Recycling*, 170:105570.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315.

Marbach, M. and Hangartner, D. (2020). Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis*, 28(3):435–444.

Margetts, E. A. and Kashima, Y. (2017). Spillover between pro-environmental behaviours: The role of resources and perceived similarity. *Journal of Environmental Psychology*, 49:30–42.

Mazar, N. and Zhong, C.-B. (2010). Do green products make us better people? *Psychological science*, 21(4):494–498.

Melnyk, V., van Herpen, E., Trijp, H., et al. (2010). The influence of social norms in consumer decision making: A meta-analysis. *ACR North American Advances*.

Mohammed, E. Y. (2012). Contingent valuation responses and hypothetical bias: mitigation effects of certainty question, cheap talk, and pledging. *Environmental Economics*, 3:62–71.

Nigbur, D., Lyons, E., and Uzzell, D. (2010). Attitudes, norms, identity and environmental behaviour: Using an expanded theory of planned behaviour to predict participation in a kerbside recycling programme. *British journal of social psychology*, 49(2):259–284.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and social psychology bulletin*, 34(7):913–923.

Ortmann, A., Ryvkin, D., Wilkening, T., and Zhang, J. (2023). Defaults and cognitive effort. *Journal of Economic Behavior & Organization*, 212:1–19.

Ready, R. C., Champ, P. A., and Lawton, J. L. (2010). Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Economics*, 86(2):363–381.

Reese, G., Loew, K., and Steffgen, G. (2014). A towel less: Social norms enhance pro-environmental behavior in hotels. *The Journal of Social Psychology*, 154(2):97–100.

Rhodes, N., Shulman, H. C., and McClaran, N. (2020). Changing norms: A meta-analytic integration of research on social norms appeals. *Human Communication Research*, 46(2-3):161–191.

Riahi, K., Schaeffer, R., Arango, J., Calvin, K., Guivarch, C., Hasegawa, T., Jiang, K., Kriegler, E., Matthews, R., Peters, G., et al. (2022). Mitigation pathways compatible with long-term goals. *IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Richter, I., Thøgersen, J., and Klöckner, C. A. (2018). A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability*, 10(8):2848.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366.

Sachdeva, S., Iliev, R., and Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science*, 20(4):523–528.

Salmivaara, L. and Lankoski, L. (2019). Promoting sustainable consumer behaviour through the activation of injunctive social norms: A field experiment in 19 workplace restaurants. *Organization & Environment*, page 1086026619831651.

Scarborough, P., Appleby, P. N., Mizdrak, A., Briggs, A. D., Travis, R. C., Bradbury, K. E., and Key, T. J. (2014). Dietary greenhouse gas emissions of meat-eaters, fish-eaters, vegetarians and vegans in the uk. *Climatic change*, 125(2):179–192.

Schultz, W. P., Khazian, A. M., and Zaleski, A. C. (2008). Using normative social influence to promote conservation among hotel guests. *Social influence*, 3(1):4–23.

Shukla, Skea, Slade, Khourdajie, A., van Diemen, McCollum, Pathak, Some, Vyas, Fradera, Belkacemi, Hasija, Lisboa, Luz, and Malley (2022). Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change. *Cambridge University Press*.

Sparkman, G. and Walton, G. M. (2017). Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychological science*, 28(11):1663–1674.

Sparkman, G., Weitz, E., Robinson, T. N., Malhotra, N., and Walton, G. M. (2020). Developing a scalable dynamic norm menu-based intervention to reduce meat consumption. *Sustainability*, 12(6):2453.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stea, S. and Pickering, G. J. (2019). Optimizing messaging to reduce red meat consumption. *Environmental Communication*, 13(5):633–648.

Stewart, C., Piernas, C., Cook, B., and Jebb, S. A. (2021). Trends in uk meat consumption: analysis of data from years 1–11 (2008–09 to 2018–19) of the national diet and nutrition survey rolling programme. *The Lancet Planetary Health*, 5(10):e699–e708.

Testa, F., Russo, M. V., Cornwell, T. B., McDonald, A., and Reich, B. (2018). Social sustainabil-

ity as buying local: effects of soft policy, meso-level actors, and social influences on purchase intentions. *Journal of Public Policy & Marketing*, 37(1):152–166.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Thøgersen, J. (1999). Spillover processes in the development of a sustainable consumption pattern. *Journal of economic psychology*, 20(1):53–81.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138.

Van Gestel, L., Adriaanse, M., and De Ridder, D. (2020). Do nudges make use of automatic processing? unraveling the effects of a default nudge under type 1 and type 2 processing. *Comprehensive Results in Social Psychology*, pages 1–21.

Van Rookhuijzen, M., De Vet, E., and Adriaanse, M. A. (2021). The effects of nudges: One-shot only? exploring the temporal spillover effects of a default nudge. *Frontiers in Psychology*, 12.

Wenzig, J. and Gruchmann, T. (2018). Consumer preferences for local food: Testing an extended norm taxonomy. *Sustainability*, 10(5):1313.

Wolstenholme, E., Poortinga, W., and Whitmarsh, L. (2020). Two birds, one stone: The effectiveness of health and environmental messages to reduce meat consumption and encourage pro-environmental behavioral spillover. *Frontiers in psychology*, 11:577111.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

3.A Appendix

3.A.A Proofs

Proof. (Proposition 8)

First, model (II.5) can be rewritten in a reduced form as below:

$$x_{2i} = \overline{\alpha} + \overline{\beta}_1 \cdot c_{1i} + \overline{\beta}_2 \cdot \theta_{1i} + \overline{\varepsilon}_i \tag{3.A.1}$$

Where:

$$\overline{\alpha} = \alpha' + \beta^{BS} \cdot \alpha \quad \overline{\beta}_1 = \beta^{BS} \cdot \beta_1 \quad \overline{\beta}_2 = \beta^{C} + \beta^{BS} \cdot \beta_2 \tag{3.A.2}$$

This implies that:

$$\beta^{BS} = \frac{\overline{\beta}_1}{\beta_1} \quad \beta^{C} = \overline{\beta}_2 - \frac{\overline{\beta}_1}{\beta_1}\beta_2 \tag{3.A.3}$$

Using ordinary least square, we can show that the coefficients of model (3.A.1) are equal to:

$$\overline{\beta}_1 = \frac{\sigma_{2c}\sigma_\theta - \sigma_{2\theta}\sigma_{\theta c}}{\sigma_c\sigma_\theta - \sigma^2_{\theta c}} \quad \overline{\beta}_2 = \frac{\sigma_{2\theta}\sigma_c - \sigma_{2c}\sigma_{\theta c}}{\sigma_c\sigma_\theta - \sigma^2_{\theta c}}$$

Where:

$$\sigma_{2\theta} = cov(\boldsymbol{x}_2, \boldsymbol{\theta}_1) \quad \sigma_{\theta c} = cov(\boldsymbol{\theta}_1, \boldsymbol{c}_1)$$

And σ_θ and σ_c denote respectively the variance of $\boldsymbol{\theta}_1$ and $\boldsymbol{c}_1$. Furthermore, using ordinary least square, we can show that the coefficients of model (II.2) are equal to:

$$\beta^{ME} = \frac{\sigma_{1\theta}}{\sigma_\theta} \quad \beta^{SE} = \frac{\sigma_{2\theta}}{\sigma_\theta}$$

Similarly, the coefficients of the first stage of model (II.5) are equal to:

$$\beta_1 = \frac{\sigma_{1c}\sigma_\theta - \sigma_{1\theta}\sigma_{\theta c}}{\sigma_c\sigma_\theta - \sigma_{\theta c}^2} \quad \beta_2 = \frac{\sigma_{1\theta}\sigma_c - \sigma_{1c}\sigma_{\theta c}}{\sigma_c\sigma_\theta - \sigma_{\theta c}^2}$$

Injecting these expressions into expression $\Xi = \beta^C + \beta^{BS} \cdot \beta^{ME} - \beta^{SE}$, one can show that

$$\Xi = 0 \Leftrightarrow \beta^{SE} = \beta^C + \beta^{BS} \cdot \beta^{ME}$$

■

3.A.B Machine Learning Procedure

Gradient tree boosting: Let $\{(x_i, y_i)\}_{i=1}^n$ be the training set with x_i the covariates of observation i and y_i its class. A decision tree is a function F which partitions the space of covariates into K regions $\{R_1, ..., R_K\}$. It predicts a single class $\hat{y}_k$ in each region, for $k \in \{1, ..., K\}$:

$$F(x) = \sum_{k=1}^{K} \hat{y}_k \mathbf{1}_{R_k}(x)$$

Where $\mathbf{1}_{R_k}(x)$ is the indicator function. We want to minimise $L(y, F(x))$ where L is a loss function. This is done in M steps such that at each step m, we fit a function $h_m \in \mathcal{H}$ to the "residuals" of the $m-1$ iteration such that:

$$F_m(x) = F_{m-1}(x) + \upsilon \cdot h_m(x, \delta_{km}) = F_{m-1}(x) + \upsilon \cdot \sum_{k=1}^{K} \delta_{km} a_{km} \mathbf{1}_{R_{km}}(x)$$

Where υ is a shrinkage parameter reducing the speed at which the model is updated. a_{km} is the value predicted by h_m in the region R_{km}. h_m is called a base learner. The scalars δ_{km} are set to minimise the loss function. For $\gamma_{km} = \delta_{km} a_{km}$:

$$\gamma_{km} = \arg\min_{\gamma} \sum_{x_i \in R_{km}} L(y_i, F_{m-1}(x_i) + \gamma)$$

The algorithm is defined as below:

Algorithm:

- Step 0: Choose a constant value γ such that:

$$F_0(x) = \arg\min_{\gamma \in \mathbb{R}} \left[\sum_{i=1}^{n} L(y_i, \gamma) \right]$$

- Step m:

  1. Compute the pseudo-residuals:

$$r_{im}(x_i) = -\frac{\partial L(y_i, F_{m-1}(x_1))}{\partial F_{m-1}(x_1)}$$

  2. Fit a base learner h_m on the pseudo-residuals.
  3. For each partition R_{km}, find the value γ_{km} such that:

$$\gamma_{km} = \arg\min_{\gamma} \sum_{x_i \in R_{km}} L(y_i, F_{m-1}(x_i) + \gamma)$$

4. Update the model:

$$F_m(x) = F_{m-1}(x) + \upsilon \cdot \sum_{k=1}^{K} \gamma_{km} \mathbf{1}_{R_{km}}(x)$$

- Step M: Output function $F_M(x)$.

■

Tuning of hyperparameters: The hyper-parameters we use in this paper are the following:

- The shrinkage parameter υ is set to 0.01. Small values allow an improvement in performance by "forcing" the algorithm to learn slower.
- The bagging fraction is set to 0.5, meaning that 50% of the training observations are randomly drawn at each iteration to train the next tree expansion. Discarding half of the observations reduces the over-fitting risk and improves computation speed.
- The minimal number of observations in each terminal node R_{km} is set to 50 when over-sampling the *unwilling* and the *transitioned* and 10 in the case without oversampling. Splits leading to nodes with numbers of observations below this threshold are discarded. This parameter is tuned using grid search.
- The size of trees K is set to 7 when oversampling the *unwilling* and the *transitioned* and 8 in the case without oversampling. The higher this number, the more numerous the interactions between covariates (the "deeper" the tree). This parameter is tuned using grid search.
- The number of trees fitted M is set to 500 when oversampling and 450 in the case without oversampling. The lower the shrinkage parameter, the higher the number of trees has to be. This parameter is tuned using grid search.

In estimating the performances of GBM, we perform nested 10×10 cross-validation. Namely, we randomly split the training set into ten subsets. First, the algorithm is fitted on nine subsets out of 10. Second, prediction errors are computed by comparing predictions made using the 10^{th} subset data with respondents' actual answers. We repeat the first and the second steps ten times, each time with a new subset, to compute the prediction errors. This process is said to be nested as, at each step, the nine subsets used to fit the model are further split into ten subsets to tune the above hyperparameters. The process to select the hyperparameters maximising the prediction performances of the algorithm is similar to the process described at the beginning of this paragraph to estimate the algorithm's performance. Here, the performance metric used is the average F1 score. Results are similar when using over metrics, such as Cohen's Kappa.

Performances estimation: In total, we considered three different metrics to estimate the performances of GBM. First, for each type, we compute the share of individuals predicted to be of type i that are actually of type i. This measure is called *precision*. It tells us about the "purity" of our predicted classes. Precision should be higher than the share of respondents in type i over the total number of respondents to perform better than chance.[24] Here, GBM performs better than chance for each type and, on average, 1.9 times better than chance across all types (see Table 3.A.1).

Nevertheless, one can achieve high precision by excluding observations that are hard to predict. This is why we also look at *recall*, a measure of performance obtained by computing the proportion of individuals of type i correctly identified as being type i. This measure tells us about how "exhaustive" each predicted class is. With four types, a ratio above 25% indicates that the algorithm is

[24]With four types, an algorithm doing as good as chance would produce a rate of true positives for type i to be $\frac{n_i}{4}$, where n_i is the number of individuals in type i. The rate of false positives would be $\frac{n-n_i}{4}$ where n is the total number of respondents. Thus precision is equal to $\frac{\frac{n_i}{4}}{\frac{n_i}{4}+\frac{n-n_i}{4}} = \frac{n_i}{n}$.

performing better than chance.[25] The average recall rate of GBM is higher than 25% for the *unwilling type*, *hesitant type* and *trying type*, and slightly higher for the *transitioned type*. On average, GBM performs 1.6 times better than chance (see Table 3.A.1).

Ideally, we would like an algorithm yielding predicted types that are both "pure" and "exhaustive". The F1 score is a measure encompassing these two aspects. It is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

For our algorithm to perform better than chance, the average F1 score in each type should be higher than the thresholds displayed in Table 3.A.1.[26] Results in Table 3.A.1 confirm that GBM does better than chance for all types and on average 1.7 times better than chance across all types.

A closer look at Table 3.A.1 reveals that GBM over-classifies respondents as *hesitant* and under-classifies respondents as *unwilling* and *transitioned*. This explains the higher recall rate of the *hesitant* type and the higher precision rate of the *unwilling* and *transitioned* types. To correct this bias, we train another GBM algorithm where, this time, we over-sample the *unwilling* and *transitioned* types in the training set. Namely, we increase the sizes of these two sub-samples by drawing new observations with replacements from the original sub-samples. The new algorithm now seems to under-predict respondents to be *hesitant* in favour of the *unwilling* and *transitioned*. Although not statistically significant, over-sampling improves the overall recall rate of the model at the expense of precision and the F1 score. Furthermore, the relative sizes of each predicted type seem closer to these of the training set when over-sampling as measured by the Euclidian distance, although,

[25] With an algorithm doing as good as chance, the rate of true positives is $\frac{n_i}{4}$, where n_i is the number of individuals in type i. The rate of false negatives is $\frac{3n_i}{4}$. Thus recall is equal to $\frac{\frac{n_i}{4}}{\frac{n_i}{4}+\frac{3n_i}{4}} = \frac{1}{4}$.

[26] The minimum thresholds for the precision and the recall of an algorithm doing as good as chance are respectively $\frac{n_i}{n}$, and $\frac{1}{4}$. As such, the F1 score of this algorithm: $2 \times \frac{\frac{n_i}{n} \times \frac{1}{4}}{\frac{n_i}{n}+\frac{1}{4}} = \frac{2 \times n_i}{n_i \times 4 + n}$.

Table 3.A.1: Estimated performance of GBM

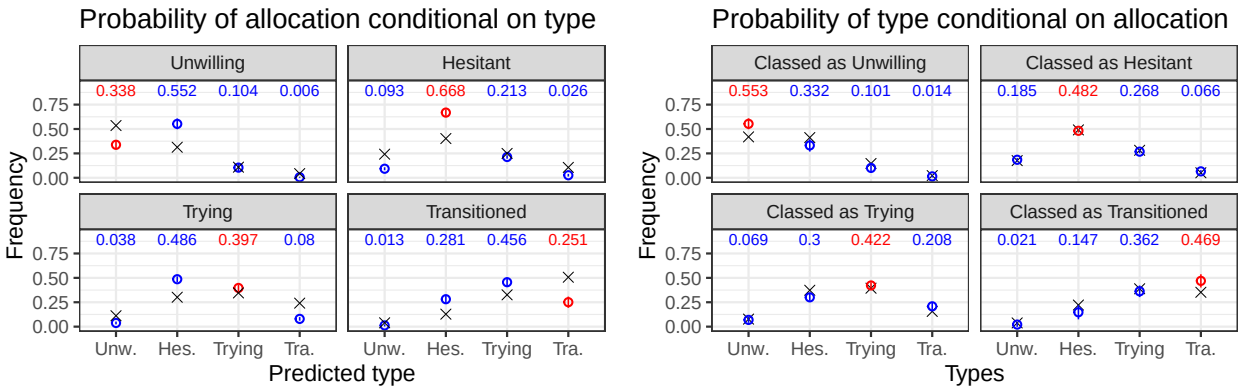
	Relative size of each type		Precision (in %)		Recall (in %)		F1 score (in %)	
	Training set	Predicted	Threshold	GBM	Threshold	GBM	Threshold	GBM
Unwilling	18.3	12 [24.2]	0.183	0.553 (0.03) [0.419*** (0.02)]	0.25	0.338 (0.02) [0.535*** (0.03)]	0.212	0.417 (0.02) [0.469 (0.02)]
Hesitant	39.4	53.5 [30.1]	0.394	0.482 (0.01) [0.491 (0.02)]	0.25	0.668 (0.02) [0.402*** (0.02)]	0.306	0.559 (0.01) [0.440*** (0.02)]
Trying	29.7	27.9 [27.0]	0.297	0.422 (0.02) [0.394 (0.01)]	0.25	0.397 (0.02) [0.346 (0.02)]	0.297	0.408 (0.02) [0.366 (0.01)]
Transitioned	12.6	6.6 [18.7]	0.126	0.469 (0.03) [0.352** (0.03))]	0.25	0.251 (0.03) [0.507*** (0.03)]	0.167	0.320 (0.03) [0.412** (0.03)]
	Euclidean distance (cross-validated)		Average					
	/	0.18 (0.02) [0.14 (0.01)]	0.25	0.481 (0.04) [0.414 (0.03)]	0.25	0.413 (0.03) [0.447 (0.03)]	0.246	0.426 (0.03) [0.421 (0.03)]

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: The columns labelled "threshold" contain the minimum performance threshold for each metric. Below these thresholds, GBM does worse than chance. Values in brackets correspond to the performance of GBM after over-sampling the *unwilling* and *transitioned* types. Stars indicate the results of simple t-tests to assess whether performances after re-sampling differ significantly from before.

here again, the difference is not statistically significant (see Table 3.A.1).

A last performance check consists of looking at whether the miss-classification errors of our two extreme types (*transitioned type* and *unwilling type*) occur in "adjacent" types. Indeed, one would prefer to avoid using an algorithm that jumbles the *transitioned* and the *unwilling* types. Here, we estimate two sets of probabilities: the probability of being classified as type j whilst being of type i, $P(class = i|type = j)$, and the probability of being of type j whilst being classified as type i, $P(type = i|class = j)$. The first set of probabilities measures the model's performance *ex-ante*: e.g., what is the probability that I will be classified in class k given my type? Symmetrically, the second set of probabilities gives us a measure of the model's performance *ex-post*: e.g., what is the probability that I am of type k given how I was classified. The left panel of Figure 3.A.1 presents the estimated first set of probabilities, whilst the right panel presents the second. Overall, misclassification errors occur less often in non-adjacent categories. Furthermore, the left panel of Figure 3.A.1 indicates that over-sampling has increased the ability of the algorithm to correctly identify



(a) Miss-classification by type

(b) Miss-classification by predicted type

Figure 3.A.1: Frequency of miss-classification errors

Note: Red dots correspond to the performance metric recall and precision on the right and left panels. Black crosses correspond to the estimates after re-sampling. 95% confidence intervals are represented by the vertical bars.

the *unwilling* and *transitioned* at the expense of its ability to identify the *hesitant* correctly. However, over-sampling has also slightly decreased its ability to produce pure predicted classes, as suggested by the right panel of Figure 3.A.1. In other words, over-sampling seems to make GBM better at detecting the *unwilling* and *transitioned* types by simply increasing the number of respondents classified in these categories. We use the predictions obtained without over-sampling to carry out the main analysis.

Predictive power of covariates: The eighteen predictors used to train the GBM algorithm can be broadly grouped into four categories displayed in Table 3.A.2. First, sociological and economic characteristics of the area of residence of respondents account for 35.67% of the relative influence of the predictors. We construct these variables by merging information from the UK 2011 census data provided by the Office for National Statistics and the postcode respondents reported. Second, respondents' attitudes towards climate change and their knowledge of the environmental impact of food represent 33.87% of the relative influence of all the predictors. In this category, respondents'

belief about whether acting against climate change is a moral duty has the greatest influence. In itself, it accounts for about 18% of the relative influence of the eighteen variables. Third, respondents' social-demographic characteristics represent 16.02% of the total influence of the predictors, followed by measures of respondents' food preferences that account for 14.43% of this influence. We excluded two predictors that contained too many missing values: respondents' income and political beliefs. We include them back when testing for the robustness of our results. Readers interested in the influence of each predictor on the likelihood of being classified in one of the four types can refer to the partial dependency plots displayed in Figures 3.A.11, 3.A.12, and 3.A.13 in Appendix 3.A.D.

Table 3.A.2: Relative influence of each predictor

Category	Predictors	Relative influence (in %)
Social-demographics of residence area	Share of unemployed among actives in residence area	11.53
	Share of students in residence area	11.16
	Proportion of rural areas in residence area	7.00
	Share of UK/EU population in residence area	5.98
Belief and knowledge on the environment	Belief moral duty to act against climate change	17.93
	Knowledge of the carbon footprint of food	7.60
	Belief climate change is exaggerated	5.37
	Confidence in one's knowledge	2.97
Personal social-demographics	Age	8.76
	Education	3.90
	Sex	1.50
	Moved out of birth area	1.04
	Caucasian	0.43
	Live in London	0.40
Food preferences	Belief British food should be meat-based	4.63
	Online food ordering habits	3.95
	Preference for meat-based food	3.37
	Follows a specific diet	2.48

3.A.C Robustness Checks

Table 3.A.3: Robustness checks of ATEs of the social norm message

Outcome	Chose vegetarian food (binary)					Food choice in kgCO2-eq		
Specification	ITT with controls	CACE w/o controls	CACE with controls	Logit w/o controls	Logit with controls	ITT with controls	CACE w/o controls	CACE with controls
Baseline	0.258***	0.136***	0.256***			22.874***	23.364***	22.889***
	(0.064)	(0.012)	(0.064)			(4.295)	(0.864)	(4.297)
Social norm	0.057***	0.083***	0.070***	0.067***	0.056***	−2.211**	−3.375**	−2.698**
	(0.016)	(0.019)	(0.020)	(0.016)	(0.016)	(0.955)	(1.137)	(1.164)
Vegetarian salient	0.124***	0.113***	0.123***	0.114***	0.121***	−7.697***	−7.802***	−7.639***
	(0.016)	(0.016)	(0.016)	(0.015)	(0.016)	(0.961)	(0.929)	(0.961)
Num.Obs.	2454	2775	2454	2775	2454	2453	2775	2453
R2	0.113					0.081		

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Effect of the social norm nudge and default menu allocation on food choices when controls are added and with non-linear probit estimation. We use the following lasso-selected controls to increase the precision of the estimates: level of hunger, how busy one is at the moment of taking the survey, knowledge of the environmental impact of food and confidence in one's knowledge, online food ordering frequency, belief that British food should be meat-based, preference for meat-based food, belief that climate change is exaggerated, belief that acting against climate change is a moral duty, income, sex, political orientation, education level and a dummy capturing the visual aspect of the menu. Robust standard errors are displayed in parentheses.

Table 3.A.4: Robustness checks of behavioural and direct spillover effects I

Outcome	Decision to donate (binary)						
Food choice	Chose vegetarian food (binary)				Food choice in kgCO2-eq		
Specification	2SLS with controls	Probit w/o controls	Probit with controls	MLE w/o controls	2SLS with controls	Probit w/o controls	Probit with controls
Baseline	0.110				0.292***		
	(0.069)				(0.075)		
Food choice	0.329**	0.355**	0.323**	0.351***	−0.005**	−0.005**	−0.005**
	(0.156)	(0.163)	(0.153)	(0.017)	(0.002)	(0.002)	(0.002)
Social norm	−0.018	−0.016	−0.017	-0.016	−0.010	−0.006	−0.010
	(0.021)	(0.022)	(0.021)	(0.018)	(0.020)	(0.020)	(0.020)
Num.Obs.	2603	2775	2603	2775	2603	2775	2603

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Behavioural and direct spillover effects of the social norm message on the decision to donate. We use the following lasso-selected controls to increase the precision of the estimates: belief that British food should be meat-based, preference for meat-based food, belief that climate change is exaggerated, belief that acting against climate change is a moral duty and political orientation. The second, third, sixth and seventh columns contain estimates obtained with a two-stage Rivers and Vuong (1988) probit estimation. The fourth column contains estimates obtained with maximum likelihood estimation *à la* Evans and Schwab (1995) with standard errors obtained using the delta method. The other standard errors are robust and displayed in parentheses.

Table 3.A.5: Robustness checks of behavioural and direct spillover effects II

Outcome	Amount donated (in £)	
Food choice	Binary food choice	Food choice in kgCO2-eq
Specification	2SLS with controls	2SLS with controls
Baseline	−0.002	1.121*
	(0.570)	(0.615)
Food choice	2.058*	−0.033*
	(1.235)	(0.020)
Social norm	−0.188	−0.136
	(0.172)	(0.160)
Num.Obs.	2602	2602

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Behavioural and direct spillover effects of the social norm message on the amount donated. We use the following lasso-selected controls to increase the precision of the estimates: belief that British food should be meat-based, preference for meat-based food, belief that climate change is exaggerated, belief that acting against climate change is a moral duty, age and political orientation. Robust standard errors are displayed in parentheses.

Table 3.A.6: Effect of food choices on perception of effort for the environment

Outcome	Perception of effort			
Food choice	Binary		In kgCO2-eq	
Specification	OLS w/o controls	OLS with controls	OLS w/o controls	OLS with controls
Baseline	2.973***	1.822***	3.153***	2.012***
	(0.020)	(0.162)	(0.021)	(0.161)
Food choice	0.310***	0.269***	−0.006***	−0.005***
	(0.042)	(0.046)	(0.001)	(0.001)
Num.Obs.	2775	2453	2775	2453
R2	0.020	0.111	0.026	0.116

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Effect of food choices on the perception of having exerted an effort for the environment. We control for the default menus, exposure to the social norm message, the appearance of menus, self-reported level of hunger and hurry, knowledge of the environmental impact of food and confidence in one's knowledge, frequency of food online delivery, income, age, education, belief that British food should be meat-based, preference for meat-based food, belief that climate change is exaggerated, belief that acting against climate change is a moral duty, gender, and political orientation. Robust standard errors are displayed in parentheses.

Table 3.A.7: Test of Assumption 1

Outcome	Amount donated (in £)		Decision to donate (binary)		Amount donated (in £)		Decision to donate (binary)	
Food choice	Binary	In kgCO2-eq	Binary	In kgCO2-eq	Binary	In kgCO2-eq	Binary	In kgCO2-eq
Specification	OLS				2SLS			
Baseline	3.064***	3.728***	0.450***	0.522***	2.795***	4.128***	0.428***	0.555***
	(0.117)	(0.138)	(0.015)	(0.017)	(0.414)	(0.651)	(0.052)	(0.081)
Food choice	1.274***	−0.022***	0.140***	−0.002***	2.678	−0.042	0.254	−0.004
	(0.285)	(0.004)	(0.034)	(0.001)	(2.096)	(0.033)	(0.260)	(0.004)
Social norm	−0.195	−0.128	−0.019	0.008	−0.004	−0.377	−0.057	0.031
	(0.167)	(0.190)	(0.021)	(0.023)	(0.607)	(0.763)	(0.077)	(0.095)
Food choice × Social norm	0.389	0.004	0.069	0.000	−0.711	0.015	0.186	−0.002
	(0.377)	(0.006)	(0.045)	(0.001)	(2.672)	(0.040)	(0.337)	(0.005)
Num.Obs.	2775	2775	2775	2775	2775	2775	2775	2775
R2	0.025	0.015	0.023	0.015				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Saturated model allowing to test Assumption 1. For the last four columns, we instrument the variables capturing food choices by the dummy equal to 1 when vegetarian choices are salient and 0 otherwise. We instrument the variables corresponding to the interaction between food choices and the social norm nudge by the dummy capturing whether vegetarian items are salient, interacted with the social norm nudge. Robust standard errors are displayed in parentheses.

Table 3.A.8: Robustness checks of the ATEs of the social norm message conditional on predicted classes I

Specification	Nested OLS model						
Outcome	Chose vegetarian food (binary)						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	0.076***	0.079***	0.073***	0.056***	0.071***	0.067***	0.114***
	(0.015)	(0.022)	(0.020)	(0.018)	(0.021)	(0.018)	(0.021)
Hesitant	0.103***	0.098***	0.098***	0.094***	0.073***	0.069***	0.038
	(0.024)	(0.027)	(0.025)	(0.022)	(0.024)	(0.022)	(0.025)
Trying	0.116***	0.160***	0.162***	0.238***	0.215***	0.219***	0.125***
	(0.025)	(0.031)	(0.030)	(0.029)	(0.031)	(0.029)	(0.032)
Transitioned	0.301***	0.421***	0.427***	0.492***	0.444***	0.451***	0.306***
	(0.035)	(0.067)	(0.065)	(0.080)	(0.064)	(0.058)	(0.046)
Social norm × Unwilling	0.042*	−0.018	0.016	0.013	−0.001	0.003	0.001
	(0.023)	(0.030)	(0.031)	(0.028)	(0.030)	(0.026)	(0.030)
	q=0.070	q=0.554	q=0.599	q=0.654	q=0.981	q=0.911	q=0.970
Social norm × Hesitant	0.027	0.037	0.029	0.056***	0.061***	0.073***	0.076***
	(0.028)	(0.022)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)
	q=0.335	q=0.102	q=0.136	q<0.01	q<0.01	q<0.01	q<0.01
Social norm × Trying	0.106***	0.095***	0.124***	0.076**	0.069**	0.061*	0.088**
	(0.031)	(0.034)	(0.033)	(0.033)	(0.034)	(0.034)	(0.036)
	q<0.01	q<0.01	q<0.01	q=0.022	q=0.045	q=0.071	q=0.018
Social norm × Transitioned	0.062	0.061	0.045	0.106	0.075	0.048	0.025
	(0.043)	(0.081)	(0.080)	(0.102)	(0.079)	(0.072)	(0.056)
	q=0.148	q=0.445	q=0.580	q=0.299	q=0.356	q=0.505	q=0.645
Num.Obs.	2730	2431	2730	2730	2730	2730	2730
R2	0.069	0.066	0.065	0.078	0.081	0.088	0.055

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: ATEs of the social norm message on the decision to choose vegetarian food. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.9: Robustness checks of the ATEs of the social norm message conditional on predicted classes II

Specification	Nested OLS model						
Outcome	Food choice in kgCO2-eq						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	25.004***	27.100***	25.206***	26.367***	25.826***	26.597***	24.119***
	(1.515)	(2.259)	(2.134)	(2.195)	(2.201)	(2.019)	(1.801)
Hesitant	−2.675	−6.831***	−4.521*	−4.858**	−4.446*	−5.368**	−3.831*
	(2.029)	(2.488)	(2.340)	(2.397)	(2.404)	(2.255)	(2.059)
Trying	−9.046***	−12.064***	−8.822***	−12.683***	−11.082***	−11.470***	−5.995***
	(1.908)	(2.538)	(2.447)	(2.446)	(2.475)	(2.315)	(2.278)
Transitioned	−12.592***	−19.403***	−16.847***	−16.701***	−16.733***	−18.594***	−13.532***
	(2.034)	(2.910)	(2.922)	(3.525)	(3.068)	(2.680)	(2.396)
Social norm × Unwilling	−2.659	1.434	0.978	0.201	1.275	−0.814	−0.584
	(2.114)	(3.370)	(3.165)	(3.204)	(3.233)	(2.865)	(2.521)
	q=0.208	q=0.672	q=0.755	q=0.950	q=0.700	q=0.772	q=0.823
Social norm × Hesitant	−3.822**	−3.069**	−3.189**	−3.524***	−3.651***	−3.645***	−2.868**
	(1.830)	(1.408)	(1.302)	(1.308)	(1.306)	(1.357)	(1.373)
	q=0.035	q=0.029	q=0.014	q<0.01	q<0.01	q<0.01	q=0.039
Social norm × Trying	−0.535	−1.912	−2.448	−1.168	−1.382	−1.889	−4.907***
	(1.694)	(1.622)	(1.668)	(1.500)	(1.610)	(1.604)	(1.863)
	q=0.753	q=0.241	q=0.151	q=0.428	q=0.390	q=0.238	q=0.010
Social norm × Transitioned	−2.844*	0.157	0.090	−3.958	−2.087	0.719	0.886
	(1.717)	(2.457)	(2.642)	(3.288)	(2.633)	(2.435)	(2.231)
	q=0.099	q=0.953	q=0.976	q=0.241	q=0.440	q=0.773	q=0.679
Num.Obs.	2730	2431	2730	2730	2730	2730	2730
R2	0.037	0.040	0.028	0.037	0.035	0.037	0.025

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: ATEs of the social norm message on the carbon footprint of respondents' food. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.10: Robustness checks of the side effects of the social norm message conditional on predicted classes I

Specification	Nested OLS model						
Outcome	Amount donated (in £)						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	1.557***	1.477***	1.273***	1.327***	1.383***	1.161***	1.747***
	(0.174)	(0.256)	(0.232)	(0.245)	(0.224)	(0.224)	(0.215)
Hesitant	1.458***	1.491***	1.649***	1.605***	1.605***	1.692***	1.435***
	(0.256)	(0.298)	(0.270)	(0.281)	(0.266)	(0.263)	(0.263)
Trying	2.612***	3.109***	3.378***	3.259***	3.282***	3.761***	2.409***
	(0.274)	(0.338)	(0.316)	(0.330)	(0.311)	(0.308)	(0.321)
Transitioned	3.340***	3.633***	3.924***	3.835***	2.858***	3.672***	2.946***
	(0.326)	(0.585)	(0.581)	(0.570)	(0.507)	(0.690)	(0.415)
Social norm × Unwilling	0.638**	0.287	0.693*	0.316	0.519	0.811**	0.313
	(0.263)	(0.392)	(0.367)	(0.364)	(0.336)	(0.358)	(0.313)
	q=0.014	q=0.483	q=0.060	q=0.375	q=0.111	q=0.022	q=0.321
Social norm × Hesitant	0.110	0.181	0.161	0.132	0.167	0.181	0.052
	(0.265)	(0.216)	(0.198)	(0.194)	(0.204)	(0.195)	(0.214)
	q=0.671	q=0.417	q=0.414	q=0.493	q=0.409	q=0.363	q=0.807
Social norm × Trying	−0.542*	−0.528*	−0.689**	−0.510	−0.655**	−0.838***	−0.192
	(0.302)	(0.314)	(0.303)	(0.310)	(0.308)	(0.295)	(0.336)
	q=0.071	q=0.098	q=0.021	q=0.100	q=0.034	q<0.01	q=0.566
Social norm × Transitioned	−0.518	−0.385	−0.920	−0.625	−0.197	−0.410	−0.698
	(0.371)	(0.672)	(0.681)	(0.676)	(0.592)	(0.864)	(0.470)
	q=0.166	q=0.565	q=0.174	q=0.359	q=0.741	q=0.636	q=0.140
Num.Obs.	2730	2431	2730	2730	2730	2730	2730
R2	0.062	0.053	0.053	0.055	0.051	0.062	0.038

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Total side effects of the social norm message on the amount donated. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.11: Robustness checks of the side effects of the social norm message conditional on predicted classes II

Specification	Nested OLS model						
Outcome	Decision to donate (binary)						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	0.239***	0.219***	0.194***	0.199***	0.202***	0.180***	0.279***
	(0.024)	(0.034)	(0.031)	(0.032)	(0.029)	(0.030)	(0.030)
Hesitant	0.223***	0.224***	0.248***	0.251***	0.257***	0.255***	0.192***
	(0.034)	(0.039)	(0.036)	(0.037)	(0.035)	(0.035)	(0.035)
Trying	0.351***	0.408***	0.439***	0.411***	0.418***	0.478***	0.286***
	(0.034)	(0.042)	(0.039)	(0.041)	(0.038)	(0.039)	(0.041)
Transitioned	0.409***	0.453***	0.473***	0.478***	0.388***	0.463***	0.350***
	(0.039)	(0.068)	(0.066)	(0.066)	(0.062)	(0.081)	(0.050)
Social norm × Unwilling	0.084**	0.026	0.087*	0.046	0.075*	0.121**	0.035
	(0.035)	(0.051)	(0.048)	(0.048)	(0.044)	(0.049)	(0.042)
	q=0.016	q=0.614	q=0.071	q=0.340	q=0.092	q=0.015	q=0.401
Social norm × Hesitant	0.028	0.031	0.029	0.020	0.021	0.027	0.011
	(0.035)	(0.028)	(0.026)	(0.026)	(0.027)	(0.026)	(0.027)
	q=0.433	q=0.271	q=0.257	q=0.454	q=0.435	q=0.284	q=0.691
Social norm × Trying	−0.074**	−0.057	−0.074**	−0.038	−0.057	−0.091***	−0.009
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.034)	(0.040)
	q=0.039	q=0.120	q=0.045	q=0.291	q=0.103	q<0.01	q=0.822
Social norm × Transitioned	−0.022	0.012	−0.073	−0.055	0.003	0.011	−0.033
	(0.042)	(0.076)	(0.076)	(0.076)	(0.071)	(0.100)	(0.055)
	q=0.604	q=0.876	q=0.325	q=0.481	q=0.972	q=0.910	q=0.553
Num.Obs.	2730	2431	2730	2730	2730	2730	2730
R2	0.066	0.058	0.052	0.052	0.052	0.058	0.037

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Total side effects of the social norm message on the decision to donate. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.12: Robustness checks of the direct spillover effect conditional on predicted classes I

Specification	Nested 2SLS model						
Outcome	Amount donated (in £)						
Food choice	Chose vegetarian food (binary)						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	1.424***	1.303***	1.137***	1.054***	1.198***	1.264***	1.529***
	(0.202)	(0.275)	(0.247)	(0.238)	(0.265)	(0.241)	(0.262)
Hesitant	1.280***	1.277***	1.466***	1.511***	1.472***	1.483***	1.362***
	(0.285)	(0.321)	(0.297)	(0.281)	(0.293)	(0.280)	(0.267)
Trying	2.411***	2.759***	3.076***	3.302***	2.865***	2.895***	2.170***
	(0.312)	(0.390)	(0.376)	(0.420)	(0.422)	(0.417)	(0.360)
Transitioned	2.817***	2.713***	3.128***	2.724***	3.020***	2.059***	2.360***
	(0.515)	(0.800)	(0.805)	(0.931)	(0.807)	(0.768)	(0.577)
Social norm × Unwilling	0.565**	0.327	0.663*	0.787**	0.318	0.513	0.310
	(0.268)	(0.399)	(0.370)	(0.358)	(0.368)	(0.340)	(0.316)
	q=0.034	q=0.411	q=0.075	q=0.028	q=0.393	q=0.129	q=0.327
Social norm × Hesitant	0.063	0.101	0.106	0.074	0.020	0.038	−0.093
	(0.267)	(0.221)	(0.201)	(0.210)	(0.211)	(0.227)	(0.236)
	q=0.803	q=0.647	q=0.591	q=0.730	q=0.923	q=0.866	q=0.695
Social norm × Trying	−0.726**	−0.737**	−0.921***	−0.983***	−0.636**	−0.763**	−0.359
	(0.334)	(0.334)	(0.341)	(0.309)	(0.322)	(0.316)	(0.356)
	q=0.030	q=0.027	q<0.01	q<0.01	q=0.044	q=0.014	q=0.298
Social norm × Transitioned	−0.625	−0.519	−1.003	−0.615	−0.762	−0.282	−0.746
	(0.381)	(0.697)	(0.695)	(0.866)	(0.679)	(0.586)	(0.467)
	q=0.097	q=0.445	q=0.153	q=0.485	q=0.264	q=0.620	q=0.105
Food choice	1.735	2.186*	1.863	1.927	1.835	1.771	1.915
	(1.317)	(1.241)	(1.268)	(1.253)	(1.270)	(1.284)	(1.299)
	q=0.011	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Direct spillover effect of the social norm message on the amount donated. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.13: Robustness checks of the direct spillover effect conditional on predicted classes II

Specification	Nested 2SLS model						
Outcome	Amount donated (in £)						
Food choice	Food choice in kgCO2-eq						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	2.170***	2.437***	1.954***	1.900***	2.016***	2.066***	2.408***
	(0.499)	(0.608)	(0.523)	(0.525)	(0.534)	(0.544)	(0.500)
Hesitant	1.393***	1.249***	1.527***	1.556***	1.487***	1.467***	1.330***
	(0.261)	(0.332)	(0.285)	(0.278)	(0.296)	(0.285)	(0.273)
Trying	2.390***	2.682***	3.140***	3.406***	2.964***	2.988***	2.245***
	(0.321)	(0.417)	(0.356)	(0.382)	(0.386)	(0.375)	(0.338)
Transitioned	3.031***	2.945***	3.469***	3.204***	3.388***	2.380***	2.575***
	(0.402)	(0.695)	(0.658)	(0.761)	(0.656)	(0.616)	(0.482)
Social norm × Unwilling	0.573**	0.337	0.719*	0.817**	0.350	0.498	0.297
	(0.272)	(0.407)	(0.375)	(0.368)	(0.382)	(0.344)	(0.320)
	q=0.034	q=0.396	q=0.053	q=0.026	q=0.353	q=0.145	q=0.348
Social norm × Hesitant	0.016	0.072	0.074	0.082	0.034	0.074	−0.027
	(0.276)	(0.226)	(0.207)	(0.208)	(0.207)	(0.216)	(0.222)
	q=0.953	q=0.747	q=0.709	q=0.695	q=0.871	q=0.729	q=0.906
Social norm × Trying	−0.555*	−0.596*	−0.755**	−0.870***	−0.547*	−0.703**	−0.326
	(0.301)	(0.318)	(0.305)	(0.295)	(0.310)	(0.309)	(0.348)
	q=0.066	q=0.062	q=0.015	q<0.01	q=0.081	q=0.023	q=0.345
Social norm × Transitioned	−0.587	−0.379	−0.917	−0.521	−0.681	−0.178	−0.674
	(0.371)	(0.661)	(0.675)	(0.873)	(0.677)	(0.592)	(0.462)
	q=0.111	q=0.576	q=0.169	q=0.559	q=0.302	q=0.768	q=0.146
Food choice	−0.025	−0.035*	−0.027	−0.028	−0.027	−0.026	−0.027
	(0.019)	(0.020)	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)
	q=0.012	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: Direct spillover effect of the social norm message on the amount donated. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.14: Robustness checks of the direct spillover effect conditional on predicted classes III

Specification	Nested 2SLS model						
Outcome	Decision to donate (binary)						
Food choice	Chose vegetarian food (binary)						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	0.218***	0.191***	0.174***	0.164***	0.179***	0.184***	0.246***
	(0.027)	(0.037)	(0.033)	(0.032)	(0.034)	(0.031)	(0.036)
Hesitant	0.196***	0.191***	0.221***	0.227***	0.231***	0.238***	0.180***
	(0.038)	(0.043)	(0.040)	(0.038)	(0.039)	(0.036)	(0.036)
Trying	0.321***	0.353***	0.394***	0.408***	0.352***	0.360***	0.249***
	(0.039)	(0.050)	(0.048)	(0.053)	(0.053)	(0.052)	(0.046)
Transitioned	0.330***	0.309***	0.354***	0.318***	0.355***	0.268***	0.260***
	(0.063)	(0.097)	(0.098)	(0.114)	(0.097)	(0.096)	(0.071)
Social norm × Unwilling	0.073**	0.032	0.082*	0.118**	0.046	0.074*	0.035
	(0.036)	(0.052)	(0.049)	(0.049)	(0.049)	(0.044)	(0.043)
	q=0.042	q=0.547	q=0.100	q=0.018	q=0.346	q=0.099	q=0.421
Social norm × Hesitant	0.021	0.018	0.021	0.011	0.003	0.001	−0.012
	(0.035)	(0.029)	(0.026)	(0.027)	(0.027)	(0.029)	(0.030)
	q=0.548	q=0.519	q=0.404	q=0.678	q=0.918	q=0.963	q=0.701
Social norm × Trying	−0.102**	−0.090**	−0.108***	−0.114***	−0.057	−0.074**	−0.035
	(0.041)	(0.039)	(0.040)	(0.036)	(0.038)	(0.037)	(0.042)
	q=0.012	q=0.025	q<0.01	q<0.01	q=0.128	q=0.041	q=0.410
Social norm × Transitioned	−0.039	−0.009	−0.085	−0.020	−0.076	−0.010	−0.040
	(0.044)	(0.081)	(0.081)	(0.103)	(0.077)	(0.071)	(0.056)
	q=0.379	q=0.900	q=0.285	q=0.821	q=0.311	q=0.877	q=0.476
Food choice	0.263	0.343**	0.278*	0.295*	0.277*	0.268*	0.295*
	(0.166)	(0.155)	(0.160)	(0.158)	(0.160)	(0.161)	(0.164)
	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Direct spillover effect of the social norm message on the amount donated. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

Table 3.A.15: Robustness checks of the direct spillover effect conditional on predicted classes IV

Specification	Nested 2SLS model						
Outcome	Decision to donate (binary)						
Food choice	Food choice in kgCO2-eq						
Algorithm	GBM 1	GBM 2	Random Forest	Ordered logit	Multinomial	LDA	QDA
Unwilling (baseline)	0.332***	0.369***	0.296***	0.293***	0.303***	0.305***	0.381***
	(0.063)	(0.077)	(0.067)	(0.068)	(0.068)	(0.069)	(0.064)
Hesitant	0.213***	0.186***	0.230***	0.234***	0.233***	0.236***	0.175***
	(0.035)	(0.045)	(0.038)	(0.038)	(0.039)	(0.037)	(0.037)
Trying	0.317***	0.341***	0.404***	0.424***	0.367***	0.374***	0.260***
	(0.041)	(0.054)	(0.045)	(0.049)	(0.049)	(0.047)	(0.043)
Transitioned	0.362***	0.346***	0.405***	0.391***	0.410***	0.316***	0.293***
	(0.049)	(0.083)	(0.077)	(0.090)	(0.077)	(0.076)	(0.059)
Social norm × Unwilling	0.074**	0.034	0.091*	0.122**	0.051	0.072	0.033
	(0.036)	(0.054)	(0.050)	(0.050)	(0.051)	(0.045)	(0.044)
	q=0.040	q=0.537	q=0.067	q=0.012	q=0.302	q=0.107	q=0.451
Social norm × Hesitant	0.014	0.014	0.017	0.012	0.005	0.007	−0.001
	(0.036)	(0.029)	(0.027)	(0.027)	(0.027)	(0.028)	(0.028)
	q=0.701	q=0.648	q=0.535	q=0.647	q=0.856	q=0.812	q=0.962
Social norm × Trying	−0.076**	−0.068*	−0.083**	−0.096***	−0.043	−0.065*	−0.030
	(0.036)	(0.037)	(0.036)	(0.034)	(0.036)	(0.036)	(0.042)
	q=0.035	q=0.067	q=0.018	q<0.01	q=0.225	q=0.073	q=0.472
Social norm × Transitioned	−0.033	0.013	−0.072	−0.006	−0.064	0.005	−0.029
	(0.042)	(0.074)	(0.076)	(0.101)	(0.076)	(0.071)	(0.054)
	q=0.418	q=0.867	q=0.346	q=0.944	q=0.393	q=0.944	q=0.585
Food choice	−0.004	−0.006**	−0.004*	−0.004*	−0.004*	−0.004*	−0.004*
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Direct spillover effect of the social norm message on the amount donated. The first column displays the results yielded by the predictions of the GBM algorithm with over-sampling and the second when political beliefs and income are used as predictors. The other columns display the results yielded by the predictions of the other algorithms. Robust standard errors are displayed in parentheses. P-values from randomisation tests with 5,000 re-sampling are displayed last (q).

3.A.D Supplementary Tables and Figures

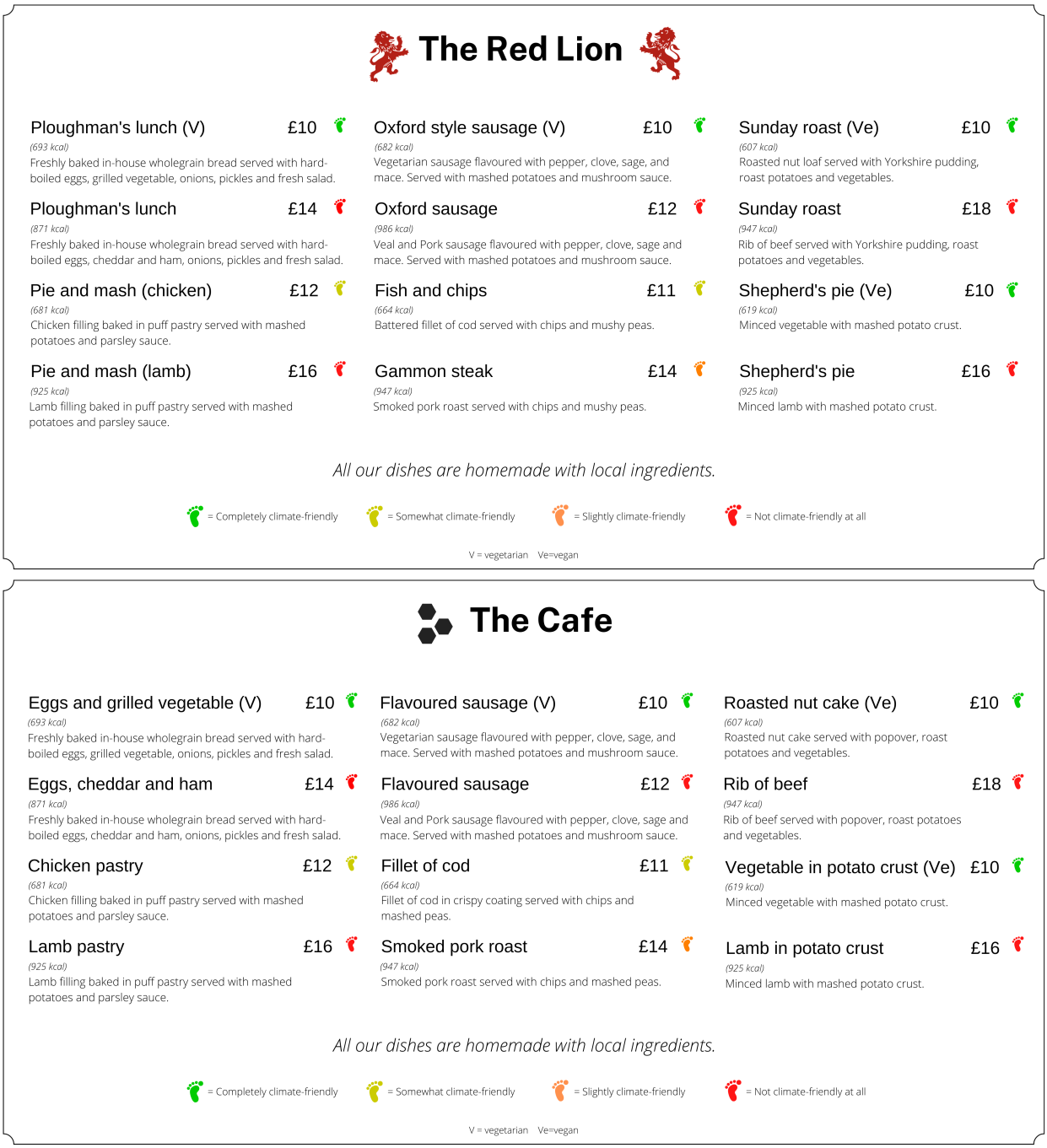


Figure 3.A.2: Full menus

Note: Two versions of the menus shown to participants. In total, we had 24 versions of the full menu in which we varied the ordering (12 versions) of the items and the menu's appearance (2 versions).

Table 3.A.16: Characteristics of the food items

Dish name	Main ingredients	Carbon footprint	Label colour
Eggs and grilled vegetable Ploughman's lunch	Eggs and vegetables	3.25	Green
Flavoured sausage Oxford style sausage	Beans	0.80	Green
Vegetable in potato crust Shepherd's pie	Vegetables	1.60	Green
Roasted nut cake Sunday roast	Nuts	2.00	Green
Chicken pastry Pie and mash	Chicken	5.40	Yellow
Fillet of cod Fish and chips	Fish	5.40	Yellow
Smoked pork roast Gammon steak	Pork	7.90	Orange
Eggs, cheddar and ham Ploughman's lunch	Ham and cheese	23.88	Red
Flavoured sausage Oxford sausage	Veal and pork	38.35	Red
Lamb in potato crust Pie and mash	Lamb	64.20	Red
Lamb pastry Shepherd's pie	Lamb	64.20	Red
Rib of beef Sunday roast	Beef	68.80	Red

Note: Based on its carbon intensity, we categorise each dish in one of four categories, corresponding to the carbon footprint labels. Carbon footprints are computed based on the main ingredients of the dishes, using Scarborough et al. (2014)'s estimates. When dishes have more than one ingredient, we take the average between the two.

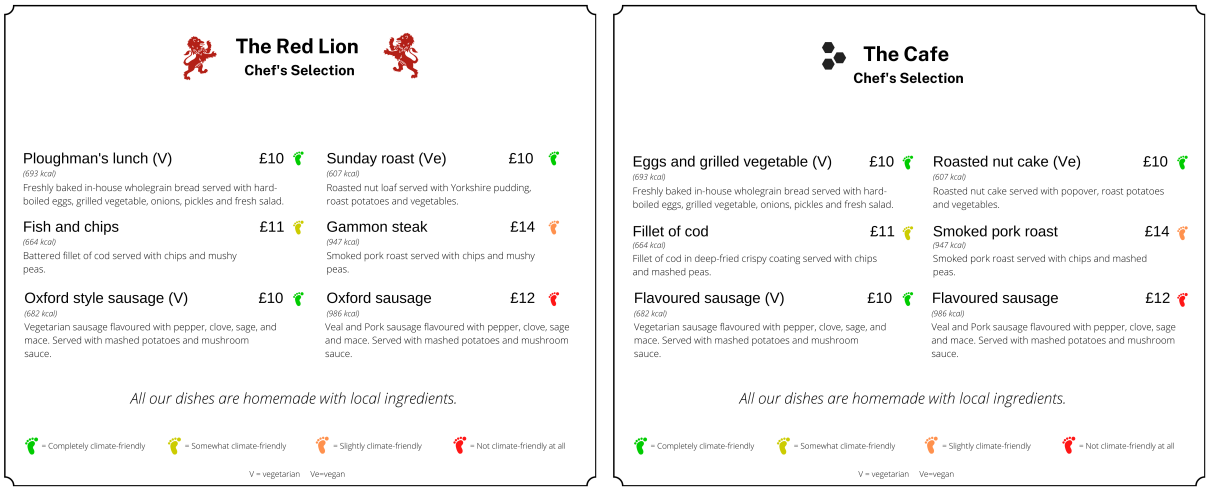


Figure 3.A.3: Plant-intensive default menus

Note: Two versions of the plant-intensive default menus that were shown to participants.

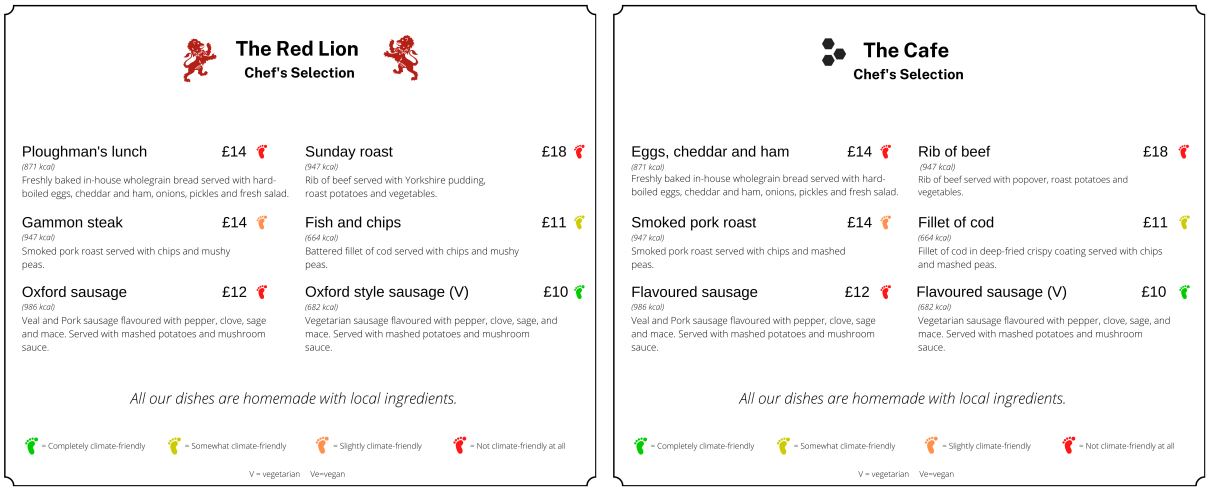


Figure 3.A.4: Meat-intensive default menus

Note: Two versions of the meat-intensive default menus that were shown to participants.

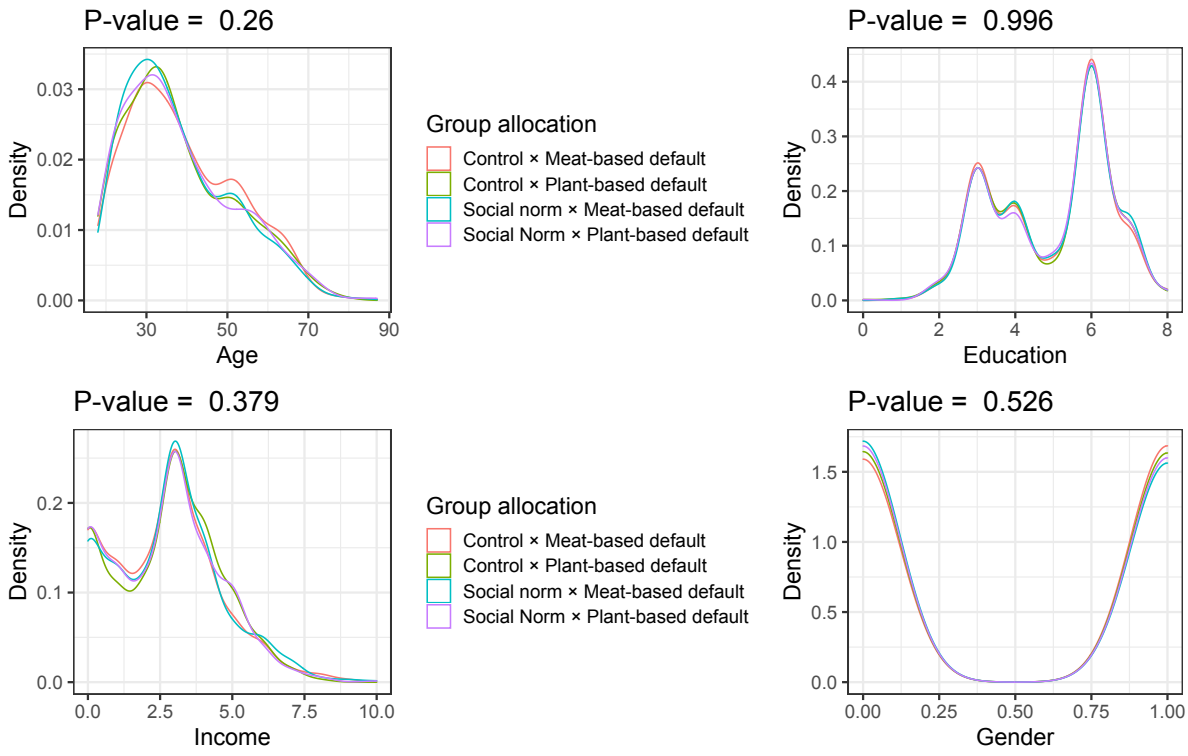


Figure 3.A.5: Distribution of the main covariates by treatment group

Note: Density plots of age, education, income and gender across the four treatment groups of the *main sample*. For education, 0 means "No education", and 8 means "PhD or equivalent". For Income, 0 means "less than £10k" and 10 means "more than £150k". For gender, 0 means female, and 1 means male.

Table 3.A.17: Descriptive statistics

Main covariates	
Age	
Mean	38 years old
Min	18 years old
Max	87 years old
SD	13.59 years old
Income	
Missing	339
$<$ £10,000	969 (18.6%)
£10,000 - £15,999	673 (12.9%)
£16,000 - £19,999	580 (11.1%)
£20,000 - £29,999	1446 (27.7%)
£30,000 - £39,999	793 (15.2%)
£40,000 - £49,999	405 (7.8%)
£50,000 - £69,999	224 (4.3%)
£70,000 - £89,999	77 (1.5%)
£90,000 - £119,999	33 (0.6%)
£120,000 - £149,999	12 (0.2%)
More than £150,000	6 (0.1%)
Gender	
Missing	1
Female	2771 (49.9%)
Male	2736 (49.2%)
Agender	1 (0.0%)
Non-binary / third gender	42 (0.8%)
Trans woman	1 (0.0%)
Prefer not to say	5 (0.1%)
Education	
Missing	29
No education	2 (0.0%)
Primary education	12 (0.2%)
Lower secondary education	137 (2.5%)
Upper secondary education	1287 (23.3%)
Post-secondary non-tertiary education	853 (15.4%)
Short-cycle tertiary education	321 (5.8%)
Bachelor or equivalent	2166 (39.2%)
Master or equivalent	663 (12.0%)
Doctoral or equivalent	87 (1.6%)

Note: Distribution of the main covariates across the 5,557 participants to the experiment.

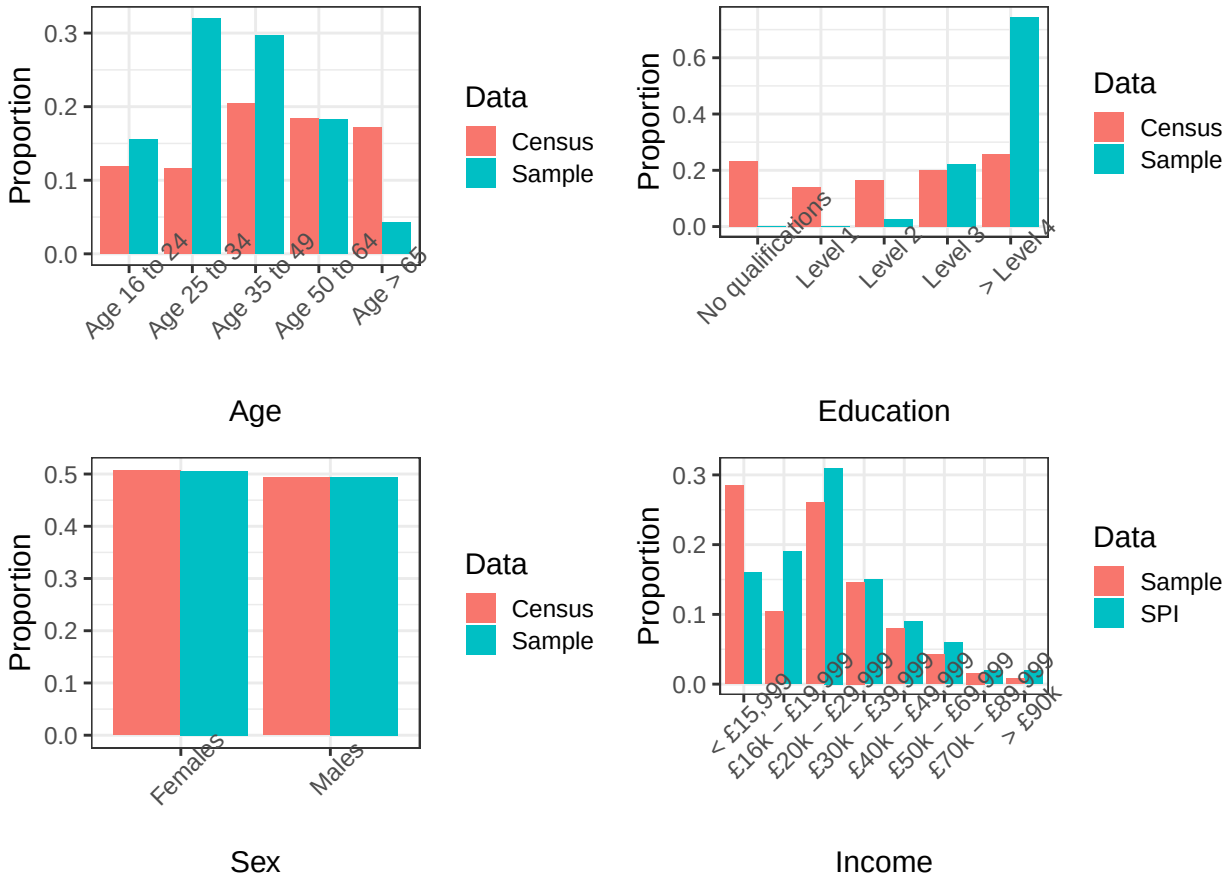


Figure 3.A.6: Comparison with UK population

Note: Comparison of the distributions of the main covariates in the sample and the UK population. We use the data from the 2011 census to plot the distribution of age, sex and education in the UK population. We use the 2020/2021 survey of personal income to plot the distribution of income in the UK population.

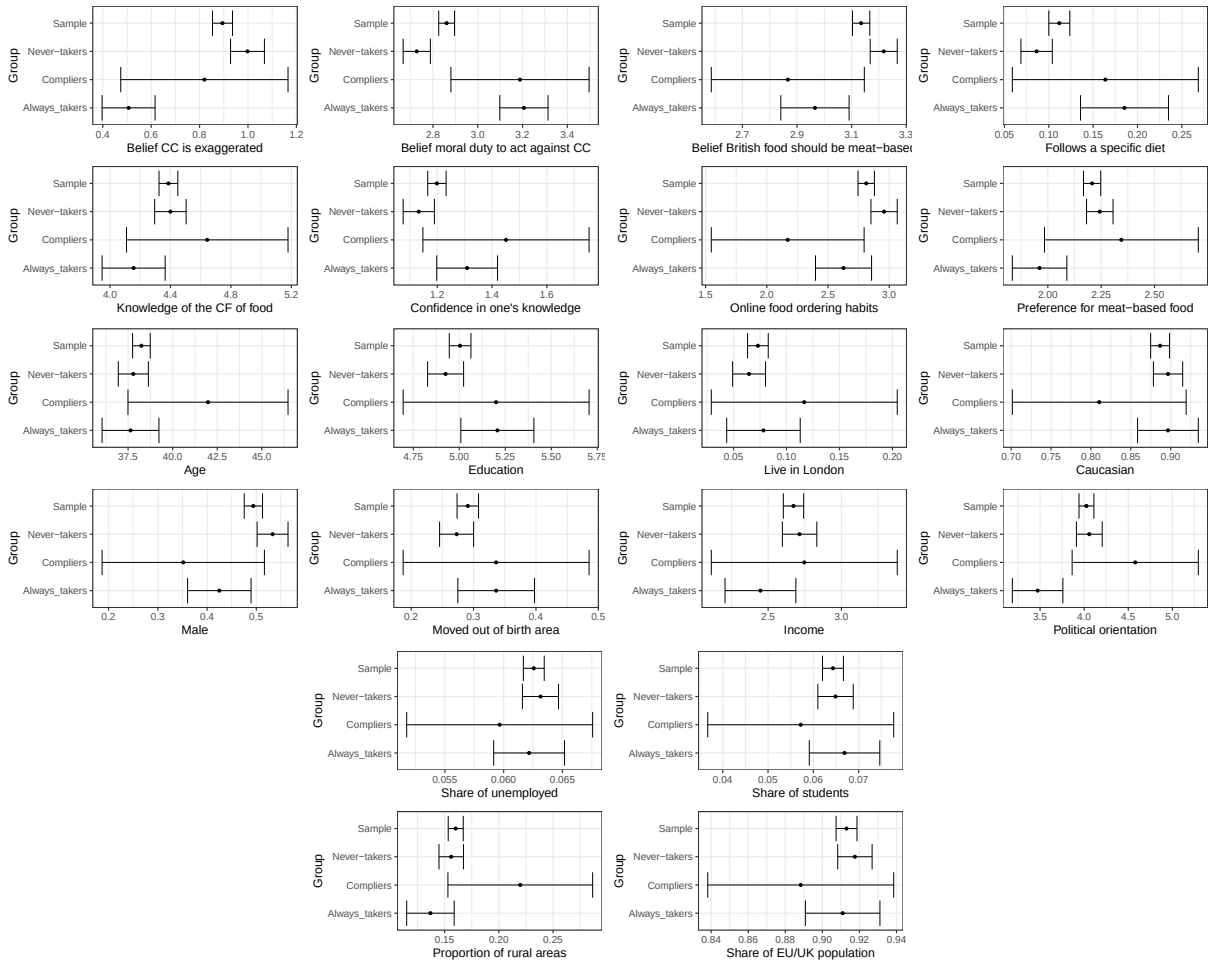


Figure 3.A.7: Profile of compliers

Note: We represent how the profile of compliers (those choosing vegetarian food when prompted to do so by the default nudge) differ from the rest of the sample, following Marbach and Hangartner (2020).

Table 3.A.18: Questions to Measure Hypothetical Bias

	WTA eating vegetarian	Feeling of Sacrifice
Constant	3.115***	2.191***
	(0.368)	(0.063)
Choose meat	1.463***	0.483***
	(0.172)	(0.053)
Social Norm	−0.038	0.007
	(0.133)	(0.045)
Num.Obs.	2775	2775
R2	0.083	0.029
Av. WTA among vegetarians	2.742***	

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Results of OLS regressions. In column one, the compensation that respondents ask for being forced to switch to vegetarian food is regressed on the dummy capturing whether they chose a meat dish and the dummy capturing their allocation to the social norm message, controlling for the amount given during the donation task. In column two, the self-reported feeling of sacrifice is regressed on the dummy capturing whether they chose a meat dish and the dummy capturing their allocation to the social norm message. The last row contains the average compensation asked by respondents who chose a vegetarian item. We perform a t-test to check if this amount is significantly different from zero.

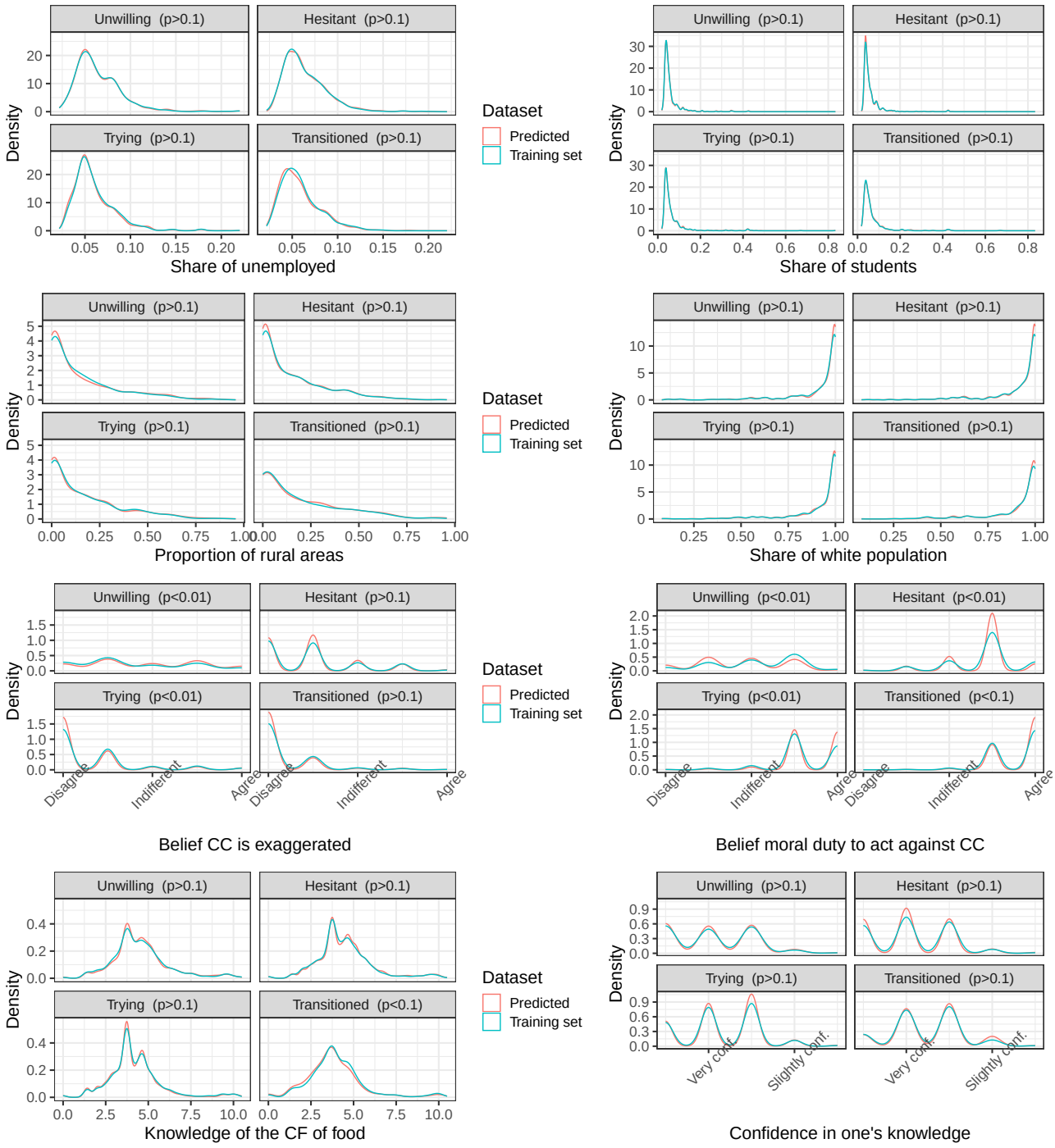


Figure 3.A.8: Distribution of the predictors by type I

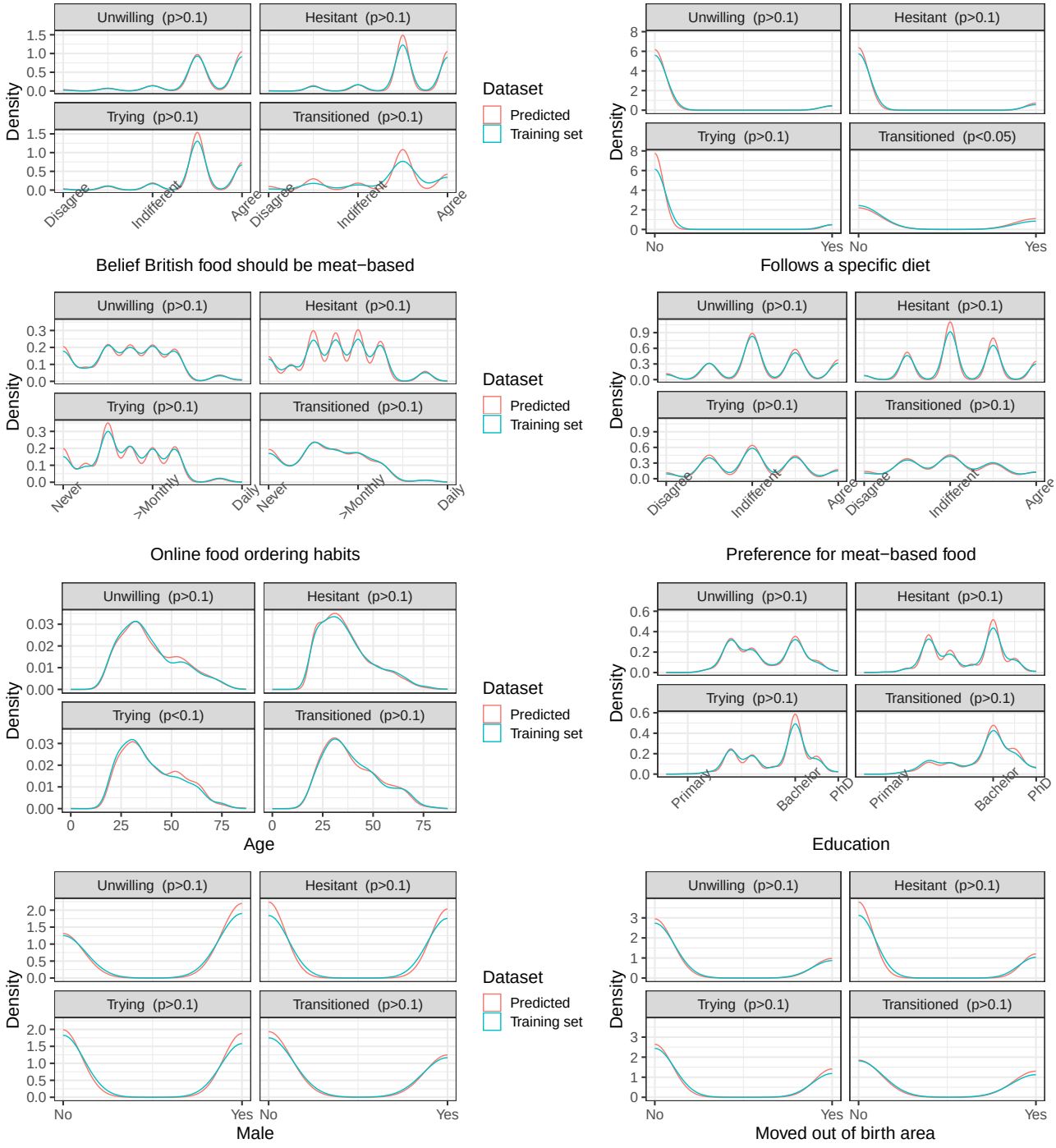


Figure 3.A.9: Distribution of the predictors by type II

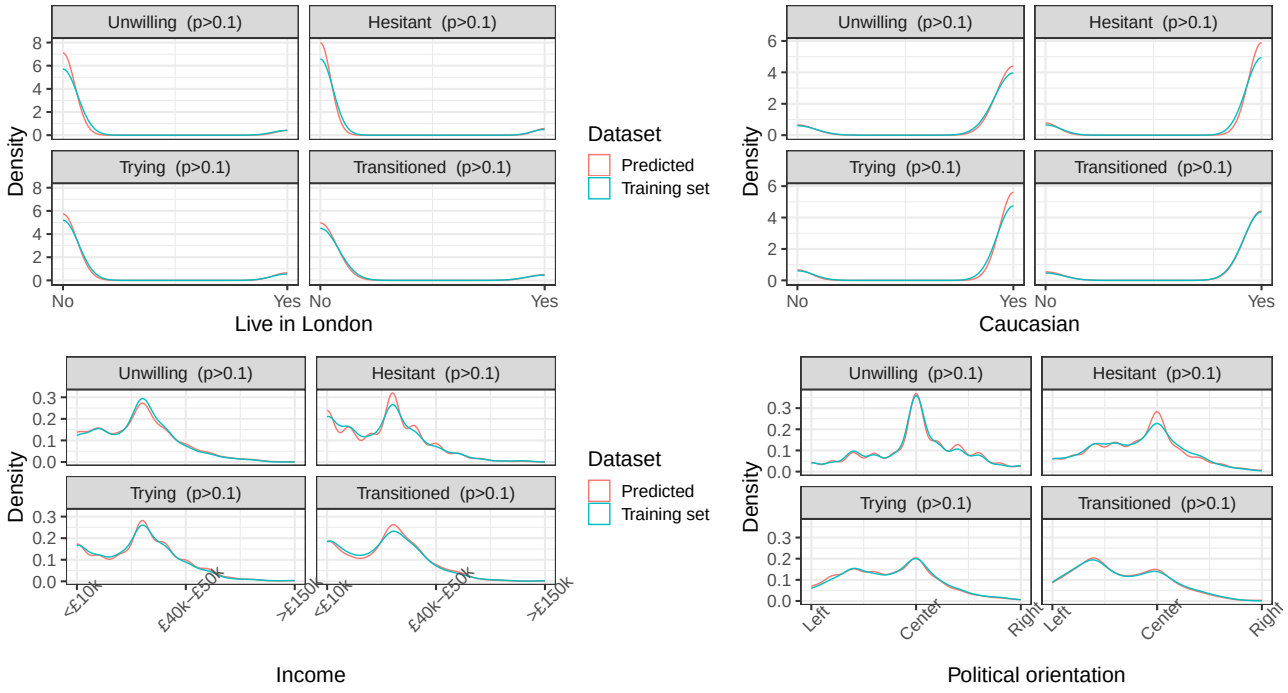


Figure 3.A.10: Distribution of the predictors by type III

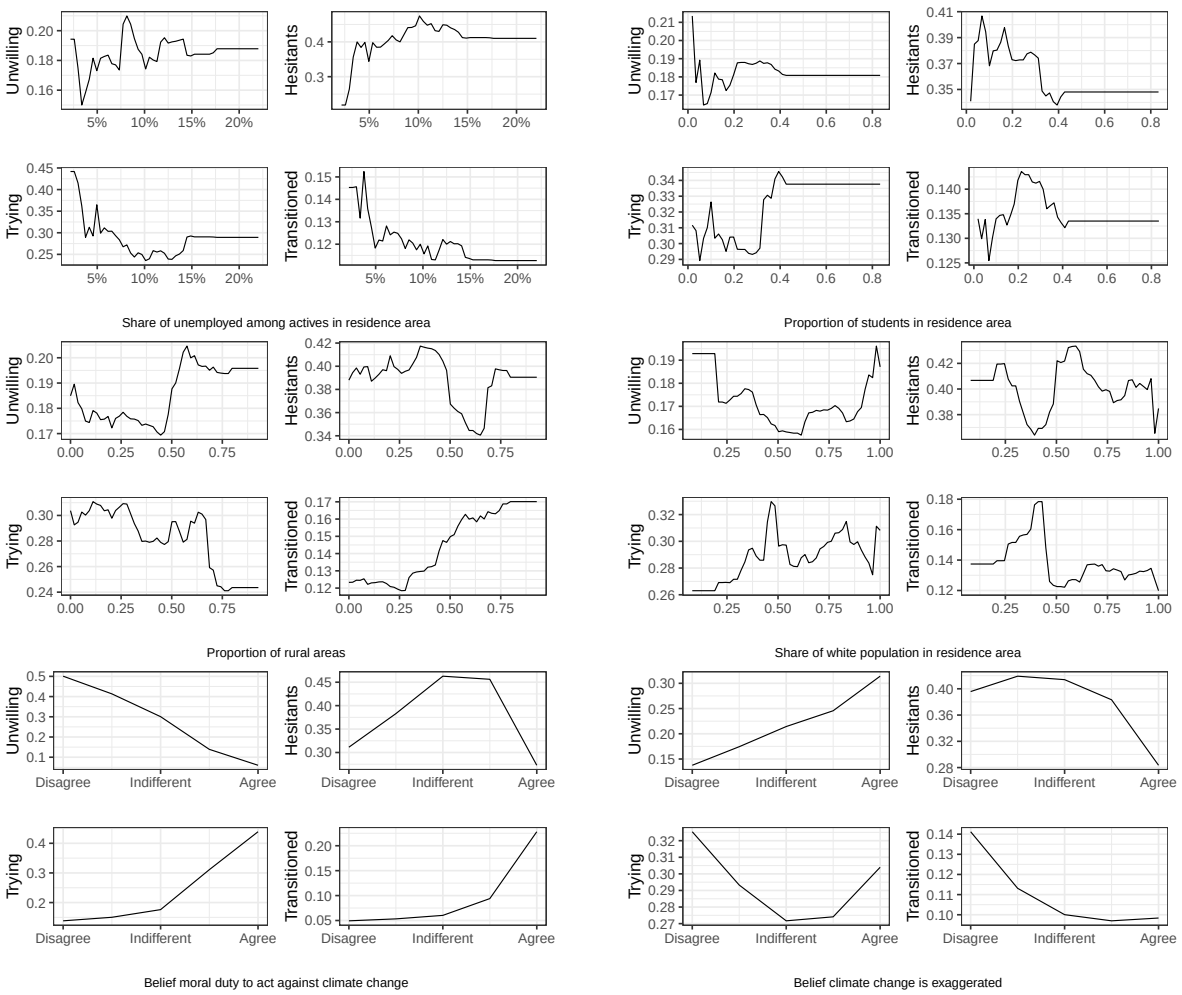


Figure 3.A.11: Partial dependence plots of the GBM algorithm I

Note: Partial independence plots visually express the likelihood of being allocated to a given class against the values a variable takes.

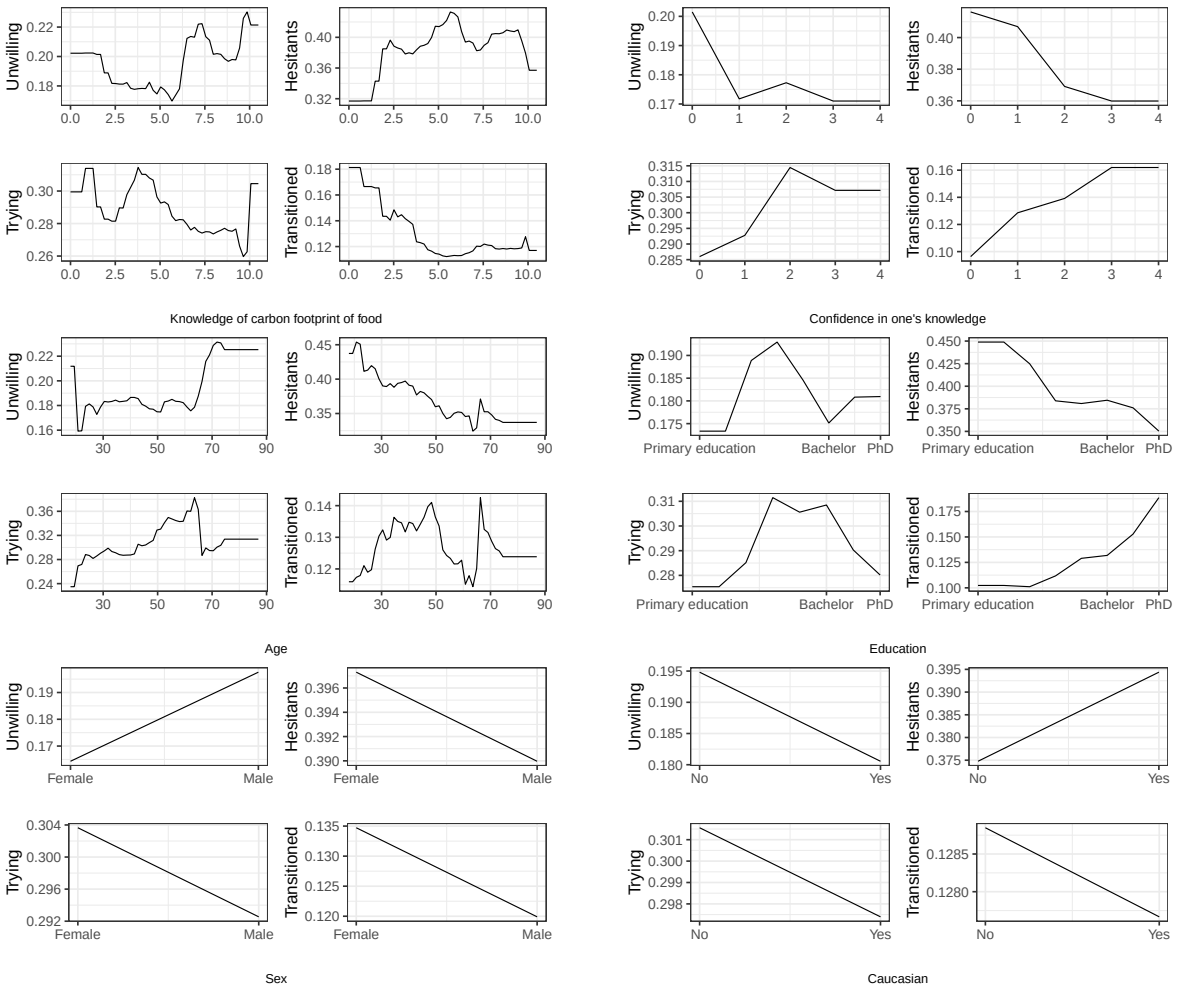


Figure 3.A.12: Partial dependence plots of the GBM algorithm II

Note: Partial independence plots visually express the likelihood of being allocated to a given class against the values a variable takes.

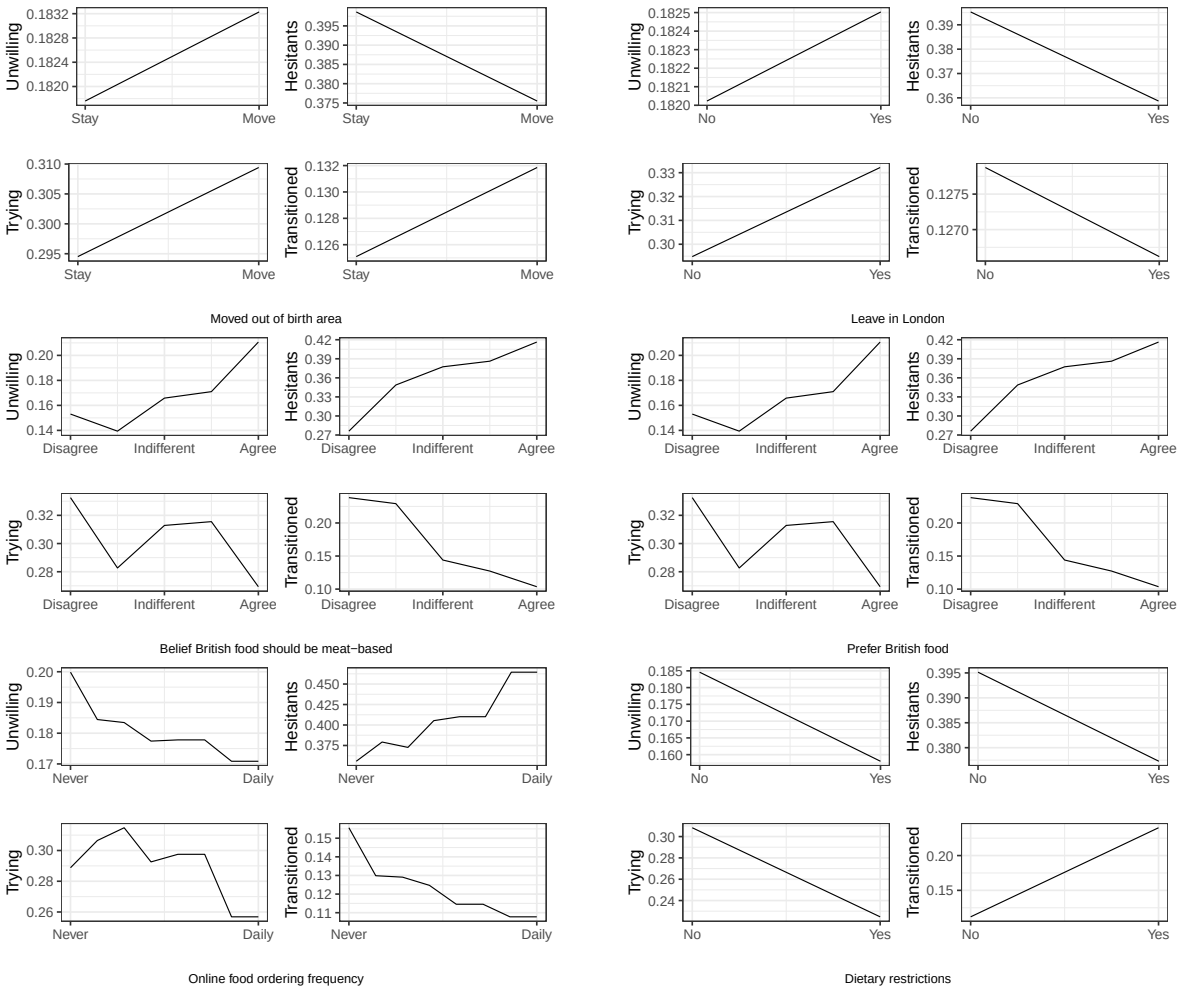


Figure 3.A.13: Partial dependence plots of the GBM algorithm III

Note: Partial independence plots visually express the likelihood of being allocated to a given class against the values a variable takes.

Table 3.A.19: Profile of each predicted type

Covariates	Unwilling	Hesitant	Trying	Transitioned
Share of EU/UK population	0.004	0.003	-0.006	-0.004
Share of unemployed	0.003	0.004***	-0.006***	-0.006*
Proportion of rural areas	0.001	-0.023**	0.012	0.062**
Share of students	-0.001	-0.009**	0.009*	0.004
Belief moral duty to act against CC	-1.942***	-0.223***	0.957***	1.025***
Belief CC is exaggerated	1.598***	0.097	-0.696***	-0.794***
Knowledge of the CF of food	0.33**	0.442***	-0.43***	-1.008***
Confident in one's knowledge	-0.12	-0.265***	0.244***	0.484***
Age	2.851**	-4.787***	4.008***	0.601
Educated	-0.331**	-0.396***	0.369***	1.01***
Male	0.191***	-0.059*	0.023	-0.126**
Moved out of birth area	-0.034	-0.121***	0.112***	0.194***
Caucasian	-0.015	-0.003	0.021	-0.023
Live in London	-0.047***	-0.019	0.047***	0.012
Income	0.074	-0.332***	0.326***	0.06
Conservative	1.248***	0.177	-0.61***	-0.878***
Belief British food should be meat-based	0.174 **	0.157 ***	-0.036	-0.78 ***
Preference for meat-based food	0.224 **	0.192 ***	-0.229 ***	-0.456 ***
Follows a specific diet	-0.047 *	-0.005	-0.092 ***	0.397 ***
Order food online frequently	-0.089	0.541 ***	-0.441 ***	-0.549 ***

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Regression coefficients from linear models where each covariate is regressed on a dummy equal to 1 if respondents are classified in a given type, and zero otherwise. Coefficients, therefore, capture how different a given type is compared to the average of the sample. P-values are adjusted using Holmes-Bonferroni correction.

Chapter 4

How Framing, Difficulty and Domain-Similarity Shape Policies' Side Effects

I Introduction

Emotions are an important driver of pro-environmental action (Davidson and Kecinski, 2022). Research shows that fear heightens the perception of the risks associated with climate change and motivates green initiatives (Wong-Parodi and Feygina, 2021; Skurka et al., 2018). Similarly, pride and hope also increase people's intentions to act (Shipley and van Riper, 2021; Nabi et al., 2018). Politicians, environmental activists, and concerned citizens frequently appeal to emotions to raise awareness of climate change. This can involve emphasising the negative outcomes and moral implications of doing nothing or praising environmental efforts and underlining their benefits.[1] The assumption is that mere information provision is not enough (Chess and Johnson, 2007; Davidson and Kecinski, 2022). This raises two questions. First, should we focus on negative or positive emotions to foster climate-friendly initiatives? Second, which approach sustains climate action beyond the behaviour that is promoted?

We aim to answer these questions using a pre-registered survey experiment.[2] We compare two narratives promoting environmental activism. The first narrative emphasises the negative consequences of inaction. This "doom-and-gloom" approach encourages pro-environmental behaviour by evoking negative emotions (e.g., guilt and fear). The second narrative emphasises the positive benefits of acting against climate change. This "win-win" approach aims to foster pro-environmental actions by triggering positive feelings such as pride from doing the right thing.

Reviewing the literature on gain and loss frames, Homar and Cvelbar (2021) find that both

[1] A striking example of an emotional appeal is Greta Thunberg's speech at the UN climate action summit 2019: *"People are suffering. People are dying. Entire ecosystems are collapsing. We are in the beginning of a mass extinction, and all you can talk about is money and fairy tales of eternal economic growth. How dare you!"*

[2] Registered report available here.

strategies boost people's pro-environmental intentions. Yet, this literature has focused mostly on stated preferences, using experiments with small sample sizes. Similarly, studies on the role of pride and guilt in inducing pro-environmental behaviours suffer from the same limitations (Shipley and van Riper, 2021). We contribute to these two strands of the literature by comparing how our two narratives foster people's participation in a consequential real-effort task. The task consists of voluntarily spending time labelling food pictures by their main ingredients. The research team will use participants' answers to create an app. This app will help people understand how their food choices affect the environment.[3] We also ensure our study is powered enough by relying on a large sample size (n=10,670).

Furthermore, we test whether exposition to one of the narratives affects respondents' likelihood of doing another pro-environmental action: signing a petition for the environment. Addressing this question is crucial to determine whether such communication strategies yield potential co-benefits or crowd out further engagement. To our knowledge, we are the first to investigate and compare the spillover effects of stressing the positive benefits of action versus the negative consequences of inaction. Whilst these two approaches can be seen as two faces of the same coin, they likely change behaviours through different psychological processes. We expect that inducing negative feelings by stressing the consequences of inaction induces an extrinsically motivated response. People act *in order to* temper these negative feelings. This can induce people to "ease off" afterwards. On the other hand, we expect that stressing the benefit of doing the right thing for the environment triggers an intrinsically motivated reaction. People *feel* emboldened, making them more likely to

[3]More precisely, participants' answers are used as a training set to train a machine learning algorithm to detect the main ingredient in a picture. The app will then display the carbon footprint of the ingredient. The training set generated by participants' answers complements another training set obtained from a previous experiment.

engage in other pro-environmental actions. In Section II, we develop the justifications underpinning our testable hypotheses.

Our experiment is structured in three parts. First, respondents read an article on how to help scientists develop solutions to mitigate people's emissions. Second, participants chose to do a real effort task similar to the one described in the article (henceforth denoted by PEB1). Third, respondents are offered to sign a petition against climate change (henceforth denoted by PEB2). We adopt a fractional design. Namely, data collection is structured in three waves. In wave 1 (n=5,492), we randomly vary the narratives in the articles. Respondents are either presented with (1) a placebo article dealing with a topic unrelated to PEB1, (2) a control article where we provide neutral information regarding PEB1, (3) the "doom-and-gloom" article where we provide neutral information and then present arguments stressing the negative consequences of inaction, or (4) the "win-win" article where we provide neutral information and then present arguments stressing the positive consequences of action. In wave 2 (n=2,622), we made the real effort task easier. In wave 3 (n=2,556), we changed the cause supported by the petition to something unrelated to the environment (i.e., fighting loneliness). We present our design in Section III.

We formalise the pathways through which we expect our narratives to trigger spillover effects on the likelihood of signing the petition as follows:[4]

$$\underbrace{\Delta_T PEB2}_{\text{Net effect}} = \underbrace{\partial_T PEB1 \times \partial_{PEB1} PEB2}_{\text{Indirect effect}} + \underbrace{\partial_T PEB2}_{\text{Direct effect}} \tag{I.1}$$

We denote by $\Delta_T PEB2$ the *net spillover effect* of narrative T on the non-targeted behaviour (i.e., signing the petition). This *net spillover effect* can be decomposed into two channels: the *indirect*

[4]See Chapter 2 for details on where this equation comes from.

spillover effect ($\partial_T PEB1 \times \partial_{PEB1} PEB2$) and the *direct spillover effect* of the narrative on the non-targeted behaviour ($\partial_T PEB2$). The indirect effect spillover effect is the behavioural spillover effect ($\partial_{PEB1} PEB2$) scaled by the main effect of the policy on the targeted pro-environmental action ($\partial_T PEB1$).

Our objectives are threefold. First, we seek to determine if our two narratives trigger different *direct spillover effects*. Second, we test whether the effort exerted when undertaking the first pro-environmental behaviour moderates the size of the *behavioural spillover effects*. Third, we investigate if *direct* and *indirect spillover effects* are restricted to behaviours belonging to the same domain (e.g., climate action) or can occur across domains (e.g., health and environment).

We rely on the methodology presented in Chapter 3. Namely, we embed an instrumental variable in the design to causally estimate the *behavioural spillover effect* ($\partial_{PEB1} PEB2$). Doing so also allows us to derive an unbiased estimate of the *direct spillover effect* ($\partial_T PEB2$). This instrumental variable consists of randomly varying the size of the button respondents press to participate in the targeted pro-environmental behaviour. Section IV presents our methodology.

We find no evidence for our pre-registered hypotheses. Stressing the benefits of climate actions or the cost of inaction does not affect respondents' likelihood to do the real effort task compared to neutral information. Moreover, providing neutral information does not alter the uptake of the real effort task compared to the placebo group. We also find no evidence for *direct* or *indirect spillover effects*. The difficulty of the real effort task does not mediate the *indirect spillover effect* either. In an exploratory analysis, we also causally show that higher performances in the real effort task increase the feeling that one has made an effort for the environment but do not increase the likelihood of signing the petition. Finally, we find suggestive evidence that stressing the cost of

inaction increased people's likelihood of signing the petition about youth loneliness. We present our results in Section V.

We see several explanations for these null results. The first relates to the nature of our treatment interventions. In two pilot studies, we fine-tuned the phrasing of our treatment texts to have the "doom-and-gloom" narrative trigger negative feelings and the "win-win" one trigger positive feelings. Nonetheless, it might still be that these manipulations were not strong enough to work. In this regard, the nature of the real effort task — our targeted pro-environmental behaviour — may have also played a role. Doing this task is costly to participants as the time spent doing it is uncorrelated with their payment. As such, whilst emotion-based appeals might work on intentions, they may not be strong enough to induce actual behaviour changes. Besides, around 70% of the sample participated. The remaining 30% may be the hardest to convince. Finally, the real effort task is peculiar and likely new to respondents. It is not something they are used to do. They do not know if doing it is a social norm or if it yields social approbation. The environmental gains are indirect and not salient. As such, it might have been that this task did not provide any other rewards than the mere knowledge that one did a good thing for the environment. It is possible that this is not enough to motivate people to act pro-environmentally or induce behavioural spillover effects. We discuss these different possibilities in Section VI.

With these caveats in mind, our study overall questions the effectiveness of pure information provision and emotion-based appeals to stir pro-environmental action. It confirms recent meta-analyses questioning the effectiveness of behavioural interventions (Nisa et al., 2019; Maier et al., 2022). We also make three contributions to the literature studying behavioural spillover effects. First, factors altering spillovers are often studied in isolation, and studies are not always comparable (Nilsson et al., 2017). In this paper, we investigate factors hypothesised to affect the sign of

spillovers (the framing of narratives) and their magnitude (difficulty and domain similarity of behaviours) in the same experimental framework. Second, experiments on spillover effects are often underpowered (Maki et al., 2019). This issue questions the reproducibility of the results reported by some studies. We are voluntarily conservative to avoid this pitfall: we choose conservative expected effect sizes and base our power analysis on two-sided tests. Furthermore, we used a registered report to pre-register our hypotheses. Finally, there is no consensual definition of behavioural spillovers, implying they have often been measured inconsistently (Maki et al., 2019; Truelove et al., 2014; Dolan and Galizzi, 2015). We provide a rigorous definition of the spillover effects that interventions trigger and a methodology to estimate them causally.

II Testable hypotheses

Based on Homar and Cvelbar (2021)'s literature review, we expect that stressing the negative consequences of inaction induces feelings such as guilt or shame. Conversely, we expect that stressing the positive aspects of taking action induces emotions such as pride. The meta-analysis of Shipley and van Riper (2021) suggests that feelings of pride and guilt moderate pro-environmental actions. As such, we posit that the positive and the negative arguments increase participation in the pro-environmental behaviour they promote (henceforth PEB1).

HYPOTHESIS 1: Positive arguments increase participation in PEB1 ($\partial_T PEB1 > 0$).

HYPOTHESIS 2: Negative arguments increase participation in PEB1 ($\partial_T PEB1 > 0$).

We expect that doing PEB1 increases participants' willingness to do another pro-environmental behaviour that we denote by PEB2 ($\partial_{PEB1} PEB2 > 0$). This hypothesis is supported by the ex-

periment presented in Chapter 3 and studies documenting such positive spillover effects in the environmental domain (e.g., Alacevich et al. 2021; Comin and Rode 2023).

HYPOTHESIS 3: The indirect spillover effect is positive ($\partial_T PEB1 \times \partial_{PEB1} PEB2 > 0$).

We hypothesise that triggering pride through positive arguments makes people's pro-environmental identity salient. In their theoretical framework, Truelove et al. (2014) argue that enhancing a social role (e.g., being pro-environmental) is likely to trigger positive spillover effects by inducing people to conform with this role (Cialdini, 1994; Cialdini et al., 1995; Festinger, 1962).

HYPOTHESIS 4: Positive arguments yield positive direct spillover effects on PEB2 ($\partial_T PEB2 > 0$).

On the other hand, we expect that negative arguments make people act pro-environmentally to reduce their guilt about not doing enough. Stressing the costs of inaction might also create mental discomfort through people's loss aversion (Kahneman et al., 1991). Truelove et al. (2014) hypothesise that behavioural interventions inducing people to act to temper negative feelings are likely to trigger negative spillover effects (Tiefenbeck et al., 2013; Mullen and Monin, 2016). Indeed, moral licensing is an underlying mechanism underpinning negative spillovers. Moral licensing describes people feeling entitled to "ease off" after doing morally virtuous behaviour to repair a deprecated identity (Tiefenbeck et al., 2013; Gneezy et al., 2012; Clot et al., 2014).

HYPOTHESIS 5: Negative arguments yield negative direct spillover effects on PEB2 ($\partial_T PEB2 < 0$).

We also aim to test if the difficulty of the first pro-environmental action affects the magnitude and the direction of *behavioural spillover effect* ($\partial_{PEB1}PEB2$). Here, different theories yield opposite predictions. The foot-in-the-door theory predicts that accepting a first easy task increases the probability of accepting a second harder one (Freedman and Fraser, 1966). On the other hand, several studies indicate that performing a first harder task increases the likelihood of doing a second easier task (Maki et al., 2019; Dolan and Galizzi, 2015; Lanzini and Thøgersen, 2014). For instance, Gneezy et al. (2012) find that pro-social actions perceived as costly consistently generate positive spillover effects. For the authors, the harder the pro-social behaviours are, the stronger they make us feel pro-social when doing them.

HYPOTHESIS 6: The difficulty of PEB1 affects the effect of doing PEB1 on PEB2 ($\partial_{PEB1}PEB2$).

Finally, we seek to test whether spillovers only occur between behaviours belonging to the same domain (i.e., environment). Evidence suggests that behaviours requiring similar resources (e.g., money, time, place of performance) are correlated (Thøgersen and Ölander, 2003; Thøgersen, 2004; Margetts and Kashima, 2017). Previous studies examining behavioural consistency have largely focused on same-domain behaviours, such as pro-environmental actions,[5] charitable donations,[6] or health-related decisions.[7] Less well understood, however, is whether spillovers occur across domains. Empirical evidence is scarce and mixed, with studies suggesting cross-domain behavioural spillovers might exist (Carrico et al., 2018; Mazar and Zhong, 2010; List and Momeni, 2020), and

[5]For instance, see Margetts and Kashima (2017); Lanzini and Thøgersen (2014); Thøgersen (2004); Truelove et al. (2016); Jessoe et al. (2021); Ek and Miliute-Plepiene (2018); Schusser and Bostedt (2019); Sintov et al. (2019).

[6]For instance, see Corazzini et al. (2015); Krieg and Samek (2017); Meer (2017); Deck and Murphy (2019); Filiz-Ozbay and Uler (2019); Carlsson et al. (2021).

[7]For instance, see Dolan et al. (2015); Bech-Larsen and Kazbare (2014).

others supporting the hypothesis that behavioural spillovers are more prevalent between choices in similar domains (Noblet and McCoy, 2018; Garvey and Bolton, 2017).

HYPOTHESIS 7: The effect of doing PEB1 on PEB2 ($\partial_{PEB1}PEB2$) is different when PEB2 is unrelated to the environment.

HYPOTHESIS 8: The direct spillover effect ($\partial_T PEB2 > 0$) is different when PEB2 is unrelated to the environment.

The following section details our experimental design to test these hypotheses.

III Design

The experiment consists of five steps: attention checks, pre-treatment surveys, treatment interventions, and targeted and non-targeted pro-environmental tasks. The attention checks and the online surveys are identical across our treatment groups. Differences in the experimental procedure only arise from the treatment interventions onwards. Details of the survey, questionnaires, treatment interventions and instructions can be accessed in the online supplementary material, here.

Attention checks

Before starting the survey, subjects answer two questions to assess if they are attentive. Respondents failing to answer them correctly were excluded from the analysis. The first attention check is a 5-Likert scale question: "People are very busy these days, and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer 'very interested'". The second attention check is a multiple-choice question: "Most modern decision-making theories recognise that decisions do not take place in

a vacuum. Individual preferences, knowledge, and situational variables can greatly impact the decision process. Select' red' among the alternatives below to demonstrate that you've read this much."

Pre-treatment survey

Before respondents are allocated to treatment arms, they answer questions about their location of residence, gender, age, income, political orientation, and education level. We also evaluate respondents' pro-environmental and altruistic attitudes using the scale provided by Bouman et al. (2018). Finally, we also measure respondents' proneness to guilt using the scale developed by Cohen et al. (2011).

Treatment interventions

In the remainder of the experiment, respondents are presented with a newspaper article covering the importance of PEB1. Then, they can choose to do PEB1. Finally, respondents are offered to undertake PEB2. We randomise the framing of the newspaper articles, the difficulty of PEB1, and the domain of PEB2.

PEB1 is a real effort task taking the form of a series of decisions in which respondents categorise food pictures by their main content (e.g., "ruminant meat", "non-ruminant meat", "fish," etc.). They have 10 seconds to determine which category a given picture is more likely to belong to. In total, they have 30 pictures to categorise. We explain to respondents that their participation in PEB1 would help us develop a dataset to train an algorithm to predict the carbon footprint of food items. This algorithm aims to help people gauge the environmental impact of their food choices.

PEB2 consists of signing a petition to redesign the German car tax by introducing a bonus-malus

system accounting for vehicles' CO2 emissions. We address the petition to the Petitions Committee of the German Bundestag. In contrast to petitions organised under private law, petitions submitted to the Petitions Committee of the German Bundestag have a legal right to be processed.

Subjects perform two filler tasks between PEB1 and PEB2. The first task consists of moving the pointers of sliders to an indicated graduation. The second task consists of sorting pictures into two categories ("big" or "small"). Separating the two behaviours helps mitigate potential experimenter demand effects.

Manipulation of the framing of narratives Participants are randomly assigned to one of four groups: two treatment groups, one control group and a placebo:

1. **Placebo:** Respondents are presented with a newspaper article covering a subject unrelated to PEB1:

   Expensive, greasy, hard to digest: the availability of unhealthy food in school kiosks is a widespread problem, but one that teachers and parents can tackle. At the Ludwig-Thoma-Realschule in Munich, healthy eating is an integral part of the school profile. Sugary drinks, fatty foods and white flour products are virtually non-existent there. Instead, fresh milk, mineral water and wholemeal bread are offered. In addition, nutritional education takes place in lessons, on project days and in cookery courses. However, this nutritional concept is a rarity in German schools, although the effects of fatty, sweet and vitamin-poor food are well known. The German Nutrition Society recommends sufficient consumption of water or other calorie-free drinks.

2. **Control group:** Respondents are presented with a newspaper article descriptively covering PEB1:

Studies show that Germans could save 67 million tonnes of CO2 annually by eating less meat. That is roughly equivalent to Portugal's total emissions. But giving up meat is difficult. Researchers at the University of Kassel are developing an innovative app to help us all make positive changes to our diet. This app allows you to take photos of meals and shows not only what you eat but also how your consumption affects our environment. But to make this app really effective, the scientists need to train algorithms that recognise food in the photos. To achieve this, the university works with volunteer lay researchers to assess tens of thousands of images. This is crucial to ensure that the image recognition algorithms work reliably and that the app can contribute to climate protection.

3. **Win-win treatment:** Respondents are presented with the same newspaper article as the control group. The following paragraph is added. It encourages participation in PEB1 by emphasising the benefit of taking action:

   Become part of the lay research community! Your support is not only important for the researchers who are developing new technologies to combat climate change. You are directly helping people who benefit from a digital app in their everyday lives overcome old habits. Become part of the journey towards a healthier and more sustainable future. Help science! Together, we can defeat climate change!

4. **Doom-and-gloom treatment:** Respondents are presented with the same newspaper article as the control group. A paragraph is added to encourage participation in PEB1 by stressing the costs of inaction:

Become an active lay researcher because inaction is deadly! It's already five past twelve on the doomsday clock. If we do nothing, climate change will have catastrophic consequences for us all. We must all change our lifestyles to stop the destruction of our planet. Take the first step now! If everyone doesn't do their bit, climate change can no longer be stopped!

Respondents read the articles allocated to them before deciding whether to undertake PEB1. Then, they answer a cloze test to ensure they read the texts. In two pilot studies, we fine-tuned the wording of the newspaper articles. Compared to the control group, reading the Win-Win text increases feelings of being capable, determined, and proud. Reading the Doom-and-Gloom text increases shame and guilt (see Figures 4.A.1 and 4.A.2 in Appendix 4.A.A).

Manipulation of the difficulty of PEB1 We allocate participants to one of the two following treatment arms:

1. **Hard PEB1:** Participants execute a series of 30 rounds where in each round, they have 10 seconds to classify pictures by their main ingredients. They have six categories: "Ruminant meat", "Non-ruminant meat", "Fish", "Dairy", "Eggs", and "Plant-based".
2. **Easy PEB1:** Participants execute a series of 30 rounds where in each round, they have 10 seconds to classify pictures by their main ingredients. Here, they have only two categories: "Animal-based products" and "Plant-based products".

In the two pilot studies, we fine-tuned the difficulty of the real effort task to ensure the easy version is perceived as easier (see Figure 4.A.3 in Appendix 4.A.A).

Manipulation of the domain of PEB2 We seek to test the existence of cross-domain behavioural spillovers. To this end, participants are randomly assigned to two groups:

1. **Environment-related PEB2:** In this condition, participants decide whether to sign a petition supporting the redesign of the German car tax by introducing a bonus-malus system that accounts for CO2 emissions.

2. **Health-related PEB2:** Participants decide whether to sign a petition supporting policies to reduce individual loneliness and social isolation

In the pilot studies, we checked that respondents perceive the environment-related petition as more similar to the real effort task than the health-related petition (see Figure 4.A.3 in Appendix 4.A.A).

Instrumental variable: We rely on an instrumental variable strategy to get causal estimates of the behavioural spillover effects. We manipulate the choice environment to unconsciously alter respondents' decisions to do PEB1 without directly affecting participation in PEB2. Namely, respondents are randomised into two conditions:

1. **Easy access to PEB1:** When deciding whether to participate in PEB1, we increase the salience of the "I want to participate" button. We also decrease the salience of the "End task" button that participants click to stop doing PEB1.

2. **Hard access to PEB1:** When deciding whether to participate in PEB1, we increase the salience of the "I do not want to participate" button. We also increase the salience of the "End task" button.

When estimating the indirect spillover effect of doing PEB1 on PEB2, we instrument the decision to do PEB1 by a dummy, capturing respondents' allocation to one of these two choice architectures. See Section IV for more details on the estimation strategy.

Data Collection and Randomisation

Our experiment follows a fractional design. Namely, data collection was planned in three waves. In wave one, we allocated respondents to our four treatment texts. We use the hard version of PEB1 and the environment-related version of PEB2. In wave two, we allocate respondents in the control or treatment text (doom-and-gloom or win-win) that yields the largest direct spillover effect in absolute terms in wave one. In this wave, respondents do the easy version of PEB1 and the environment-related version of PEB2. In wave three, we allocate respondents to the control group or the selected treatment text. Here, participants do the easy version of PEB1 and the health-related version of PEB2. Within each wave, participants' allocation to the treatment texts is randomised. We worked with Norstat, our panel data provider, to randomise respondents' allocation to the waves.

IV Analysis Plan

To estimate the effect of the win-win framing (hypothesis 1) and the doom-and-gloom framing versus the control (hypothesis 2), we use probability linear models estimated by ordinary least squares:

$$PEB1_i = \alpha + \beta \cdot T_i + \upsilon_i \tag{IV.1}$$

In testing hypothesis 1, the dummy T_1 is equal to 1 if respondent i is in the win-win group and zero if i is in the control group. Symmetrically, T_1 equals one if respondent i is in the doom-and-gloom group and zero if i is in the control group for hypothesis 2. The dependent variable is either binary and captures the decision to do PEB1 or continuous, capturing the number of rounds participants execute.

For testing hypotheses 3, 4 and 5, we use linear models estimated by two-stage least-squares, such as:

$$PEB1_i = \alpha + \beta \cdot IV_i + \delta \cdot T_1 + \upsilon_i \quad \text{(Stage 1)} \tag{IV.2}$$

$$PEB2_i = \alpha' + \beta' \cdot \widehat{PEB1}_i + \delta' \cdot T_1 + \upsilon'_i \quad \text{(Stage 2)} \tag{IV.3}$$

The estimate of β' corresponds to the behavioural spillover effect $\partial_{PEB1} PEB2$. The estimate of δ' corresponds to the direct spillover effect $\partial_T PEB2$. When testing hypothesis 3, T equals one when respondents are allocated to either the win-win or the doom-and-gloom groups and zero if they are in the control group. For testing hypothesis 4, the dummy T_i equals one if respondent i is in the win-win group and zero if in the control group. Symmetrically, T_i equals one if respondent i is in the doom-and-gloom group and zero if in the control group when testing hypothesis 5. IV_i is equal to 1 when respondent i is allocated to a condition where doing PEB1 is facilitated by the salience nudge, 0 otherwise. $\widehat{PEB1}_i$ are the predicted values from regression (IV.2).

To estimate the indirect spillover effect $\partial_T PEB1 \times \partial_{PEB1} PEB2$, we fit the additional regression:

$$PEB2_i = \alpha'' + \beta'' \cdot T_i + \upsilon''_i \tag{IV.4}$$

The indirect spillover effect is estimated as $\beta'' - \delta'$. We calculate a bootstrapped 95%-CI for $\beta'' - \delta'$ based on 10.000 resamples of the original sample. We consider the effect significant if the 95%-CI does not include 0.

In testing hypothesis 6, we also use linear models estimated by two-stage least squares, such as:

$$PEB1_i = \alpha + \beta_1 \cdot IV_i + \beta_2 \cdot T_1 + \beta_3 \cdot D_i + \upsilon_i \quad \text{(Stage 1a)} \tag{IV.5}$$

$$(PEB1 \cdot D)_i = \alpha' + \beta'_1 \cdot (IV \cdot D)_i + \beta'_2 \cdot T_1 + \beta'_3 \cdot D_i + \upsilon'_i \quad \text{(Stage 1b)} \tag{IV.6}$$

$$PEB2_i = \alpha'' + \beta''_1 \cdot \widehat{PEB1}_i + \beta''_2 \cdot T_1 + \beta''_3 \cdot D_i + \beta''_4 \cdot (\widehat{PEB1 \cdot D})_i + \upsilon''_i \quad \text{(Stage 2)} \tag{IV.7}$$

Here, D_i is a dummy equal to 1 when respondent i is allocated to the easy PEB1 condition.

Estimating coefficient β''_4 allows us to test hypothesis 6: whether the difficulty of PEB1 mediates the effect of doing PEB1 on PEB2.

Finally, in testing hypotheses 7 and 8, we use the following linear models estimated by two-stage least squares:

$$PEB1_i = \alpha + \beta_1 \cdot IV_i + \beta_2 \cdot T_1 + \beta_3 \cdot H_i + \beta_4 \cdot (T \cdot H)_i + \upsilon_i \quad \text{(Stage 1a)} \tag{IV.8}$$

$$(PEB1 \cdot H)_i = \alpha' + \beta'_1 \cdot (IV \cdot H)_i + \beta'_2 \cdot T_1 + \beta'_3 \cdot H_i + \beta'_4 \cdot (T \cdot H)_i + \upsilon'_i \quad \text{(Stage 1b)} \tag{IV.9}$$

$$PEB2_i = \alpha'' + \beta''_1 \cdot \widehat{PEB1}_i + \beta''_2 \cdot T_1 + \beta''_3 \cdot H_i + \beta''_4 \cdot (\widehat{PEB1 \cdot H})_i + \beta''_5 \cdot (T \cdot H)_i + \upsilon''_i \quad \text{(Stage 2)} \tag{IV.10}$$

Here, H_i is a dummy equal to 1 when respondent i is allocated to the condition where PEB2 consists of signing a petition for supporting policies to reduce individual social isolation and loneliness.

Estimating coefficient β''_4 allows us to test hypothesis 7: whether the framing of PEB2 mediates the effect of doing PEB1 on PEB2. Hypothesis 8 is tested by estimating coefficient β''_5.

As part of an exploratory analysis, we investigate the heterogeneity of our treatment effects. To do so, we interact our moderators with the dummy variable capturing respondents' allocation to the treatment texts (T_i). More specifically, we interact the variables capturing pro-environmental and altruistic values with the dummy capturing allocation to the "win-win" narrative. We also interact the variable capturing guilt-proneness with the dummy capturing allocation to the "doom-and-gloom" narrative.[8] This heterogeneity analysis is to be conducted for hypotheses 1, 2, 4 and 5.

As part of another exploratory analysis, we test whether the mere exposure to information on PEB1 triggers spillover effects. In doing so, we compare respondents allocated to the control group with those assigned to the placebo group. We use similar statistical models as those used to investigate hypotheses 1, 2, 3, 4 and 5.

For robustness checks, we control for respondents' social-demographic and attitudinal information. We also compute re-randomised p-values based on Young (2019)'s procedure to ensure outliers do not drive statistical significance. Using linear models with dichotomous variables is valid as long as the predicted values of these models are bounded between 0 and 1, which is the case in our study. Yet, we run further robustness checks to ensure that our results are not an artefact of the statistical models chosen. Namely, we fit probit models to estimate equations (IV.1) and rely on Rivers and Vuong (1988)'s procedure to estimate equations (IV.3), (IV.7), and (IV.10).

Table 4.1: Descriptive Statistics by Wave

	Wave 1 (N=5,492)	Wave 2 (N=2,622)	Wave 3 (N=2,556)	Total (N=10,670)	p-value
Age					< 0.001
Median	45-54 yo (18.8%)	45-54 yo (23.3%)	45-54 yo (23.0%)	45-54 yo (20.9%)	
Min	18-24 yo (6.5%)	18-24 yo (4.6%)	18-24 yo (6.3%)	18-24 yo (6.0%)	
Max	>65 yo (26.5%)	>65 yo (18.1%)	>65 yo (18.5%)	>65 yo (22.5%)	
Income					< 0.001
Median	40k-49k€ (11.9%)	40k-49k€ (12.3%)	40k-49k€ (12.8%)	40k-49k€ (12.2%)	
Min	<10k€ (7.4%)	<10k€ (7.3%)	<10k€ (6.7%)	<10k€ (7.2%)	
Max	>150k€ (2.4%)	>150k€ (2.2%)	>150k€ (2.5%)	>150k€ (2.4%)	
Gender					0.208
Female	2782 (51.0%)	1282 (49.3%)	1262 (49.9%)	5326 (50.3%)	
Male	2612 (47.9%)	1282 (49.3%)	1224 (48.4%)	5118 (48.4%)	
Other	59 (1.1%)	34 (1.3%)	41 (1.6%)	134 (1.3%)	
Education					0.006
Median	Abitur or eq. (24.1%)	Abitur or eq. (24.5%)	Abitur or eq. (24.6%)	Abitur or eq. (24.3%)	
Min	No education (0.4%)	No education (0.4%)	No education (0.1%)	No education (0.4%)	
Max	PhD (1.6%)	PhD (2.1%)	PhD (1.7%)	PhD (1.7%)	
Political Belief					0.401
Mean	4.720	4.697	4.646	4.697	
Median	5 (31.5%)	5 (29.6%)	5 (27.6%)	5 (30.1%)	
Min	0 (2.7%)	0 (2.9%)	0 (2.8%)	0 (2.8%)	
Max	10 (2.7%)	10 (3.1%)	10 (2.3%)	10 (2.7%)	
SD	2.180	2.217	2.227	2.200	

Note: We use a Wilcoxon test to check for differences in political beliefs across waves. We use a Chi-square test to check for gender differences. We use trend tests to check for differences in education, income, and age.

V Results

V.A Sample Characteristics

We collected the first wave of data between 26^{th} October 2023 and 8^{th} January 2024. In total, 7,837 respondents took the survey, and 29.9% failed the attention checks, leaving us with a final sample of 5,492 respondents. Following the criteria set in the registered report, we focused on the doom-and-gloom narrative and the control group in the next two waves. We collected the second wave of data between 9^{th} January 2024 and 6^{th} February 2024. In total, 3,724 respondents took

[8]To construct the moderating variables capturing pro-environmental values, altruistic values and guilt-proneness, we average the responses to the questions measuring each of these respective elements. We then create a dummy equal to one when respondents score above the median for each measure and zero otherwise (see Appendix B in the supplementary information).

the survey. We excluded 29.6% of participants as they did not pass the attention checks. This left us with a final sample size of 2,622 respondents. Wave three was conducted simultaneously with wave two. Overall, 3,769 respondents took the survey, and 32.2% were excluded, leaving us with 2,556 participants. The final sample comprises 10,670 observations, above our target of 10,000 respondents.

Table 4.1 presents descriptive statistics by wave. The median respondent has a high school diploma (abitur in German) or equivalent, earns between 40,000 and 49,999 euros per year, is 45 to 54 years old, and is centrist on the political spectrum. Our sample is gender balanced (49.9% female, 48% male, 1.3% other). To check if randomisation worked, we tested differences in gender, age, income, education and political beliefs across treatment groups. Within each wave, randomisation was successful despite a small imbalance in income in wave three. Being richer positively correlates with being in the condition where the salience nudge hardens the uptake of PEB1 (Cohen's d=0.088, p-value = 0.023). We observe small statistical differences between waves. Participants in wave one are slightly poorer than participants in wave two (d=-0.081, $p<0.01$) and three (d=-0.129, $p<0.01$). They are also slightly older than participants in wave two (d=0.053, p=0.024) and three (d=0.086, $p<0.01$) and less educated than participants in waves two and three combined (d=-0.062, p=0.006). Participants in wave two are also slightly poorer than those in wave three (d=-0.048, p=0.094). Despite being statistically significant, these differences remain small. Furthermore, Figure 4.A.7 in Appendix 4.A.A shows similar distributions of these covariates across waves.

On average, around 70% of the full sample did the real-effort task. 85% of the participants who did the real effort task completed it by doing the 30 rounds (see upper panel of Figure 4.1). We observe more variations when looking at respondents' performances during the real-effort task (see lower panel of Figure 4.1). We construct the performance score by averaging the correct rounds

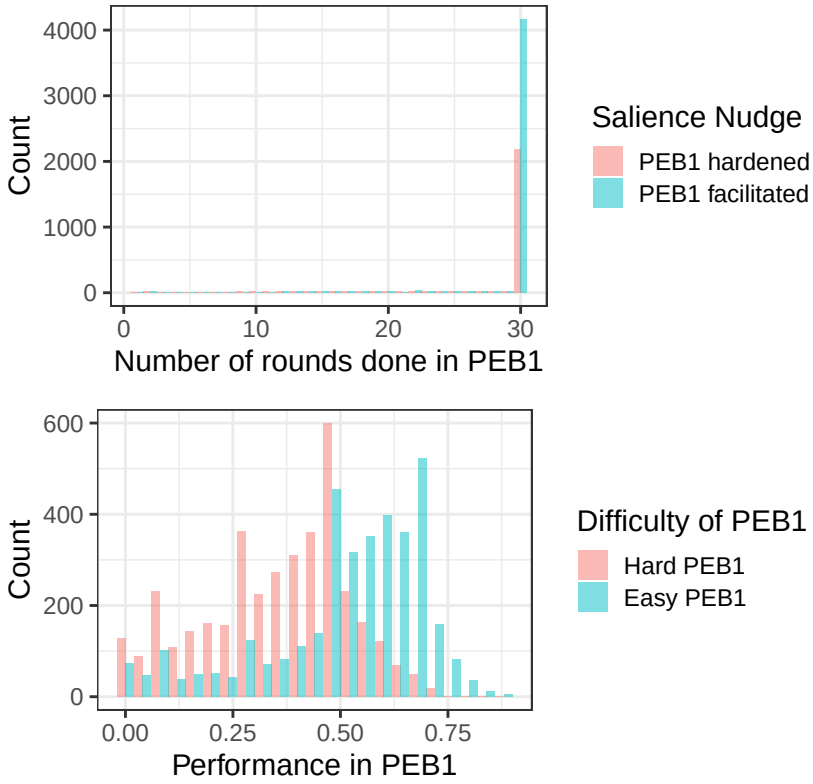


Figure 4.1: Histograms of the number of rounds (up) and performances (down) for PEB1

over 30. We deem an answer correct when it matches what was reported by respondents in a previous experiment.[9] Finally, on average, 29% of our sample signed the petition. We observe a large difference between the environment-related and the health-related petition, with 27% of signatories in the former case and 38% in the latter.

V.B Effect of treatment texts on PEB1

Win-Win narrative vs neutral information: The first and the second columns of Table 4.2 present

[9]These respondents were followed over several months and reported the food they were eating as well as a picture of the food (that we used for the present experiment). We use their answers to check the performance of the participants in the present study. When participants did fewer than 30 rounds, we counted the rounds they skipped as incorrect. As such, doing 15 correct rounds before stopping yields a score of 0.5.

Table 4.2: Effect of Treatment Texts on PEB1 - Hypotheses 1 and 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.525***	13.810***	0.518***	13.520***	0.491***	12.997***	0.503***	13.283***
	(0.016)	(0.452)	(0.016)	(0.449)	(0.015)	(0.440)	(0.013)	(0.389)
Salience nudge	0.367***	11.093***	0.381***	11.670***	0.414***	12.251***	0.389***	11.658***
	(0.016)	(0.478)	(0.016)	(0.471)	(0.015)	(0.460)	(0.011)	(0.333)
	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01	q<0.01
Win-Win vs.	−0.023	−0.425						
Control	(0.016)	(0.475)						
	q=0.159	q=0.372						
Doom & Gloom vs.			−0.024	−0.367				
Control			(0.016)	(0.472)				
			q=0.147	q=0.433				
Control vs.					0.011	0.230		
Placebo					(0.016)	(0.463)		
					q=0.488	q=0.609		
All vs.							−0.004	−0.022
Placebo							(0.013)	(0.379)
							q<0.01	q<0.01
Num Obs	2727	2727	2760	2760	2757	2757	5492	5492
R2	0.161	0.167	0.173	0.182	0.205	0.203	0.178	0.181
F-stat of salient nudge	515.097	538.467	576.188	615.208	720.309	709.462	1195.218	1222.825
Data included	Wave 1 without placebo & doom-and-gloom		Wave 1 without placebo & win-win		Wave 1 without win-win & doom-and-gloom		Wave 1	
Outcome	PEB1 (binary)	PEB1 (continuous)	PEB1 (binary)	PEB1 (continuous)	PEB1 (binary)	PEB1 (continuous)	PEB1 (binary)	PEB1 (continuous)

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of the win-win, doom-and-gloom, and neutral information texts on respondents' participation in the real effort task. Columns one and two present the average treatment effects (ATEs) of adding win-win arguments to neutral information. Columns three and four present the ATEs of adding doom-and-gloom arguments to neutral information. Columns five and six present the ATEs of providing neutral information compared to the placebo group. Columns seven and eight present the ATEs of being in a treatment condition or the control group compared to the placebo group. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

estimates of the effect of reading Win-Win arguments on the binary decision to do PEB1 and the number of rounds executed.[10] Stressing the benefit of action does not increase respondents' participation in the real effort task compared to providing neutral information. There is no support for hypothesis 1. This null finding is not an artefact of our statistical method. It remains non-significant when using probit models or controlling for social-demographic covariates and attitudinal information (see Table 4.A.8 in Appendix 4.A.C). In table 4.A.1, we present the results of an exploratory analysis testing if respondents' altruism and pro-environmental attitude mediate this effect. We do not find evidence for heterogeneity in this result.

Doom-and-Gloom narrative vs neutral information: The third and the fourth columns of Table 4.2 present the effect of reading Doom-and-Gloom arguments on the decision to do PEB1 and the number of rounds executed. Emphasising the cost of inaction does not increase respondents' participation in the real effort task compared to providing neutral information. As such, there is no support for hypothesis 2. We even detect a backfiring effect when pooling data from all the waves together: reading Doom-and-Gloom arguments slightly reduces participation in the real effort task (see Table 4.A.9 in Appendix 4.A.C). In table 4.A.1, we present the results of an exploratory analysis testing if respondents' proneness to guilt mediates this effect. We do not find evidence for heterogeneity in this result.

Neutral information vs placebo: The fifth and the sixth columns of Table 4.2 present the effect of providing neutral information on the decision to do PEB1 and the number of rounds executed. This analysis is exploratory. Providing neutral information does not affect participation in PEB1 compared to the placebo group. Columns seven and eight of table 4.2 present the results from an

[10] Analyses were conducted on R using the package *estimatr* (Blair et al., 2022).

exploratory analysis where we compare the placebo group with all the other groups. Again, we do not detect a significant difference in participation in the real effort task.

Looking at performances: We also explore if our treatment interventions affect performances in the real effort task. We do not detect any significant effects of the win-win and doom-and-gloom arguments or neutral information (see Table 4.A.4 in Appendix 4.A.B).

Effect of the Salient Nudge: We find a strong and significant effect of the salient nudge on respondents' participation in the real effort task. Facilitating participation in PEB1 increases the likelihood of doing it by 37.1 percentage points ($p<0.01$) compared to hardening participation in PEB1. It also makes respondents do 11 more rounds on average ($p<0.01$). The F-statistics associated with the dummy capturing respondents' allocation in the salience nudge is way above the convention threshold of 10. This indicates that our instrumental variable is strong (Bound et al., 1995; Staiger and Stock, 1997; Stock and Yogo, 2002).

V.C Spillover Effects on PEB2

Indirect Spillover Effect: Table 4.3 presents estimates of the effect of doing the real effort task on respondents' likelihood of signing the environment-related petition ($\partial_{PEB1}PEB2$). We also report 95% confidence intervals of the indirect spillover effect $\partial_T PEB1 \times \partial_{PEB1}PEB2$ estimated with bootstrap. Participating in the real effort task does not causally increase respondents' likelihood of signing the environment-related petition, despite positive correlations (see Table 4.A.3 in Appendix 4.A.B). As such, there is no support for hypothesis 3. This null finding is not an artefact of the statistical method used. It remains non-significant when using Rivers and Vuong (1988)'s specification or when controlling for social-demographic covariates and attitudinal information (see Table 4.A.10 in Appendix 4.A.C). In an exploratory analysis, we further tested whether higher per-

Table 4.3: Behavioural Spillover Effects on PEB2 - Hypotheses 3, 4 and 5

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.300***	0.297***	0.298***	0.295***	0.270***	0.268***	0.252***	0.252***
	(0.036)	(0.033)	(0.034)	(0.031)	(0.030)	(0.028)	(0.024)	(0.022)
PEB1	−0.040	−0.001	−0.037	−0.001	−0.035	−0.001	−0.010	0.000
($\partial_{PEB1} PEB2$)	(0.047)	(0.002)	(0.044)	(0.001)	(0.040)	(0.001)	(0.031)	(0.001)
	q=0.092	q=0.087	q=0.101	q=0.101	q=0.084	q=0.083	q=0.515	q=0.516
Win-Win vs.	0.003	0.003						
Control ($\partial_T PEB1$)	(0.017)	(0.017)						
	q=0.865	q=0.849						
Doom & Gloom vs.			−0.011	−0.010				
Control ($\partial_T PEB1$)			(0.017)	(0.017)				
			q=0.533	q=0.538				
Control vs.					0.027	0.027		
Placebo ($\partial_T PEB1$)					(0.017)	(0.017)		
					q=0.105	q=0.109		
All vs.							0.024	0.024
Placebo ($\partial_T PEB1$)							(0.014)	(0.014)
							q=1.000	q<0.01
Num Obs	2727	2727	2760	2760	2757	2757	5492	5492
Data included	Wave 1 without placebo & doom-and-gloom		Wave 1 without placebo & win-win		Wave 1 without win-win & doom-and-gloom		Wave 1	
Outcome	PEB2							
PEB1	Binary	Continuous	Binary	Continuous	Binary	Continuous	Binary	Continuous
95% bootstrapped CI of $\partial_T PEB1 \times \partial_{PEB1} PEB2$	[-0.003, 0.004]	[-0.003, 0.002]	[-0.003, 0.004]	[-0.003, 0.003]	[-0.003, 0.002]	[-0.002, 0.002]	[-0.002, 0.002]	[-0.001, 0.001]

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of doing the real effort task — instrumented by the salience nudge — on respondents' likelihood of signing the environment-related petition. It also presents the effect of the win-win, doom-and-gloom, and neutral information texts on respondents' likelihood of signing the petition. Columns one and two present the direct spillover effects triggered by win-win arguments compared to the control condition. Columns three and four present the direct spillover effects triggered by doom-and-gloom arguments compared to the control condition. Columns five and six present the direct spillover effect triggered by providing neutral information compared to the placebo group. Columns seven and eight present the direct spillover effect of being in a treatment condition or the control group compared to the placebo group. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.4: Effect of PEB1 Difficulty and Domain Similarity - Hypotheses 6, 7 and 8

	(1)	(2)	(3)	(4)
Constant	0.252***	0.252***	0.284***	0.282***
	(0.024)	(0.022)	(0.026)	(0.024)
Difficulty	0.002	0.002	0.005	0.006
	(0.042)	(0.040)	(0.012)	(0.012)
	q=0.972	q=0.963	q=0.672	q=0.625
PEB1	−0.010	0.000	−0.023	−0.001
	(0.031)	(0.001)	(0.033)	(0.001)
All vs. Placebo	0.024*	0.024*		
	(0.014)	(0.014)		
PEB1 x Difficulty	0.001	0.000		
	(0.058)	(0.002)		
	q=0.989	q=0.986		
Doom & Gloom vs. Control			−0.001	−0.001
			(0.012)	(0.012)
Health Petition			0.016	0.020
			(0.047)	(0.044)
			q=0.730	q=0.648
PEB1 x			0.106	0.003
Health Petition			(0.061)	(0.002)
			q=0.079	q=0.081
Doom & Gloom vs. Control x			0.042	0.041
Health Petition			(0.023)	(0.023)
			q=0.064	q=0.073
Num Obs	8112	8112	7936	7936
Data included	Wave 1+2		Wave 1+2+3 without placebo & win-win	
Outcome	PEB2			
PEB1	Binary	Continuous	Binary	Continuous

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table displays the effects of varying the difficulty of the real-effort task and the cause supported by the petition on direct and indirect spillover effects. In all columns, the variables capturing respondents' participation in PEB1 are instrumented by the salient nudge. The variable *Difficulty* is a dummy equal to one when the difficulty of PEB1 is reduced, zero otherwise. The variable *Health Petition* is a dummy equal to one when the petition is about loneliness. Columns one and two present results from statistical model (IV.7). Columns three and four present results from statistical model (IV.10). Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

formances in the real effort task increased respondents' likelihood of signing the petition. Results are presented in Table 4.A.5 in Appendix 4.A.B. To causally estimate the effect of higher performances on signing the petition, we instrumented performances by the difficulty of PEB1. Again, we find no evidence that doing well in the real effort task increased respondents' likelihood of signing the petition.

Direct Spillover Effect: Table 4.3 presents estimates of the direct effects triggered by our treatment texts on PEB2 ($\partial_T PEB2$). We find no evidence that the Win-Win or Doom-and-Gloom treatments directly affected respondents' likelihood of signing the environment-related petition. As such, there is no support for hypotheses 4 and 5. This null finding is not an artefact of the statistical method used. It remains non-significant when using Rivers and Vuong (1988)'s specification or when controlling for social-demographic covariates and attitudinal information (see Table 4.A.11 and 4.A.12 in Appendix 4.A.C). Further exploratory analyses suggest a negative direct effect of the win-win treatment for people being more altruistic than the median of respondents. However, this effect does not pass multiple hypothesis correction. We find no further heterogeneity.

Difficulty of PEB1: The first two columns of Table 4.4 presents estimate of the effect of making PEB1 easier on the indirect behavioural spillover effect ($\partial_{PEB1} PEB2$). We do not find that making PEB1 easier changed the indirect spillover effect. As such, there is no support for hypothesis 6. Here again, the statistical method used does not influence our results (see Table 4.A.13 in Appendix 4.A.C).

Similarity between PEB1 and PEB2: We find suggestive evidence that the indirect spillover effect is higher with the health-related petition. We also find suggestive evidence that the direct spillover effect triggered by the Doom-and-Gloom narrative is higher with the health-related pe-

tition. Although these effects do not pass p-value correction, they are robust to non-linear and linear specifications where controls are added (see Table 4.A.14 in Appendix 4.A.C). Re-randomise p-values also indicate that outliers do not drive statistical significance.

VI Discussion

Our findings align with recent meta-analyses indicating that information nudges are ineffective (Maier et al., 2022; Nisa et al., 2019). We see five non-mutually exclusive explanations for our results. The first relates to participants' attention. Respondents may not have read or processed the information presented in the articles. However, our data does not support this explanation. On average, 75% of the sample correctly answered our manipulation checks. Furthermore, our treatment effects are not mediated by the time spent reading the articles, whilst we would have expected that the shorter the time spent reading, the smaller the effect (see Table 4.A.6 in Appendix 4.A.B).

The second explanation regards heterogeneity. Our treatment interventions may have yielded opposite effects on different subsamples, which would average to zero on the full sample. Our exploratory analyses do not support this explanation either.

Third, we designed a consequential real-effort task. In that, we align with other experimental approaches seeking to move beyond pro-environmental intentions (e.g., Berger and Wyss 2021; Lange et al. 2018; Lange and Dewitte 2022). In our case, respondents' payment was uncorrelated with the time spent doing the survey. In other words, they faced an opportunity cost from participating in the real-effort task. Previous studies stressing the benefit of acting or the costs of inaction looked at intentions (Homar and Cvelbar, 2021). Our results might thus indicate that such approaches are

not powerful enough to translate people's intentions into actions. This might be even more the case that most respondents did the real effort task. The remaining 30% that did not do it might be the hardest to convince. Furthermore, amongst those that did the task, only 15% did fewer than 30 rounds. This might not have provided enough variations to observe an effect at the intensive margin.

The fourth explanation concerns respondents' trust in the source of information. Previous work shows that who delivers the information matters (Alsan and Eichmeyer, 2024; Banerjee et al., 2022). The success of win-win or doom-and-gloom arguments might depend on who initiates such attempts. It might have been that respondents exposed to these arguments sensed the experimenter was trying to steer them to do the real effort task by playing on their feelings. This could explain why we do not observe an effect of the win-win narrative and a (small) backfiring effect of the doom-and-gloom narrative. Similarly, not trusting the sender could have induced respondents not to believe that doing the real effort task would impact the environment. Yet, among respondents who did the task, those in the treated and control groups were more likely to feel they made an effort for the environment compared to the placebo group (see Table 4.A.7 in Appendix 4.A.B). This contradicts this explanation.

Fifth, the information treatment may have provided respondents with enough information to update their beliefs without significantly changing their intentions to act pro-environmentally. This is the case when respondents are motivated by things other than a genuine desire to solve environmental issues (e.g., warm glow, boost in self-esteem, social recognition). Our real effort task did not provide such contingent rewards. Contrary to other well-known pro-environmental behaviours, it was new to respondents. They could not have priors on whether doing it is a social norm or form any habits. Furthermore, respondents acted on their "carbon handprint" when doing the

real effort task, i.e., others' carbon footprint. The pathway through which it reduces carbon emissions is indirect. As such, doing it may not have yielded a strong feeling of achievement. In other words, it is possible that there were no "ropes" that our treatment interventions could pull to spur participation.

This fifth explanation could also explain why we did not observe any indirect spillover effects. In Chapter 2, we modelled behavioural spillover effects to occur for "identity-enhancing" behaviours. Again, contrary to vegetarian food choices, as tested in Chapter 3, its "artificial" nature might not have induced the processes leading to such spillover effects.

Another explanation for spillover effects is that *successfully* doing first pro-environmental deeds simply motivates people to do more (Lauren et al., 2016). Here, our exploratory analyses seem to rule out this explanation. Higher performances increase the feeling that one has made an effort for the environment. Yet, it does not increase respondents' likelihood of signing the environment-related petition (see Table 4.A.5 in appendix 4.A.B).

Truelove et al. (2014) hypothesise that the difficulty of the first action moderates behavioural spillover effects. We do not find evidence for this assumption. However, we only studied one dimension of behavioural difficulty. For instance, respondents might perceive some pro-environmental behaviours as difficult because doing them implies transgressing a social norm, irrespective of how easy it is to execute the behaviour.

Finally, we find suggestive evidence of "cross-domain" spillover effects. However, our categorisation of the two tasks in separate domains can differ from respondents' perceptions. Indeed, the real effort task could be perceived as more pro-social than pro-environmental, given that respondents benevolently helped us encode food pictures. Similarly, the health-related petition about youth

loneliness might have been perceived as more pro-social than the environment-related one, calling for a carbon tax on cars. Therefore, the pro-social nature of behaviours might be a stronger driver of spillover effects than their pro-environmental nature. These interpretations should be taken cautiously, as our results did not pass multiple hypothesis correction.

VII Conclusion

Providing information as a standalone intervention or coupling it with doom-and-gloom or win-win arguments does not foster pro-environmental action. Despite positive correlations, we also do not find causal evidence of behavioural spillover effects. In other words, doing the first pro-environmental action did not alter respondents' likelihood to sign an environment-related petition, irrespective of the difficulty of the first action. Finally, we find suggestive evidence of cross-domain spillovers, with the caveat that this result does not pass multiple hypothesis correction.

Two main implications can be derived from our findings. First, our experiment questions the effectiveness of communication campaigns relying on information provision, alarmist warnings, or blissful optimistic messages. Second, our findings suggest that doing a pro-environmental action does not necessarily trigger behavioural spillover effects.

We see two explanations for these conclusions. First, we relied on a real effort task to proxy pro-environmental action instead of measuring intentions. Our treatment interventions might only shift intentions without altering actual behaviours. This would explain our inability to replicate previous findings. Second, our setting was peculiar as the real effort task was *new* to participants. People did not know if doing it was socially desirable or did not form any habits. The only reward associated with doing it is the mere knowledge that one made an effort for the environment. This

might not be enough to motivate people to act. This may explain why we did not observe any first- and second-order effects.

These two explanations imply different recommendations. Whilst the first suggests that the effect of soft policies has been inflated, the second implies that the proxies used to measure pro-environmental actions miss some important psychological dimensions associated with doing them. Future work should investigate which channel is most likely to be at play.

Informed consent and ethics approval: The ethical approval was obtained from the German Association for Experimental Economic Research e.V. (No. bh41mpGC) and the London School of Economics (reference 133444).

Funding Statement: The German Federal Ministry of Education and Research funded this study.

Code and Data Availability: The code and data for the analysis are available upon request.

Bibliography

Alacevich, C., Bonev, P., and Söderberg, M. (2021). Pro-environmental interventions and behavioral spillovers: Evidence from organic waste sorting in sweden. *Journal of Environmental Economics and Management*, 108:102470.

Alsan, M. and Eichmeyer, S. (2024). Experimental evidence on the effectiveness of nonexperts for improving vaccine demand. *American Economic Journal: Economic Policy*, 16(1):394–414.

Banerjee, A., Alsan, M., BREZA, E., Chowdhury, A., Duflo, E., Olken, B., Chandrasekhar, A., and Goldsmith-Pinkham, P. (2022). Can a trusted messenger change behavior when information is

plentiful? evidence from the first months of the covid-19 pandemic in west bengal. *Technical report*.

Bech-Larsen, T. and Kazbare, L. (2014). Spillover of diet changes on intentions to approach healthy food and avoid unhealthy food. *Health Education*. Publisher: Emerald Group Publishing Limited.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Berger, S. and Wyss, A. M. (2021). Measuring pro-environmental behavior using the carbon emission task. *Journal of environmental psychology*, 75:101613.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). *estimatr: Fast Estimators for Design-Based Inference*. https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr.

Bouman, T., Steg, L., and Kiers, H. A. (2018). Measuring values in environmental research: a test of an environmental portrait value questionnaire. *Frontiers in psychology*, 9:564. Publisher: Frontiers Media SA.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

Carlsson, F., Jaime, M., and Villegas, C. (2021). Behavioral spillover effects from a social information campaign. *Journal of Environmental Economics and Management*, 109:102325. Publisher: Elsevier.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018). Putting Your Money Where Your

Mouth Is: An Experimental Test of Pro-Environmental Spillover From Reducing Meat Consumption to Monetary Donations. *Environment and Behavior*, 50(7):723–748. Publisher: SAGE Publications Inc.

Chess, C. and Johnson, B. B. (2007). *Information is not enough*, pages 223–233. Cambridge University Press.

Cialdini, R. B. (1994). *Interpersonal influence*, pages 195—-217. Allyn & Bacon.

Cialdini, R. B., Trost, M. R., and Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of personality and social psychology*, 69(2):318. Publisher: American Psychological Association.

Clot, S., Grolleau, G., and Ibanez, L. (2014). Smug alert! Exploring self-licensing behavior in a cheating game. *Economics Letters*, 123(2):191–194. Publisher: Elsevier.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the GASP scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947. Publisher: American Psychological Association.

Comin, D. A. and Rode, J. (2023). Do green users become green voters? Technical report, National Bureau of Economic Research.

Corazzini, L., Cotton, C., and Valbonesi, P. (2015). Donor coordination in project funding: Evidence from a threshold public goods experiment. *Journal of Public Economics*, 128:16–29. Publisher: Elsevier.

Davidson, D. J. and Kecinski, M. (2022). Emotional pathways to climate change responses. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2):e751.

Deck, C. and Murphy, J. J. (2019). Donors change both their level and pattern of giving in response to contests among charities. *European Economic Review*, 112:91–106. Publisher: Elsevier.

Dolan, P. and Galizzi, M. M. (2015). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16. Publisher: Elsevier.

Dolan, P., Galizzi, M. M., and Navarro-Martinez, D. (2015). Paying people to eat or not to eat? Carryover effects of monetary incentives on eating behaviour. *Social Science & Medicine*, 133:153–158. Publisher: Elsevier.

Ek, C. and Miliute-Plepiene, J. (2018). Behavioral spillovers from food-waste collection in Swedish municipalities. *Journal of Environmental Economics and Management*, 89:168–186. Publisher: Elsevier.

Festinger, L. (1962). Cognitive dissonance. *Scientific American*, 207(4):93–106. Publisher: JSTOR.

Filiz-Ozbay, E. and Uler, N. (2019). Demand for giving to multiple charities: An experimental study. *Journal of the European Economic Association*, 17(3):725–753. Publisher: Oxford University Press.

Freedman, J. L. and Fraser, S. C. (1966). Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, 4(2):195. Publisher: American Psychological Association.

Garvey, A. and Bolton, L. (2017). The licensing effect revisited: How virtuous behavior heightens the pleasure derived from subsequent hedonic consumption. *Journal of Marketing Behavior, Forthcoming*.

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., and Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, 58(1):179–187. Publisher: INFORMS.

Homar, A. R. and Cvelbar, L. K. (2021). The effects of framing on environmental decisions: A systematic literature review. *Ecological Economics*, 183:106950.

Jessoe, K., Lade, G. E., Loge, F., and Spang, E. (2021). Spillovers from behavioral interventions: Experimental evidence from water and energy use. *Journal of the Association of Environmental and Resource Economists*, 8(2):315–346. Publisher: The University of Chicago Press Chicago, IL.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206.

Krieg, J. and Samek, A. (2017). When charities compete: A laboratory experiment with simultaneous public goods. *Journal of behavioral and experimental economics*, 66:40–57. Publisher: Elsevier.

Lange, F. and Dewitte, S. (2022). The work for environmental protection task: A consequential web-based procedure for studying pro-environmental behavior. *Behavior research methods*, 54(1):133–145.

Lange, F., Steinke, A., and Dewitte, S. (2018). The pro-environmental behavior task: A laboratory measure of actual pro-environmental behavior. *Journal of Environmental Psychology*, 56:46–54.

Lanzini, P. and Thøgersen, J. (2014). Behavioural spillover in the environmental domain: An intervention study. *Journal of Environmental Psychology*, 40:381–390.

Lauren, N., Fielding, K. S., Smith, L., and Louis, W. R. (2016). You did, so you can and you will: self-efficacy as a mediator of spillover from easy to more difficult pro-environmental behaviour. *Journal of Environmental Psychology*, 48:191–199.

List, J. A. and Momeni, F. (2020). Leveraging upfront payments to curb employee misbehavior: Evidence from a natural field experiment. *European Economic Review*, 130:103601. Publisher: Elsevier.

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31):e2200300119.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315. Publisher: Nature Publishing Group.

Margetts, E. A. and Kashima, Y. (2017). Spillover between pro-environmental behaviours: The role of resources and perceived similarity. *Journal of Environmental Psychology*, 49:30–42. Publisher: Elsevier.

Mazar, N. and Zhong, C.-B. (2010). Do green products make us better people? *Psychological science*, 21(4):494–498. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Meer, J. (2017). Does fundraising create new giving? *Journal of Public Economics*, 145:82–93. Publisher: Elsevier.

Mullen, E. and Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual review of psychology*, 67(1):363–385.

Nabi, R. L., Gustafson, A., and Jensen, R. (2018). Framing climate change: Exploring the role of emotion in generating advocacy behavior. *Science Communication*, 40(4):442–468.

Nilsson, A., Bergquist, M., and Schultz, W. P. (2017). Spillover effects in environmental behaviors, across time and context: a review and research agenda. *Environmental Education Research*, 23(4):573–589. Publisher: Taylor & Francis.

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., and Faller, D. G. (2019). Meta-analysis of random-

ised controlled trials testing behavioural interventions to promote household action on climate change. *Nature communications*, 10(1):1–13.

Noblet, C. L. and McCoy, S. K. (2018). Does one good turn deserve another? Evidence of domain-specific licensing in energy behavior. *Environment and Behavior*, 50(8):839–863. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366. Publisher: Elsevier.

Schusser, S. and Bostedt, G. (2019). Green Behavioural (In) consistencies: Are Pro-environmental Behaviours in Different Domains Substitutes or Complements? *Environmental Economics*, 10(1):23–47.

Shipley, N. J. and van Riper, C. J. (2021). Pride and guilt predict pro-environmental behavior: A meta-analysis of correlational and experimental evidence. *Journal of Environmental Psychology*, page 101753. Publisher: Elsevier.

Sintov, N., Geislar, S., and White, L. V. (2019). Cognitive accessibility as a new factor in proenvironmental spillover: results from a field study of household food waste management. *Environment and Behavior*, 51(1):50–80. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Skurka, C., Niederdeppe, J., Romero-Canyas, R., and Acup, D. (2018). Pathways of influence in emotional appeals: Benefits and tradeoffs of using fear or humor to promote climate change-related intentions and risk perceptions. *Journal of Communication*, 68(1):169–193.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Thøgersen, J. (2004). A cognitive dissonance interpretation of consistencies and inconsistencies in environmentally responsible behavior. *Journal of environmental Psychology*, 24(1):93–103. Publisher: Elsevier.

Thøgersen, J. and Ölander, F. (2003). Spillover of environment-friendly consumer behaviour. *Journal of environmental psychology*, 23(3):225–236. Publisher: Elsevier.

Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013). For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy*, 57:160–171. Publisher: Elsevier.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138. Publisher: Elsevier.

Truelove, H. B., Yeung, K. L., Carrico, A. R., Gillis, A. J., and Raimi, K. T. (2016). From plastic bottle recycling to policy support: An experimental test of pro-environmental spillover. *Journal of Environmental Psychology*, 46:55–66. Publisher: Elsevier.

Wong-Parodi, G. and Feygina, I. (2021). Engaging people on climate change: The role of emotional responses. *Environmental Communication*, 15(5):571–593.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

4.A Appendix

4.A.A Figures

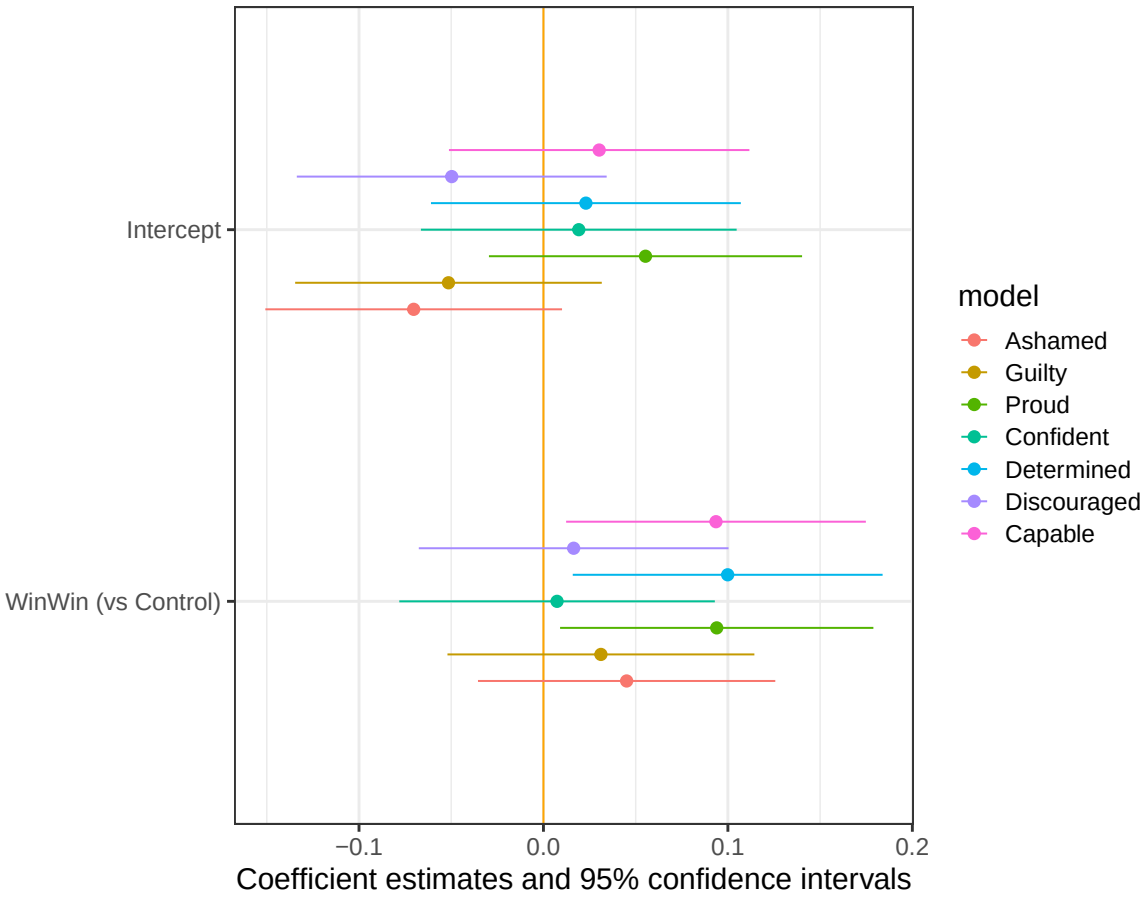


Figure 4.A.1: Feelings Associated with Win-win Arguments

Note: Results from the second pilot session. The graph displays coefficients from OLS regressions where respondents' reported feelings are regressed on their treatment allocations. There were 278 respondents in the control group and 265 in the win-win group.

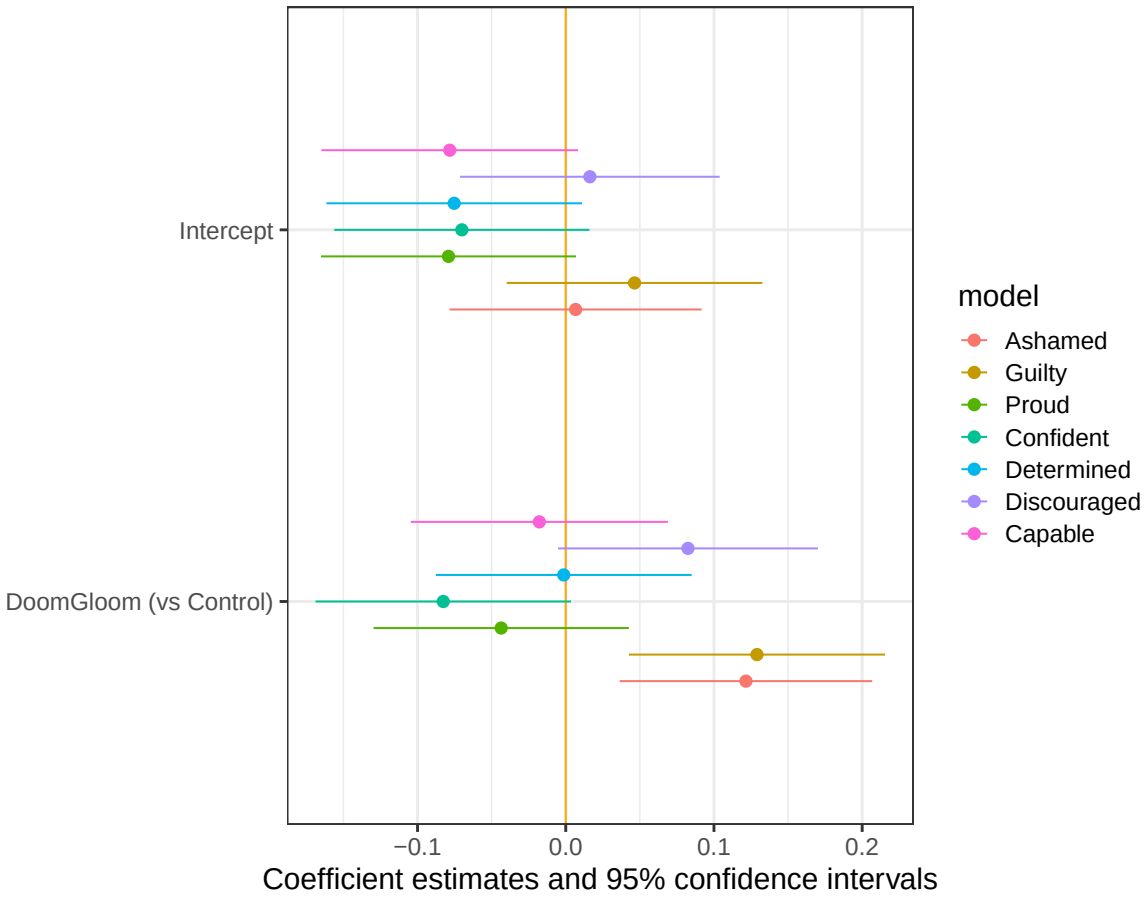


Figure 4.A.2: Feelings Associated with Doom-and-gloom Arguments

Note: Results from the second pilot session. The graph displays coefficients from OLS regressions where respondents' reported feelings are regressed on their treatment allocations. There were 278 respondents in the control group and 271 in the doom-and-gloom group.

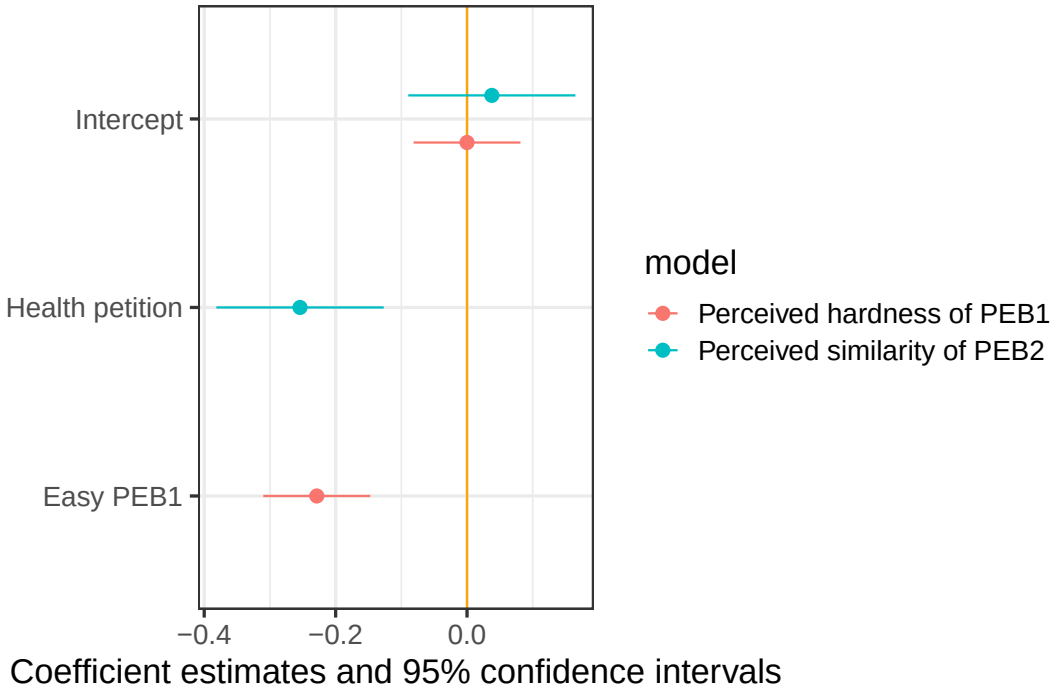


Figure 4.A.3: Perception that PEB1 is Difficult and Similar to PEB2

Note: Results from the first pilot session. The graph displays coefficients from OLS regressions where respondents' perception of the difficulty of PEB1 is regressed on their allocation to the easy or the hard version of PEB1 (red dots), and respondents' perception that PEB1 and PEB2 are similar is regressed on their allocation to the health or the environment-related version of PEB2 (blue dots). There were 574 respondents in the hard version of PEB1 and 260 in the easy version. There were 703 respondents in the environment-related version of PEB2 and 360 in the health-related version.

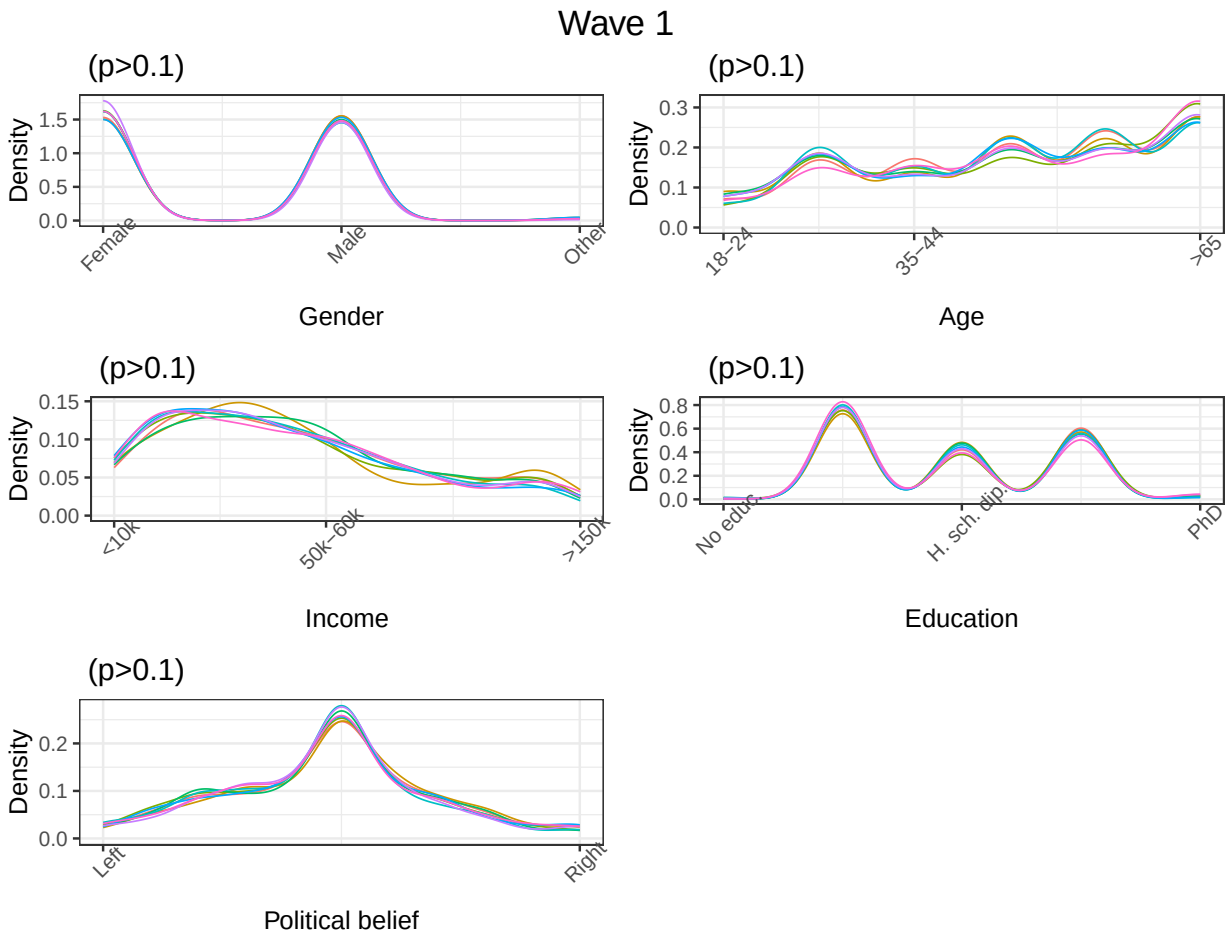


Figure 4.A.4: Distribution of Covariates Within Wave 1

Note: We use a Chi-square test to check for gender differences across treatment groups. We use a Kruskal-Wallis Rank Sum Test for the other covariates.

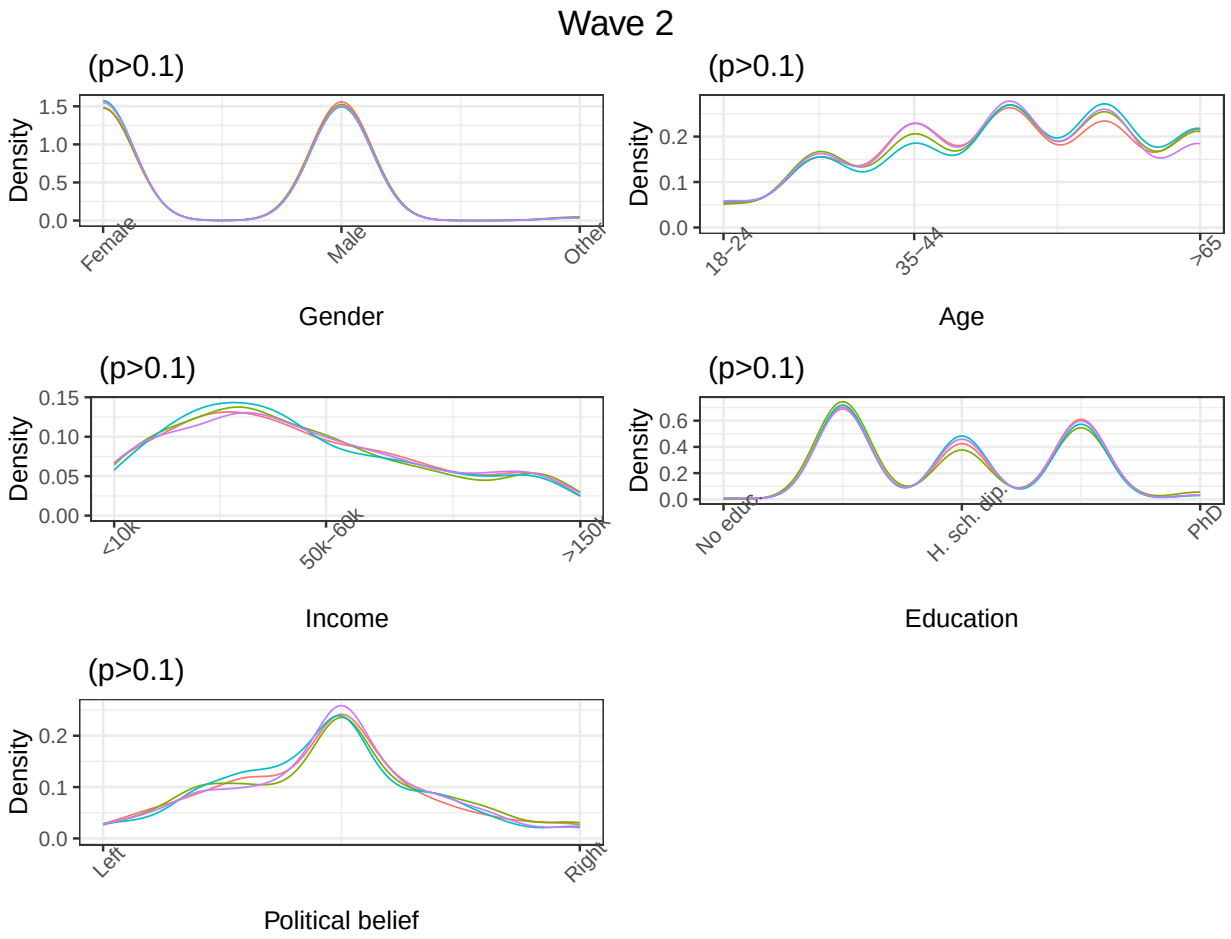


Figure 4.A.5: Distribution of Covariates Within Wave 2

Note: We use a Chi-square test to check for gender differences across treatment groups. We use a Kruskal-Wallis Rank Sum Test for the other covariates.

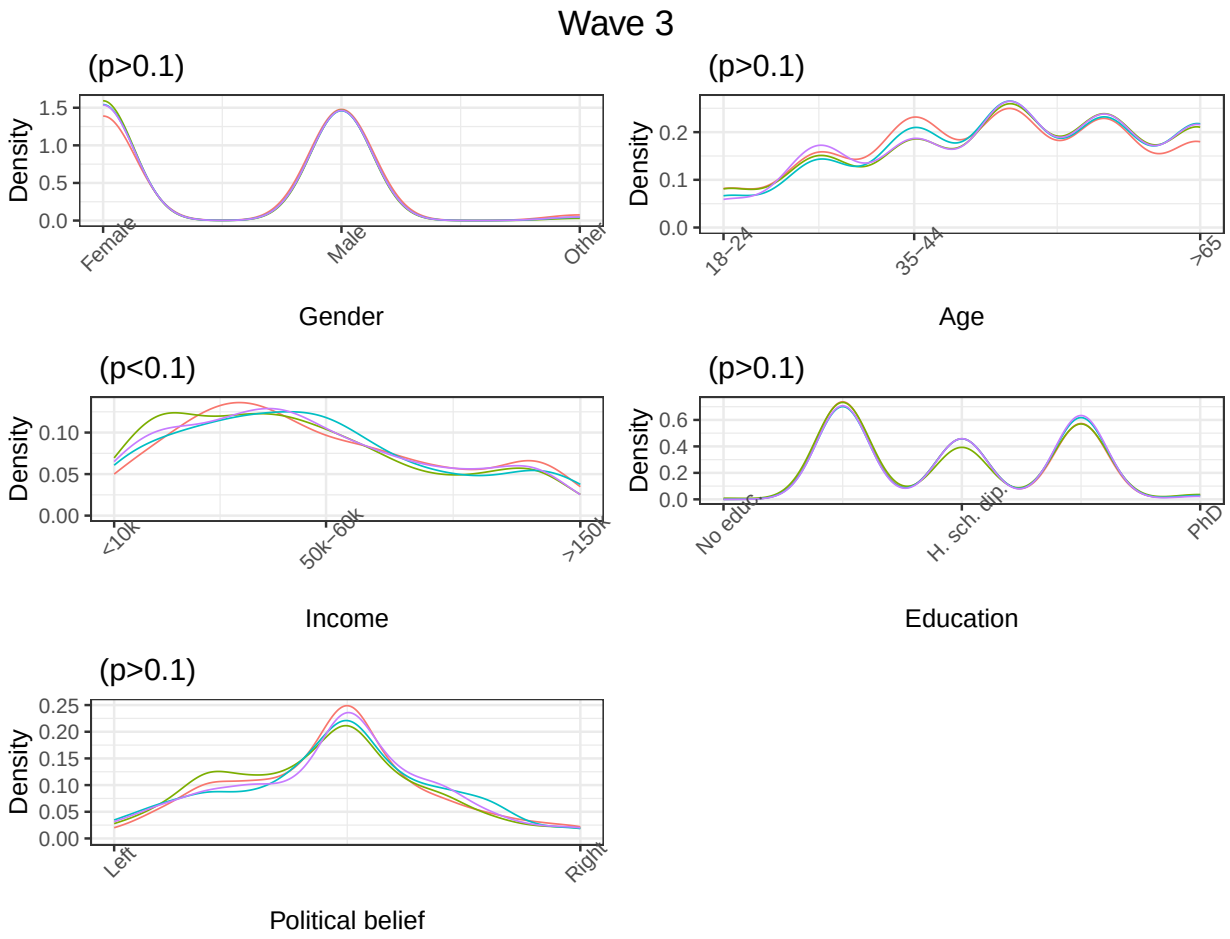


Figure 4.A.6: Distribution of Covariates Within Wave 3

Note: We use a Chi-square test to check for gender differences across treatment groups. We use a Kruskal-Wallis Rank Sum Test for the other covariates.

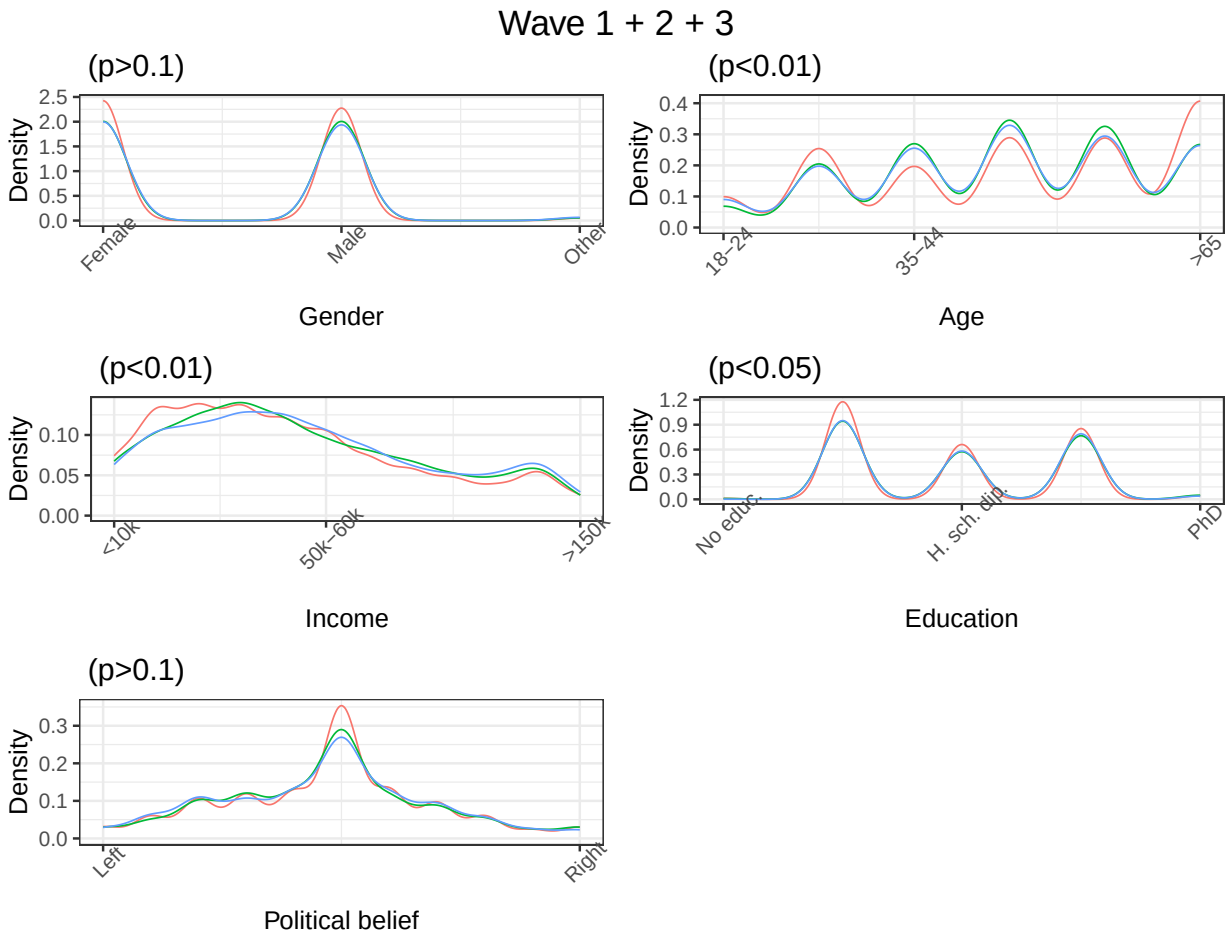


Figure 4.A.7: Distribution of Covariates Across Waves

Note: We use a Chi-square test to check for gender differences across waves. We use a Kruskal-Wallis Rank Sum Test for the other covariates.

4.A.B Exploratory Analyses

Table 4.A.1: Heterogeneity Analysis - Hypothesis 1 and 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.468***	0.477***	11.823***	12.260***	0.490***	0.478***	12.682***	12.323***
	(0.021)	(0.020)	(0.602)	(0.564)	(0.020)	(0.021)	(0.582)	(0.617)
Salience nudge	0.370***	0.369***	11.185***	11.151***	0.383***	0.383***	11.710***	11.711***
	(0.016)	(0.016)	(0.475)	(0.475)	(0.016)	(0.016)	(0.470)	(0.469)
Win-Win vs. Control	0.008	−0.002	0.565	−0.103				
	(0.026)	(0.024)	(0.766)	(0.706)				
Altruism	0.094***		3.257***					
	(0.023)		(0.669)					
Pro-environment		0.091***		2.927***				
		(0.022)		(0.654)				
Altruism x Win-Win vs. Control	−0.052		−1.683					
	(0.033)		(0.974)					
	q=0.113		q=0.081					
Pro-environment x Win-Win vs. Control		−0.042		−0.679				
		(0.032)		(0.950)				
		q=0.184		q=0.474				
Doom & Gloom vs. Control					−0.010	−0.012	−0.105	−0.281
					(0.024)	(0.026)	(0.705)	(0.774)
Guilt (NBE)					0.050**		1.494**	
					(0.022)		(0.660)	
Guilt (Repair)						0.063***		1.907***
						(0.023)		(0.678)
Guilt (NBE) x Doom & Gloom vs. Control					−0.025		−0.477	
					(0.032)		(0.948)	
					q=0.428		q=0.617	
Guilt (Repair) x Doom & Gloom vs. Control						−0.019		−0.143
						(0.033)		(0.975)
						q=0.563		q=0.887
Num Obs	2727	2727	2727	2727	2760	2760	2760	2760
R2	0.167	0.168	0.176	0.176	0.174	0.176	0.184	0.186
Data included	Wave 1 without placebo and doom-and-gloom groups				Wave 1 without placebo and win-win groups			
Outcome	PEB1 (binary)		PEB1 (continuous)		PEB1 (binary)		PEB1 (continuous)	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table displays the effect of respondents' attitudes regarding guilt, altruism and the environment on the ATEs of win-win and doom-and-gloom arguments on PEB1. We use OLS regressions. Columns one to four present the interaction between allocation to win-win arguments and scoring above the median for altruistic and pro-environmental attitudes. Columns five to eight present the interaction between allocation to doom-and-gloom arguments and scoring above the median for guilt-proneness attitudes. NBE stands for Negative-Behaviour Evaluation. It captures the extent to which one feels bad about one's behaviour. "Guilt (Repair)" captures respondents' tendency to act to temper guilt. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values of the interacted terms (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.A.2: Heterogeneity Analysis - Hypothesis 4 and 5

	(1)	(2)	(3)	(4)
Constant	0.202***	0.191***	0.264***	0.229***
	(0.032)	(0.031)	(0.033)	(0.032)
PEB1	−0.001	−0.001	−0.001	−0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Win-Win vs. Control	0.045*	0.010		
	(0.025)	(0.021)		
Altruism	0.148***			
	(0.024)			
Pro-environment		0.191***		
		(0.024)		
Altruism x Win-Win vs. Control	−0.071			
	(0.034)			
	q=0.037			
Pro-environment x Win-Win vs. Control		−0.017		
		(0.033)		
		q=0.598		
Doom & Gloom vs. Control			−0.026	−0.003
			(0.024)	(0.025)
Guilt (NBE)			0.052**	
			(0.024)	
Guilt (Repair)				0.101***
				(0.024)
Guilt (NBE) x Doom & Gloom vs. Control			0.029	
			(0.034)	
			q=0.382	
Guilt (Repair) x Doom & Gloom vs. Control				−0.012
				(0.034)
				q=0.720
Num Obs	2727	2727	2760	2760
Data included	Wave 1 without placebo and doom-and-gloom groups		Wave 1 without placebo and win-win groups	
Outcome	PEB2			

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table displays the effect of respondents' attitudes regarding guilt, altruism and the environment on the direct effects of win-win and doom-and-gloom arguments on PEB2. We use 2SLS regressions where PEB1 is instrumented by allocation to the salience nudge. Columns one to four present the interaction between allocation to win-win arguments and scoring above the median for altruistic and pro-environmental attitudes. Columns five to eight present the interaction between allocation to doom-and-gloom arguments and scoring above the median for guilt-proneness attitudes. NBE stands for Negative-Behaviour Evaluation. It captures the extent to which one feels bad about one's behaviour. "Guilt (Repair)" captures respondents' tendency to act to temper guilt. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values of the interacted terms (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.A.3: Correlations between PEB1 and PEB2

	(1)	(2)
Constant	0.241***	0.240***
	(0.010)	(0.010)
PEB1	0.032*	0.001**
	(0.013)	(0.000)
Num Obs	5492	5492
R2	0.001	0.001
Data included	Wave 1	
Outcome	PEB2	
PEB1	Binary	Continuous

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

Note: This table displays the correlation between the decision to do PEB2 and the decision to do PEB1 (first column) and the number of rounds done in PEB1 (second column). We use OLS regressions. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p).

Table 4.A.4: Hypothesis 1 and 2 with Performances as an Outcome

	(1)	(2)	(3)	(4)
Constant	0.198***	0.196***	0.182***	0.186***
	(0.007)	(0.007)	(0.007)	(0.006)
Salience nudge	0.154***	0.159***	0.166***	0.159***
	(0.008)	(0.008)	(0.007)	(0.005)
Win-Win vs.	−0.001			
Control	(0.008)			
	q=0.883			
Doom & Gloom vs.		−0.006		
Control		(0.008)		
		q=0.433		
Control vs.			0.010	
Placebo			(0.008)	
			q=0.182	
All vs.				0.008
Placebo				(0.006)
				q=1.000
Num Obs	2727	2760	2757	5492
R2	0.125	0.134	0.152	0.136
Outcome	Performances in PEB1			
Data included	Wave 1			

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the ATEs of win-win and doom-and-gloom arguments on respondents' performances in the real effort task. We used OLS regressions where respondents' performances are regressed on the dummy variables, capturing their allocation to treatment texts. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.A.5: Effect of Performances in PEB1 on PEB2 and Perception of Effort

	(1)	(2)	(3)
Constant	0.231***	2.323***	0.248***
	(0.008)	(0.046)	(0.019)
Difficulty	0.310***		
	(0.005)		
	q<0.01		
Performance		0.456***	0.013
		(0.101)	(0.043)
		q<0.01	q=0.550
Num Obs	5671	5671	5671
R2	0.366	0.001	0.001
F-stat of difficulty	3663.575		
Data included	Participants who did PEB1 in waves 1 and 2		
Outcome	Performance	Effort	PEB2

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the ATE of varying the difficulty of the real-effort task on respondents' performances (column one), the effect of performances — instrumented by the difficulty of PEB1 — on respondents' perception of having made an effort for the environment (column two), and the effect of performances — instrumented by the difficulty of PEB1 — on respondents' likelihood of signing the environment-related petition (column 3). We fitted an OLS regression for column one and 2SLS regressions for columns two and three. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.A.6: Time Spent Reading the Articles and Treatment Effects

	(1)	(2)
Constant	0.683***	18.665***
	(0.014)	(0.415)
Time reading	0.004	0.118
	(0.003)	(0.094)
All vs. Placebo	0.014	0.519
	(0.016)	(0.474)
Time reading ×	−0.005	−0.150
All vs. Placebo	(0.003)	(0.101)
	q<0.01	q<0.01
Num Obs	5492	5492
R2	0.0004	0.0004
Data included	Wave 1	
Outcome	PEB1 (binary)	PEB1 (continuous)

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of spending more time reading the texts on the ATEs of being allocated to the treatment or control groups on PEB1. We fitted OLS regressions to estimate the interaction between allocation in one of the treated or control groups and the time spent reading the articles. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values of the interacted terms (p). P-values of randomisation tests with 5,000 re-sampling are displayed last following Young (2019)'s procedure (q).

Table 4.A.7: Correlation between Perception of Effort and Information Treatments

	(1)
Constant	2.441***
	(0.035)
All vs. Placebo	0.230***
	(0.040)
Num Obs	3806
R2	0.009
Data included	Participants who did PEB1 in wave 1
Outcome	Effort

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table displays the effect of being shown information on PEB1 on the perception of having made an effort for the environment when participating in PEB1. Robust standard errors are in parentheses. We apply Benjamini and Hochberg (1995) correction to conventional p-values (p).

4.A.C Robustness Checks

Table 4.A.8: Robustness Checks - Hypothesis 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant		0.553***	0.197**	0.465***	15.195***	4.582	12.300***
		(0.051)	(0.096)	(0.032)	(1.516)	(2.837)	(0.943)
Salience nudge	0.340***	0.356***	0.359***	0.365***	10.692***	10.796***	11.003***
	(0.013)	(0.017)	(0.017)	(0.012)	(0.512)	(0.508)	(0.344)
Win-Win vs. Control	−0.025	−0.026	−0.024	−0.020	−0.648	−0.600	−0.354
	(0.016)	(0.017)	(0.017)	(0.016)	(0.512)	(0.509)	(0.477)
Num Obs	2727	2356	2356	5228	2356	2356	5228
R2		0.171	0.180	0.165	0.175	0.185	0.170
Model	Probit	OLS	OLS	OLS	OLS	OLS	OLS
Control for social-demographics		X	X	X	X	X	X
Control for attitudes			X			X	
Data included	Wave 1	Wave 1	Wave 1	Wave 1+2+3	Wave 1	Wave 1	Wave 1+2+3
Outcome	PEB1 (binary)				PEB1 (continuous)		

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the ATE of win-win arguments on participation in PEB1. When using the full sample, we add wave fixed effects.

Table 4.A.9: Robustness Checks - Hypothesis 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant		0.518***	0.278***	0.438***	14.036***	5.750**	11.348***
		(0.051)	(0.099)	(0.026)	(1.512)	(2.873)	(0.781)
Salience nudge	0.355***	0.376***	0.376***	0.371***	11.489***	11.487***	11.250***
	(0.012)	(0.017)	(0.017)	(0.009)	(0.505)	(0.503)	(0.282)
Doom & Gloom vs. Control	−0.025	−0.023	−0.024	−0.025***	−0.410	−0.466	−0.613**
	(0.016)	(0.017)	(0.017)	(0.009)	(0.511)	(0.509)	(0.282)
Num Obs	5492	2358	2358	7843	2358	2358	7843
R2		0.182	0.186	0.168	0.190	0.197	0.174
Model	Probit	OLS	OLS	OLS	OLS	OLS	OLS
Control for social-demographics		X	X	X	X	X	X
Control for attitudes			X			X	
Data included	Wave 1	Wave 1	Wave 1	Wave 1+2+3	Wave 1	Wave 1	Wave 1+2+3
Outcome	PEB1 (binary)				PEB1 (continuous)		

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the ATE of doom-and-gloom arguments on participation in PEB1. When using the full sample, we add wave fixed effects.

Table 4.A.10: Robustness Checks - Hypothesis 3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant		0.255***	−0.307***	0.168***		0.254***	−0.308***	0.167***
		(0.045)	(0.070)	(0.028)		(0.045)	(0.070)	(0.027)
PEB1	−0.010	−0.023	−0.017	−0.008	0.000	−0.001	−0.001	0.000
	(0.031)	(0.034)	(0.033)	(0.026)	(0.001)	(0.001)	(0.001)	(0.001)
All vs. Placebo	0.025*	0.025*	0.021	0.025*	0.025*	0.025*	0.021	0.025*
	(0.014)	(0.015)	(0.014)	(0.013)	(0.014)	(0.015)	(0.014)	(0.013)
Num Obs	5492	4704	4704	8112	5492	4704	4704	8112
Model	R&V (1998)	OLS	OLS	OLS	R&V (1998)	OLS	OLS	OLS
Control for social-demographics		X	X	X		X	X	X
Control for attitudes			X				X	
Data included	Wave 1	Wave 1	Wave 1	Wave 1+2+3	Wave 1	Wave 1	Wave 1	Wave 1+2+3
Outcome	PEB2							
PEB1	Binary				Continuous			

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the effect of participating in the real effort task — instrumented by allocation to the salience nudge — on respondents' likelihood of signing the environment-related petition, controlling for their allocation to one of the treatment texts. When using the full sample, we add wave fixed effects.

Table 4.A.11: Robustness Checks - Hypothesis 4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant		0.354***	−0.170*	0.206***		0.352***	−0.172*	0.204***
		(0.067)	(0.097)	(0.041)		(0.066)	(0.097)	(0.039)
PEB1	−0.041	−0.062	−0.043	−0.022	−0.001	−0.002	−0.001	−0.001
	(0.046)	(0.052)	(0.050)	(0.038)	(0.002)	(0.002)	(0.002)	(0.001)
Win-Win vs. Control	0.003	0.004	0.004	0.002	0.003	0.004	0.004	0.002
	(0.017)	(0.019)	(0.018)	(0.017)	(0.017)	(0.019)	(0.018)	(0.017)
Num Obs	2727	2356	2356	4022	2727	2356	2356	4022
Model	R&V (1998)	OLS	OLS	OLS	R&V (1998)	OLS	OLS	OLS
Control for social-demographics		X	X	X		X	X	X
Control for attitudes			X				X	
Data included	Wave 1	Wave 1	Wave 1	Wave 1+2+3	Wave 1	Wave 1	Wave 1	Wave 1+2+3
Outcome	PEB2							
PEB1	Binary				Continuous			

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the direct effect of win-win arguments on respondents' likelihood of signing the environment-related petition, controlling for participation in the real effort task — instrumented by allocation to the salience nudge. When using the full sample, we add wave fixed effects.

Table 4.A.12: Robustness Checks - Hypothesis 5

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant		0.290***	−0.403***	0.186***		0.287***	−0.408***	0.184***
		(0.065)	(0.101)	(0.034)		(0.063)	(0.100)	(0.032)
PEB1	−0.037	−0.051	−0.052	−0.021	−0.001	−0.002	−0.002	−0.001
	(0.044)	(0.049)	(0.047)	(0.033)	(0.001)	(0.002)	(0.002)	(0.001)
Doom & Gloom vs. Control	−0.011	−0.007	−0.011	−0.001	−0.010	−0.006	−0.011	−0.001
	(0.017)	(0.019)	(0.018)	(0.012)	(0.017)	(0.018)	(0.018)	(0.012)
Num Obs	2760	2358	2358	5380	2760	2358	2358	5380
Model	R&V (1998)	OLS	OLS	OLS	R&V (1998)	OLS	OLS	OLS
Control for social-demographics		X	X	X		X	X	X
Control for attitudes			X				X	
Data included	Wave 1	Wave 1	Wave 1	Wave 1+2+3	Wave 1	Wave 1	Wave 1	Wave 1+2+3
Outcome	PEB2							
PEB1	Binary				Continuous			

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the direct effect of doom-and-gloom arguments on respondents' likelihood of signing the environment-related petition, controlling for participation in the real-effort task — instrumented by allocation to the salience nudge. When using the full sample, we add wave fixed effects.

Table 4.A.13: Robustness Checks - Hypothesis 6

	(1)	(2)	(3)	(4)	(5)	(6)
Constant		0.244***	−0.310***		0.243***	−0.311***
		(0.040)	(0.062)		(0.039)	(0.061)
PEB1	−0.012	−0.023	−0.019	0.000	−0.001	−0.001
	(0.032)	(0.034)	(0.033)	(0.001)	(0.001)	(0.001)
All vs. Placebo	0.025*	0.025*	0.020	0.025*	0.025*	0.021
	(0.014)	(0.015)	(0.014)	(0.014)	(0.015)	(0.014)
Difficulty	−0.002	0.015	0.013	−0.001	0.016	0.013
	(0.043)	(0.048)	(0.047)	(0.041)	(0.046)	(0.045)
PEB1 x Difficulty	0.005	−0.018	−0.021	0.000	−0.001	−0.001
	(0.059)	(0.065)	(0.064)	(0.002)	(0.002)	(0.002)
Num Obs	8112	6874	6855	8112	6874	6855
Model	R&V (1998)	OLS	OLS	R&V (1998)	OLS	OLS
Control for social-demographics		X	X		X	X
Control for attitudes			X			X
Data included			Wave 1+2			
Outcome			PEB2			
PEB1		Binary			Continuous	

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the effect of varying the difficulty of the real-effort task on the effect of doing the real-effort task — instrumented by allocation to the salience nudge — on the likelihood of signing the environment-related petition.

Table 4.A.14: Robustness Checks - Hypothesis 7 and 8

	(1)	(2)	(3)	(4)	(5)	(6)
Constant		0.317***	−0.317***		0.314***	−0.321***
		(0.042)	(0.065)		(0.041)	(0.064)
PEB1	−0.025	−0.045	−0.045	0.001	−0.001	−0.001
	(0.034)	(0.037)	(0.036)	(0.001)	(0.001)	(0.001)
Doom & Gloom vs. Control	−0.001	−0.003	−0.002	0.000	−0.003	−0.002
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Petition	0.016	−0.016	−0.025	0.057	−0.011	−0.019
	(0.045)	(0.053)	(0.052)	(0.036)	(0.051)	(0.050)
PEB1 x Health Petition	0.101*	0.138**	0.156**	0.002	0.005**	0.005**
	(0.058)	(0.070)	(0.069)	(0.002)	(0.002)	(0.002)
Doom & Gloom vs. Control x	0.038*	0.054**	0.057**	0.036*	0.054**	0.057**
Health Petition	(0.022)	(0.025)	(0.025)	(0.022)	(0.025)	(0.025)
Num Obs	7936	6628	6609	7936	6628	6609
Model	R&V (1998)	OLS	OLS	R&V (1998)	OLS	OLS
Control for social-demographics		X	X		X	X
Control for attitudes			X			X
Data included			Wave 1+2+3			
Outcome			PEB2			
PEB1		Binary			Continuous	

* p < 0.1, ** p < 0.05, *** p < 0.01

Note: This table presents different specifications estimating the effect of varying the cause supported by the petition on the effect of doing the real-effort task — instrumented by allocation to the salience nudge — and the effect of doom-and-gloom arguments on the likelihood of signing the petition. For all specifications, we control for the difficulty of PEB1.

Chapter 5

Conclusion

I believe an ideal experiment is a triptych. First, there should be a theory. The theory should be positive for the problems I am interested in, i.e., it should describe people's behaviour to explain what is happening in the experiment. Positive theories define a set of necessary and sufficient conditions $\mathcal{C}$ under which one can observe a given phenomenon P (e.g., moral licensing happens if and only if decisions are extrinsically motivated). When these conditions are not met (when we are in $\overline{\mathcal{C}}$), then the phenomenon cannot happen ($\overline{P}$ happens). We can disprove or corroborate a theory if we measure all the conditions.

Unfortunately, delineating these conditions is impossible for the questions I study (e.g., I cannot measure "extrinsic motivations") and, more generally, for most studies looking at moral behaviours. We are not in people's minds. We can only assume the existence of some latent conditions that induce the phenomena we measure.

I am conscious of this limit when relying on economic models. My approach when developing the theory in Chapter 2 was to obtain an "isomorphism" between the conditions $\mathcal{C}$ (i.e., intrinsic and extrinsic motivations) and the phenomena $\mathcal{P}$ (i.e., negative and positive spillover effects). Then, I would proceed by backward induction. From the phenomena I observed, I would infer the condition at play, assuming that they are right. As such, the model is just an imperfect tool to make sense of what I measured.

Second, once we have the theory that maps the unobservable conditions to the observable phenomena, we need to consider how to measure the observable phenomena in an experiment. Observing pro-environmental decisions is hard. Three criteria should be kept in mind.

1. *Allow people to make actual pro-environmental actions*. In other words, people's choices should

have consequences, or at least, the experiment should be set to make people behave *as if* their decisions had consequences. Otherwise, external validity is reduced.

2. *Re-create the context for all the motivations underpinning pro-environmental actions to be at play*. Context matters. My decisions may have consequences, but if I am in a context that I would never encounter in my daily life (i.e. if there is no parallel I can draw between the current contexts and other situations in which I have found myself), then again, external validity is reduced.

3. *Design the experiment to estimate the effects of interests without biases*. This relates to internal validity. For instance, I relied on an instrumental variable embedded in my experimental designs to disentangle direct from indirect spillover effects in Chapters 3 and 4. This criterion must be fulfilled to derive any meaningful results in the first place.

Field experiments are the gold standard to satisfy criteria (1) and (2). Yet, they do not always allow for criterion (3) to be fulfilled. For constraints inherent to doctoral studies (e.g., time and money), for the sake of maximising power, but also because deriving causal estimates of spillover effects is not trivial, I chose to use online experiments.

In hindsight, the two experiments presented in this dissertation show a trade-off between criteria (1) and (2). In Chapter 3, food choices were hypothetical, but people could easily apprehend them. It was a familiar decision, and people's choices were close to what one would do in an online delivery app. In other words, criterion (1) was not perfectly satisfied because of hypothetical choices, but people could think of contexts where they would make such a decision (criterion (2) was likely satisfied). In Chapter 4, choices were consequential, but the task was new to respondents and maybe too artificial to reproduce the rich processes underpinning a selfless act. Here, criterion (1)

was fulfilled (choices were consequential), but maybe not criterion (2). This is something I learnt and a piece of experience I will carry with me in other projects.

The last part of this triptych is the process through which data is collected, analysed and presented. This process ought to be as transparent as possible. Here, the gold standard, in my opinion, is the use of registered reports. Indeed, to increase transparency, minimise publication biases and increase the reproducibility of experimental results, the review process should be based on the question and the method, not the results. This is what I am trying to do with the experiment presented in Chapter 4 and with other projects I am currently working on.

However, registered reports are not yet the norm in economics. When they cannot be used, one ought to be transparent about the thought process leading to the outcome presented to the research community. This entails using pre-analysis plans, conducting power analyses, and correcting for multiple hypothesis testing. In this regard, my research is not perfect, and what I planned on analysing when designing the experiments did not always turn out to be the most interesting part. I seek to be transparent about this by clearly stating what is exploratory. I also believe in using post-analysis plans to document deviations from the initial plan that would not necessarily have their place in a research paper.

Correcting for multiple hypothesis testing entails a trade-off between minimising false positives whilst not creating too many false negatives. This trade-off is particularly important when estimating effects with no priors from the literature, which was my case when disentangling direct from indirect spillover effects. In this situation, I was interested in maximising the number of discoveries while controlling the number of false positives. This is why I decided to control for the false discovery rate (i.e., keeping the expected number of false positives under 5%) rather than

the family-wise error rate (i.e., keeping the probability of having at least one false positive under 5%). I completed this approach with robustness checks (i.e., using different statistical models and placebo tests based on re-randomising treatment assignments).

Now, what about the next steps? In this dissertation, I have been abstracting from others' influence on individuals' decisions (albeit scratching the surface of this topic in Chapter 3 with social norm messaging). I applied Ockham's razor on purpose. The task was already complex enough. Yet, understanding how others influence the way someone perceives a policy (and so the psychological mechanism of the policy) and how a policy applied to someone can influence others is key to answering the big picture question: how to make pro-environmental behaviours the social norm? This will be the next step in my research agenda.

Bibliography

Abrahamse, W. and Steg, L. (2013). Social influence approaches to encourage resource conservation: A meta-analysis. *Global environmental change*, 23(6):1773–1785.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The quarterly journal of economics*, 115(3):715–753.

Alacevich, C., Bonev, P., and Söderberg, M. (2021). Pro-environmental interventions and behavioral spillovers: Evidence from organic waste sorting in sweden. *Journal of Environmental Economics and Management*, 108:102470.

Allcott, H. (2011). Social norms and energy conservation. *Journal of public Economics*, 95(9-10):1082–1095.

Alsan, M. and Eichmeyer, S. (2024). Experimental evidence on the effectiveness of nonexperts for improving vaccine demand. *American Economic Journal: Economic Policy*, 16(1):394–414.

Alt, M., Bruns, H., and DellaValle, N. (2023). The more the better?-synergies of prosocial interventions and effects on behavioral spillovers. *Synergies of prosocial interventions and effects on behavioral spillovers (June 27, 2023)*.

Alt, M. and Gallier, C. (2022). Incentives and intertemporal behavioral spillovers: A two-period experiment on charitable giving. *Journal of Economic Behavior & Organization*, 200:959–972.

Andersson, M. and von Borgstede, C. (2010). Differentiation of determinants of low-cost and high-cost recycling. *Journal of Environmental Psychology*, 30(4):402–408.

Andor, M. A. and Fels, K. M. (2018). Behavioral economics and energy conservation–a systematic review of non-price interventions and their causal effects. *Ecological economics*, 148:178–210.

Andor, M. A., Frondel, M., and Vance, C. (2017). Mitigating hypothetical bias: Evidence on the effects of correctives from a large field study. *Environmental and Resource Economics*, 68(3):777–796.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Baca-Motes, K., Brown, A., Gneezy, A., Keenan, E. A., and Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, 39(5):1070–1084.

Banerjee, A., Alsan, M., BREZA, E., Chowdhury, A., Duflo, E., Olken, B., Chandrasekhar, A., and Goldsmith-Pinkham, P. (2022). Can a trusted messenger change behavior when information is plentiful? evidence from the first months of the covid-19 pandemic in west bengal. *Technical report*.

Banerjee, S. and Picard, J. (2023). Thinking through norms can make them more effective. experimental evidence on reflective climate policies in the uk. *Journal of Behavioral and Experimental Economics*, page 102024.

Bech-Larsen, T. and Kazbare, L. (2014). Spillover of diet changes on intentions to approach healthy food and avoid unhealthy food. *Health Education*. Publisher: Emerald Group Publishing Limited.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Berger, S. and Wyss, A. M. (2021). Measuring pro-environmental behavior using the carbon emission task. *Journal of environmental psychology*, 75:101613.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., and Sonnet, L. (2022). *estimatr: Fast Estimators for Design-Based Inference*. https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr.

Blondin, S., Attwood, S., Vennard, D., and Mayneris, V. (2022). Environmental messages promote plant-based food choices: An online restaurant menu study. *World Resources Institute*.

Bonev, P. (2023). Behavioral spillovers. PsyArXiv.

Bonnet, C., Bouamra-Mechemache, Z., Réquillart, V., and Treich, N. (2020). Regulating meat consumption to improve health, the environment and animal welfare. *Food Policy*, 97:101847.

Bouman, T., Steg, L., and Kiers, H. A. (2018). Measuring values in environmental research: a test of an environmental portrait value questionnaire. *Frontiers in psychology*, 9:564. Publisher: Frontiers Media SA.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

Bovens, L. (2009). The ethics of nudge. In *Preference change*, pages 207–219. Springer.

Bratt, C. (1999). The impact of norms and assumed consequences on recycling behavior. *Environment and behavior*, 31(5):630–656.

Bryan, C. J., Tipton, E., and Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989.

Cameron, J. and Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational research*, 64(3):363–423.

Carlsson, F., Jaime, M., and Villegas, C. (2021). Behavioral spillover effects from a social information campaign. *Journal of Environmental Economics and Management*, 109:102325. Publisher: Elsevier.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018a). Putting your money where your mouth is: an experimental test of pro-environmental spillover from reducing meat consumption to monetary donations. *Environment and Behavior*, 50(7):723–748.

Carrico, A. R., Raimi, K. T., Truelove, H. B., and Eby, B. (2018b). Putting Your Money Where Your Mouth Is: An Experimental Test of Pro-Environmental Spillover From Reducing Meat Consumption to Monetary Donations. *Environment and Behavior*, 50(7):723–748. Publisher: SAGE Publications Inc.

Carrico, A. R. and Riemer, M. (2011). Motivating energy conservation in the workplace: An evaluation of the use of group-level feedback and peer education. *Journal of environmental psychology*, 31(1):1–13.

Champ, P. A., Moore, R., and Bishop, R. C. (2009). A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review*, 38(2):166–180.

Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.

Chell, K., Davison, T. E., Masser, B., and Jensen, K. (2018). A systematic review of incentives in blood donation. *Transfusion*, 58(1):242–254.

Chess, C. and Johnson, B. B. (2007). *Information is not enough*, pages 223–233. Cambridge University Press.

Cialdini, R. B. (1994). *Interpersonal influence*, pages 195—–217. Allyn & Bacon.

Cialdini, R. B. and Jacobson, R. P. (2021). Influences of social norms on climate change-related behaviors. *Current Opinion in Behavioral Sciences*, 42:1–8.

Cialdini, R. B., Trost, M. R., and Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of personality and social psychology*, 69(2):318. Publisher: American Psychological Association.

Clot, S., Della Giusta, M., and Jewell, S. (2022). Once good, always good? testing nudge's spillovers on pro environmental behavior. *Environment and Behavior*, 54(3):655–669.

Clot, S., Grolleau, G., and Ibanez, L. (2014). Smug alert! Exploring self-licensing behavior in a cheating game. *Economics Letters*, 123(2):191–194. Publisher: Elsevier.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the GASP scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947. Publisher: American Psychological Association.

Comin, D. A. and Rode, J. (2023). Do green users become green voters? Technical report, National Bureau of Economic Research.

Corazzini, L., Cotton, C., and Valbonesi, P. (2015). Donor coordination in project funding: Evidence from a threshold public goods experiment. *Journal of Public Economics*, 128:16–29. Publisher: Elsevier.

Cornelissen, G., Bashshur, M. R., Rode, J., and Le Menestrel, M. (2013). Rules or consequences? the role of ethical mind-sets in moral dynamics. *Psychological Science*, 24(4):482–488.

Costa, D. L. and Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3):680–702.

Davidson, D. J. and Kecinski, M. (2022). Emotional pathways to climate change responses. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2):e751.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627.

Deck, C. and Murphy, J. J. (2019). Donors change both their level and pattern of giving in response to contests among charities. *European Economic Review*, 112:91–106. Publisher: Elsevier.

Dolan, P. and Galizzi, M. M. (2014). Because i'm worth it: a lab-field experiment on the spillover effects of incentives in health. Centre for Economic Performance, London School of Economics and Political Science.

Dolan, P. and Galizzi, M. M. (2015a). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16.

Dolan, P. and Galizzi, M. M. (2015b). Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16. Publisher: Elsevier.

Dolan, P., Galizzi, M. M., and Navarro-Martinez, D. (2015). Paying people to eat or not to eat? Carryover effects of monetary incentives on eating behaviour. *Social Science & Medicine*, 133:153–158. Publisher: Elsevier.

Douenne, T. and Fabre, A. (2022). Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1):81–110.

Effron, D. A., Cameron, J. S., and Monin, B. (2009). Endorsing obama licenses favoring whites. *Journal of experimental social psychology*, 45(3):590–593.

Ek, C. and Miliute-Plepiene, J. (2018). Behavioral spillovers from food-waste collection in Swedish municipalities. *Journal of Environmental Economics and Management*, 89:168–186. Publisher: Elsevier.

Elliot, A. J. and Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of personality and social psychology*, 67(3):382.

Evans, L., Maio, G. R., Corner, A., Hodgetts, C. J., Ahmed, S., and Hahn, U. (2013). Self-interest and pro-environmental behaviour. *Nature Climate Change*, 3(2):122–125.

Evans, W. N. and Schwab, R. M. (1995). Finishing high school and starting college: Do catholic schools make a difference? *The Quarterly Journal of Economics*, 110(4):941–974.

Farrow, K., Grolleau, G., and Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, 140:1–13.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, 101(3):318–22.

Festinger, L. (1962a). Cognitive dissonance. *Scientific American*, 207(4):93–106.

Festinger, L. (1962b). Cognitive dissonance. *Scientific American*, 207(4):93–106. Publisher: JSTOR.

Filiz-Ozbay, E. and Uler, N. (2019). Demand for giving to multiple charities: An experimental study. *Journal of the European Economic Association*, 17(3):725–753. Publisher: Oxford University Press.

Fiske, A. P. and Tetlock, P. E. (1997). Taboo trade-offs: reactions to transactions that transgress the spheres of justice. *Political psychology*, 18(2):255–297.

Fornara, F., Carrus, G., Passafaro, P., and Bonnes, M. (2011). Distinguishing the sources of normative influence on proenvironmental behaviors: The role of local norms in household waste recycling. *Group Processes & Intergroup Relations*, 14(5):623–635.

Freedman, J. L. and Fraser, S. C. (1966). Compliance without pressure: the foot-in-the-door tech-

nique. *Journal of personality and social psychology*, 4(2):195. Publisher: American Psychological Association.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gärtner, M. (2018). The prosociality of intuitive decisions depends on the status quo. *Journal of Behavioral and Experimental Economics*, 74:127–138.

Garvey, A. and Bolton, L. (2017). The licensing effect revisited: How virtuous behavior heightens the pleasure derived from subsequent hedonic consumption. *Journal of Marketing Behavior, Forthcoming*.

Geiger, S. J., Brick, C., Nalborczyk, L., Bosshard, A., and Jostmann, N. B. (2021). More green than gray? toward a sustainable overview of environmental spillover effects: A bayesian meta-analysis. *Journal of Environmental Psychology*, 78:101694.

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., and Norton, M. I. (2012a). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, 58(1):179–187.

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., and Norton, M. I. (2012b). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, 58(1):179–187. Publisher: INFORMS.

Gneezy, U., Imas, A., and Madarász, K. (2014). Conscience accounting: Emotion dynamics and social behavior. *Management Science*, 60(11):2645–2658.

Goetz, A., Mayr, H., and Schubert, R. (2022). One thing leads to another: Evidence on the scope and persistence of behavioral spillovers. *Available at SSRN 4479949*.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482.

Gonzalez, N. I. V., Kee, J. Y., Palma, M. A., and Pruitt, J. R. (2023). The relationship between monetary incentives, social status, and physical activity. *Journal of Behavioral and Experimental Economics*, page 102155.

Green, R., Milner, J., Dangour, A. D., Haines, A., Chalabi, Z., Markandya, A., Spadaro, J., and Wilkinson, P. (2015). The potential to reduce greenhouse gas emissions in the uk through healthy and realistic dietary change. *Climatic Change*, 129(1):253–265.

Grubb, M. (2014). *Planetary economics: energy, climate change and the three domains of sustainable development*. Routledge.

Handgraaf, M. J., De Jeude, M. A. V. L., and Appelt, K. C. (2013). Public praise vs. private pay: Effects of rewards on energy conservation in the workplace. *Ecological Economics*, 86:86–92.

Hansen, P. G. and Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3–28.

Herweg, F. and Schmidt, K. M. (2022). How to regulate carbon emissions with climate-conscious consumers. *The Economic Journal*, 132(648):2992–3019.

Homar, A. R. and Cvelbar, L. K. (2021). The effects of framing on environmental decisions: A systematic literature review. *Ecological Economics*, 183:106950.

Jessoe, K., Lade, G. E., Loge, F., and Spang, E. (2021a). Spillovers from behavioral interventions: Experimental evidence from water and energy use. *Journal of the Association of Environmental and Resource Economists*, 8(2):315–346.

Jessoe, K., Lade, G. E., Loge, F., and Spang, E. (2021b). Spillovers from behavioral interventions: Experimental evidence from water and energy use. *Journal of the Association of Environmental and Resource Economists*, 8(2):315–346. Publisher: The University of Chicago Press Chicago, IL.

Jordan, J., Leliveld, M. C., and Tenbrunsel, A. E. (2015). The moral self-image scale: Measuring and understanding the malleability of the moral self. *Frontiers in Psychology*, 6:1878.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206.

Krieg, J. and Samek, A. (2017). When charities compete: A laboratory experiment with simultaneous public goods. *Journal of behavioral and experimental economics*, 66:40–57. Publisher: Elsevier.

Kristofferson, K., White, K., and Peloza, J. (2014). The nature of slacktivism: How the social observability of an initial act of token support affects subsequent prosocial action. *Journal of Consumer Research*, 40(6):1149–1166.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Lacasse, K. (2016). Don't be satisfied, identify! strengthening positive spillover by connecting pro-

environmental behaviors to an "environmentalist" label. *Journal of Environmental Psychology*, 48:149–158.

Landry, C. E., Lange, A., List, J. A., Price, M. K., and Rupp, N. G. (2010). Is a donor in hand better than two in the bush? evidence from a natural field experiment. *American Economic Review*, 100(3):958–983.

Lange, F. and Dewitte, S. (2022). The work for environmental protection task: A consequential web-based procedure for studying pro-environmental behavior. *Behavior research methods*, 54(1):133–145.

Lange, F., Steinke, A., and Dewitte, S. (2018). The pro-environmental behavior task: A laboratory measure of actual pro-environmental behavior. *Journal of Environmental Psychology*, 56:46–54.

Lanzini, P. and Thøgersen, J. (2014). Behavioural spillover in the environmental domain: An intervention study. *Journal of Environmental Psychology*, 40:381–390.

Lapinski, M. K., Rimal, R. N., DeVries, R., and Lee, E. L. (2007). The role of group orientation and descriptive norms on water conservation attitudes and behaviors. *Health communication*, 22(2):133–142.

Lauren, N., Fielding, K. S., Smith, L., and Louis, W. R. (2016). You did, so you can and you will: self-efficacy as a mediator of spillover from easy to more difficult pro-environmental behaviour. *Journal of Environmental Psychology*, 48:191–199.

Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. (2022). Valid t-ratio inference for iv. *American Economic Review*, 112(10):3260–3290.

Lee, S. W. and Schwarz, N. (2010). Dirty hands and dirty mouths: Embodiment of the moral-purity

metaphor is specific to the motor modality involved in moral transgression. *Psychological science*, 21(10):1423–1425.

List, J. A. and Momeni, F. (2020). Leveraging upfront payments to curb employee misbehavior: Evidence from a natural field experiment. *European Economic Review*, 130:103601. Publisher: Elsevier.

Liu, Y., Kua, H., and Lu, Y. (2021). Spillover effects from energy conservation goal-setting: A field intervention study. *Resources, Conservation and Recycling*, 170:105570.

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31):e2200300119.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019a). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019b). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315. Publisher: Nature Publishing Group.

Marbach, M. and Hangartner, D. (2020). Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis*, 28(3):435–444.

Margetts, E. A. and Kashima, Y. (2017a). Spillover between pro-environmental behaviours: The role of resources and perceived similarity. *Journal of Environmental Psychology*, 49:30–42.

Margetts, E. A. and Kashima, Y. (2017b). Spillover between pro-environmental behaviours: The

role of resources and perceived similarity. *Journal of Environmental Psychology*, 49:30–42. Publisher: Elsevier.

Mazar, N. and Zhong, C.-B. (2010a). Do green products make us better people? *Psychological science*, 21(4):494–498.

Mazar, N. and Zhong, C.-B. (2010b). Do green products make us better people? *Psychological science*, 21(4):494–498. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Meer, J. (2017). Does fundraising create new giving? *Journal of Public Economics*, 145:82–93. Publisher: Elsevier.

Melnyk, V., van Herpen, E., Trijp, H., et al. (2010). The influence of social norms in consumer decision making: A meta-analysis. *ACR North American Advances*.

Mohammed, E. Y. (2012). Contingent valuation responses and hypothetical bias: mitigation effects of certainty question, cheap talk, and pledging. *Environmental Economics*, 3:62–71.

Monin, B. and Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology*, 81(1):33.

Mullen, E. and Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual review of psychology*, 67(1):363–385.

Nabi, R. L., Gustafson, A., and Jensen, R. (2018). Framing climate change: Exploring the role of emotion in generating advocacy behavior. *Science Communication*, 40(4):442–468.

Nigbur, D., Lyons, E., and Uzzell, D. (2010). Attitudes, norms, identity and environmental beha-

viour: Using an expanded theory of planned behaviour to predict participation in a kerbside recycling programme. *British journal of social psychology*, 49(2):259–284.

Nilsson, A., Bergquist, M., and Schultz, W. P. (2017). Spillover effects in environmental behaviors, across time and context: a review and research agenda. *Environmental Education Research*, 23(4):573–589. Publisher: Taylor & Francis.

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., and Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature communications*, 10(1):1–13.

Noblet, C. L. and McCoy, S. K. (2018). Does one good turn deserve another? Evidence of domain-specific licensing in energy behavior. *Environment and Behavior*, 50(8):839–863. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and social psychology bulletin*, 34(7):913–923.

Ockenfels, A., Werner, P., and Edenhofer, O. (2020). Pricing externalities and moral behaviour. *Nature Sustainability*, 3(10):872–877.

Oliver, A. (2013). From nudging to budging: using behavioural economics to inform public sector policy. *Journal of Social Policy*, 42(4):685–700.

Ortmann, A., Ryvkin, D., Wilkening, T., and Zhang, J. (2023). Defaults and cognitive effort. *Journal of Economic Behavior & Organization*, 212:1–19.

Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human communication research*, 39(1):47–73.

Ready, R. C., Champ, P. A., and Lawton, J. L. (2010). Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Economics*, 86(2):363–381.

Reese, G., Loew, K., and Steffgen, G. (2014). A towel less: Social norms enhance pro-environmental behavior in hotels. *The Journal of Social Psychology*, 154(2):97–100.

Rhodes, N., Shulman, H. C., and McClaran, N. (2020). Changing norms: A meta-analytic integration of research on social norms appeals. *Human Communication Research*, 46(2-3):161–191.

Riahi, K., Schaeffer, R., Arango, J., Calvin, K., Guivarch, C., Hasegawa, T., Jiang, K., Kriegler, E., Matthews, R., Peters, G., et al. (2022). Mitigation pathways compatible with long-term goals. *IPCC, 2022: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Richter, I., Thøgersen, J., and Klöckner, C. A. (2018). A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability*, 10(8):2848.

Rivers, D. and Vuong, Q. H. (1988a). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366.

Rivers, D. and Vuong, Q. H. (1988b). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366. Publisher: Elsevier.

Rosendahl, K. E. (2019). Eu ets and the waterbed effect. *Nature Climate Change*, 9(10):734–735.

Sachdeva, S., Iliev, R., and Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science*, 20(4):523–528.

Salmivaara, L. and Lankoski, L. (2019). Promoting sustainable consumer behaviour through the ac-

tivation of injunctive social norms: A field experiment in 19 workplace restaurants. *Organization & Environment*, page 1086026619831651.

Sapolsky, R. M. (2018). *Behave: The biology of humans at our best and worst*. Penguin.

Scarborough, P., Appleby, P. N., Mizdrak, A., Briggs, A. D., Travis, R. C., Bradbury, K. E., and Key, T. J. (2014). Dietary greenhouse gas emissions of meat-eaters, fish-eaters, vegetarians and vegans in the uk. *Climatic change*, 125(2):179–192.

Schultz, W. P., Khazian, A. M., and Zaleski, A. C. (2008). Using normative social influence to promote conservation among hotel guests. *Social influence*, 3(1):4–23.

Schusser, S. and Bostedt, G. (2019). Green Behavioural (In) consistencies: Are Pro-environmental Behaviours in Different Domains Substitutes or Complements? *Environmental Economics*, 10(1):23–47.

Shipley, N. J. and van Riper, C. J. (2021). Pride and guilt predict pro-environmental behavior: A meta-analysis of correlational and experimental evidence. *Journal of Environmental Psychology*, page 101753. Publisher: Elsevier.

Shukla, Skea, Slade, Khourdajie, A., van Diemen, McCollum, Pathak, Some, Vyas, Fradera, Belkacemi, Hasija, Lisboa, Luz, and Malley (2022). Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change. *Cambridge University Press*.

Sintov, N., Geislar, S., and White, L. V. (2019). Cognitive accessibility as a new factor in proenvironmental spillover: results from a field study of household food waste management. *Environment and Behavior*, 51(1):50–80. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Skurka, C., Niederdeppe, J., Romero-Canyas, R., and Acup, D. (2018). Pathways of influence in emotional appeals: Benefits and tradeoffs of using fear or humor to promote climate change-related intentions and risk perceptions. *Journal of Communication*, 68(1):169–193.

Smith, A. (1853). *The theory of moral sentiments*. HG Bohn.

Sparkman, G. and Walton, G. M. (2017). Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychological science*, 28(11):1663–1674.

Sparkman, G., Weitz, E., Robinson, T. N., Malhotra, N., and Walton, G. M. (2020). Developing a scalable dynamic norm menu-based intervention to reduce meat consumption. *Sustainability*, 12(6):2453.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stea, S. and Pickering, G. J. (2019). Optimizing messaging to reduce red meat consumption. *Environmental Communication*, 13(5):633–648.

Steinhorst, J. and Matthies, E. (2016). Monetary or environmental appeals for saving electricity?–potentials for spillover on low carbon policy acceptability. *Energy Policy*, 93:335–344.

Stern, P. C. (2020). A reexamination on how behavioral interventions can promote household action to limit climate change. *Nature communications*, 11(1):1–3.

Stewart, C., Piernas, C., Cook, B., and Jebb, S. A. (2021). Trends in uk meat consumption: analysis of data from years 1–11 (2008–09 to 2018–19) of the national diet and nutrition survey rolling programme. *The Lancet Planetary Health*, 5(10):e699–e708.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Tajfel, H., Turner, J. C., Austin, W. G., and Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16.

Testa, F., Russo, M. V., Cornwell, T. B., McDonald, A., and Reich, B. (2018). Social sustainability as buying local: effects of soft policy, meso-level actors, and social influences on purchase intentions. *Journal of Public Policy & Marketing*, 37(1):152–166.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Thøgersen, J. (1999). Spillover processes in the development of a sustainable consumption pattern. *Journal of economic psychology*, 20(1):53–81.

Thøgersen, J. (2004). A cognitive dissonance interpretation of consistencies and inconsistencies in environmentally responsible behavior. *Journal of environmental Psychology*, 24(1):93–103. Publisher: Elsevier.

Thøgersen, J. and Ölander, F. (2003). Spillover of environment-friendly consumer behaviour. *Journal of environmental psychology*, 23(3):225–236. Publisher: Elsevier.

Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013a). For better or for worse? empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy*, 57:160–171.

Tiefenbeck, V., Staake, T., Roth, K., and Sachs, O. (2013b). For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy*, 57:160–171. Publisher: Elsevier.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014a). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138.

Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T., and Vandenbergh, M. P. (2014b). Positive and negative spillover of pro-environmental behavior: An integrative review and theoretical framework. *Global Environmental Change*, 29:127–138. Publisher: Elsevier.

Truelove, H. B., Yeung, K. L., Carrico, A. R., Gillis, A. J., and Raimi, K. T. (2016). From plastic bottle recycling to policy support: An experimental test of pro-environmental spillover. *Journal of Environmental Psychology*, 46:55–66. Publisher: Elsevier.

Ulph, A., Panzone, L., and Hilton, D. (2023). Do rational people sometimes act irrationally? a dynamic self-regulation model of sustainable consumer behavior. *Economic Modelling*, 126:106384.

Van de Ven, D.-J., González-Eguino, M., and Arto, I. (2018). The potential of behavioural change for climate change mitigation: a case study for the european union. *Mitigation and adaptation strategies for global change*, 23:853–886.

Van Gestel, L., Adriaanse, M., and De Ridder, D. (2020). Do nudges make use of automatic processing? unraveling the effects of a default nudge under type 1 and type 2 processing. *Comprehensive Results in Social Psychology*, pages 1–21.

Van Rookhuijzen, M., De Vet, E., and Adriaanse, M. A. (2021). The effects of nudges: One-shot only? exploring the temporal spillover effects of a default nudge. *Frontiers in Psychology*, 12.

Wenzig, J. and Gruchmann, T. (2018). Consumer preferences for local food: Testing an extended norm taxonomy. *Sustainability*, 10(5):1313.

Wolstenholme, E., Poortinga, W., and Whitmarsh, L. (2020). Two birds, one stone: The effectiveness of health and environmental messages to reduce meat consumption and encourage pro-environmental behavioral spillover. *Frontiers in psychology*, 11:577111.

Wong-Parodi, G. and Feygina, I. (2021). Engaging people on climate change: The role of emotional responses. *Environmental Communication*, 15(5):571–593.

Xu, L., Zhang, X., and Ling, M. (2018). Spillover effects of household waste separation policy on electricity consumption: evidence from hangzhou, china. *Resources, Conservation and Recycling*, 129:219–231.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.