Grantham Research Institute on Climate Change and the Environment

Centre for Climate Change Economics and Policy

Behavioural spillovers unpacked: estimating the side effects of social norm nudges

Julien Picard and Sanchayan Banerjee

October 2023

Centre for Climate Change Economics and Policy Working Paper No. 427
ISSN 2515-5709 (Online)

Grantham Research Institute on Climate Change and the Environment
Working Paper No. 402
ISSN 2515-5717 (Online)

Economic and Social Research Council

Grantham Foundation for the Protection of the Environment

University of Leeds

The London School of Economics and Political Science

The Centre for Climate Change Economics and Policy (CCCEP) was established by the University of Leeds and the London School of Economics and Political Science in 2008 to advance public and private action on climate change through innovative, rigorous research. The Centre is funded by the UK Economic and Social Research Council. Its third phase started in October 2018 with seven projects:

1. Low-carbon, climate-resilient cities
2. Sustainable infrastructure finance
3. Low-carbon industrial strategies in challenging contexts
4. Integrating climate and development policies for 'climate compatible development'
5. Competitiveness in the low-carbon economy
6. Incentives for behaviour change
7. Climate information for adaptation

More information about CCCEP is available at www.cccep.ac.uk

The Grantham Research Institute on Climate Change and the Environment was established by the London School of Economics and Political Science in 2008 to bring together international expertise on economics, finance, geography, the environment, international development and political economy to create a world-leading centre for policy-relevant research and training. The Institute is funded by the Grantham Foundation for the Protection of the Environment and a number of other sources. It has 13 broad research areas:

1. Biodiversity
2. Climate change adaptation and resilience
3. Climate change governance, legislation and litigation
4. Climate, health and environment
5. Environmental behaviour
6. Environmental economic theory
7. Environmental policy evaluation
8. International climate politics
9. Science and impacts of climate change
10. Sustainable public and private finance
11. Sustainable natural resources
12. Transition to zero emissions growth
13. UK national and local climate policies

More information about the Grantham Research Institute is available at: www.lse.ac.uk/GranthamInstitute

Suggested citation:
Picard J and Banerjee S (2023) *Behavioural spillovers unpacked: estimating the side effects of social norm nudges*. Centre for Climate Change Economics and Policy Working Paper 427/Grantham Research Institute on Climate Change and the Environment Working Paper 402. London: London School of Economics and Political Science

Behavioural Spillovers Unpacked:

Estimating the Side Effects of Social Norm Nudges[†]

Julien Picard[a*], Sanchayan Banerjee[b]

(Last version available here)

Abstract

Fighting the climate crisis requires changing many aspects of our consumption habits. Previous studies show that a first pro-environmental action can lead to another. But does this spillover effect persist when nudges foster the initial action? We model the mechanisms leading nudges to alter such behavioural spillovers. In an online experiment (n=2775), we test if encouraging vegetarianism with a social norm nudge alters environmental donations. The nudge is effective in increasing intentions to choose vegetarian food. Using machine learning, we find that a subgroup drives this effect. We also see a positive spillover effect: choosing vegetarian food increases donations. However, the nudge crowds out this spillover effect for the subgroup identified with machine learning. Our results suggest that social norm nudges are effective but crowd out people's willingness to do more. *JEL* codes: C36, C93, D91, Z18.

Keywords: social norm; meat; climate change; behavioural spillovers; side effects

Credit author statement (details here):

JP: conceptualisation, data curation, formal analysis, investigation, funding acquisition, software, methodology, project administration, visualisation, writing.
SB: funding acquisition.

[†]We are particularly grateful to Eugenie Dugoua, Marion Dumas, Elisabeth Gsottbauer and Gregor Singer for their invaluable advice. We also thank Susana Mourato, Kelsey Jack, Romain Espinosa, Jean-Marc Bourgeon, Mark Andor, Meryem Yankol-Schalck, Sarah Gharbi and many conference participants for their insightful comments on the paper. We also thank the Royal Geographic Society for its financial support.
[a]Department of Geography and Environment, London School of Economics, London.
[b]Department of Environmental Economics, Vrije Universiteit Amsterdam.
*Corresponding author: Julien, Picard; `j.r.picard@lse.ac.uk`.

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 increase people's likelihood to vote for green parties. Conversely, Mazar and Zhong (2010) find that people become less altruistic after buying green products. Thus, promoting actions yielding large decreases in carbon emissions is not enough. These actions should inspire people to do more for the environment to foster lifestyle changes. But do positive behavioural spillovers persist when policies cause the initial pro-environmental action?

In this paper, we develop a model and 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). Cutting on meat yields large reductions in greenhouse gas emissions (Green et al., 2015; Riahi et al., 2022). However, we do not know whether choosing vegetarian food 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 recycling,[2] promote sustainable diets,[3] improve water and electricity consumption,[4]

[1]See Rhodes, Shulman, and McClaran (2020) and Melnyk, van Herpen, Trijp, 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 Abrahamse and Steg (2013); Andor and Fels (2018); Cialdini and Jacobson (2021); Farrow, Grolleau, and Ibanez (2017).

[2]See for instance Andersson and von Borgstede (2010); Bratt (1999); Fornara, Carrus, Passafaro, and Bonnes (2011); Nigbur, Lyons, and Uzzell (2010).

[3]See for instance Richter, Thøgersen, and Klöckner (2018); Salmivaara and Lankoski (2019); Sparkman and Walton (2017); Sparkman, Weitz, Robinson, Malhotra, and Walton (2020); Stea and Pickering (2019); Testa, Russo, Cornwell, McDonald, and Reich (2018); Wenzig and Gruchmann (2018).

[4]See for instance Allcott (2011); Carrico and Riemer (2011); Costa and Kahn (2013); Ferraro, Miranda, and Price (2011); Handgraaf, De Jeude, and Appelt (2013); Lapinski, Rimal, DeVries, and Lee (2007); Nolan, Schultz, Cialdini, Goldstein, and Griskevicius (2008).

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.

We model the effect of nudges on food choices and the effect of food choices on donations in a utility maximisation framework. To our knowledge, only Goetz, Mayr, and Schubert (2022) and Alt and Gallier (2022) use theoretical models to rationalise behavioural spillovers. Our model is novel in that it links the signs of the side effects of nudges with their causal mechanisms. Our approach draws upon the theories proposed by Akerlof and Kranton (2000) and Bénabou and Tirole (2011) on how social context and personal values influence individual choices. In our model, the motives behind the first pro-environmental action determine behavioural spillovers.[6] Acting pro-environmentally out of intrinsic motivation (e.g. because it is "something we care about") raises the cost of reverting to self-serving behaviours (positive behavioural spillover). For instance, quitting meat due to conviction may encourage us to do more to stay true to our beliefs. Acting pro-environmentally out of extrinsic motivation (e.g., as "a means to an end") reduces the cost of reverting to self-serving behaviours (negative behavioural spillover). For instance, quitting meat out of social pressure can license us to do less as the initial action is not self-driven.

In the model, nudges affect non-targeted pro-environmental behaviours through two channels. The first is indirect. Nudges foster the initial pro-environmental decision, triggering behavioural spillovers. The second is direct and either amplifies or weakens the spillover effect. Its sign depends on whether nudges play on intrinsic or extrinsic motivations to foster the initial pro-environmental action. We label this direct effect a crowding-in/out effect. The sign of

[5] See for instance Goldstein, Cialdini, and Griskevicius (2008); Reese, Loew, and Steffgen (2014); Schultz, Khazian, and Zaleski (2008).

[6] For an extended version of the model, see Picard (2023).

crowding-in/out effects indicates the mechanisms through which nudges operate. Crowding-in effects imply that nudges have led individuals to act out of intrinsic motivation. Crowding-out effects imply that nudges have led individuals to act out of extrinsic motivation.

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 dataset to predict this inclination based on respondents' social-demographic characteristics, attitudinal information and self-reported beliefs (Friedman, 2001). We then use this algorithm to classify respondents in 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, Sekhon, Bickel, and Yu (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 do not choose more vegetarian food but make less carbon-intensive food choices when nudged. The nudge does not affect the choices of respondents predicted to be unwilling to conform and respondents predicted to be already conforming. Bryan, Tipton, and Yeager (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 crowds out donations of those predicted to be trying to conform. This crowding-out effect dominates the positive behavioural spillover effect. Our model 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". Whenever the social norm nudge succeeds in increasing vegetarian food choices, it is at the cost of crowding out this willingness to do more.

We contribute to a burgeoning literature studying the side effects of policies. The meta-analyses of Maki et al. (2019) and Geiger, Brick, Nalborczyk, Bosshard, and Jostmann (2021) find only weak evidence for behavioural spillovers. However, methodological discrepancies make studies hard to compare.[7] This could explain this scarcity of compelling evidence.

[7] Some studies compare respondents exposed to a policy with those allocated to a control group (Carrico, Raimi, Truelove, & Eby, 2018; Goetz et al., 2022; Jessoe, Lade, Loge, & Spang, 2021; Liu, Kua, & Lu, 2021; Van Rookhuijzen, De Vet, & Adriaanse, 2021; Wolstenholme, Poortinga, & Whitmarsh, 2020). This method does not distinguish policies' crowding-in/out effect from behavioural spillovers. Other studies randomly offer participants the targeted behaviours to estimate spillover effects (Alt & Gallier, 2022; Clot, Della Giusta, & Jewell, 2022; Margetts & Kashima, 2017). This design supposes that choosing (not) to do the targeted behaviour is the same as (not) being proposed to do it. This assumption is, however, debatable.

Our estimation strategy aligns with Bonev (2023)'s recommendations to estimate behavioural spillovers. To our knowledge, only Alacevich, Bonev, and Söderberg (2021), Comin and Rode (2023), and Alt, Bruns, and DellaValle (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 sign of their crowding-in/out 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. We present our theoretical model in Section II. In Section III, we present our empirical framework. In Section IV, we share the experimental results from a social norm nudge promoting vegetarianism. Section V concludes.

II Modelling Behavioural Spillovers

Positive behavioural spillovers are often explained using cognitive dissonance theory (Festinger, 1962). Cognitive dissonance theory posits that people prefer staying consistent across their choices. This is particularly true when pro-environmental deeds signal an altruistic identity. This raises the mental discomfort of behaving at odds with this identity. Conversely, moral licensing theory explains negative behavioural spillovers (Merritt, Effron, & Monin, 2010). In this case, a first pro-environmental deed allows decision-makers to subsequently act selfishly.[8]

In our model, the motivation for the first pro-environmental action determines the sign of behavioural spillovers. This allows us to reconcile the predictions made by these two theories. Behavioural policies interact with spillovers by altering these motivations.

[8]For extensive discussions on the psychological drivers of behavioural spillover effects, see Dolan and Galizzi (2015) and Truelove, Carrico, Weber, Raimi, and Vandenbergh (2014).

Motivations: We consider two motives for acting pro-environmentally. Individuals' decisions are intrinsically motivated or extrinsically motivated. Intrinsic motivations imply "identity-driven" decisions[9] (we act pro-environmentally because it is *who we are*). For instance, imagine the fictional character of Anne. She stops eating meat because she cares about the environment. Her colleagues praise her when eating with them, reinforcing her convictions. This may induce her to do other pro-environmental behaviours to avoid feeling inconsistent.[10] Experimental evidence shows that consistent behaviours are more frequent when reminding people of their past actions (e.g., Gneezy, Imas, Brown, Nelson, and Norton 2012; Lacasse 2016; Van der Werff, Steg, and Keizer 2013, 2014), after pledging (e.g., Banerjee, Galizzi, John, and Mourato 2022; Lokhorst, Werner, Staats, van Dijk, and Gale 2013), or when labelling people as pro-environmental (e.g., Baca-Motes, Brown, Gneezy, Keenan, and Nelson 2013; Lacasse 2016).

On the other hand, extrinsic motivations imply decisions done as *a means to an end*. Extrinsic motivations range from seeking material rewards (e.g., a tax rebate) to more intangible rewards (e.g., not feeling excluded). In such instances, moral licensing is more likely to occur. To illustrate this, imagine now the fictional character of Bob, a colleague of Anne. Bob chooses the vegetarian option at the cafeteria because he does not want Anne to judge him. Eating vegetarian does not enhance his pro-environmental identity. Bob acts pro-environmentally to avoid disapproving looks. After eating vegetarian, Bob can feel relieved and indulge in other self-serving behaviours. Some studies show that social pressure or monetary incentives can lead people to slacken after an effort (e.g., Dolan and Galizzi 2014; Hartmann, Marcos, and Barrutia 2023; Kristofferson, White, and Peloza 2014; Steinhorst and Matthies 2016; Tiefenbeck, Staake, Roth, and Sachs 2013; Xu, Zhang, and Ling 2018).

[9]Following Akerlof and Kranton (2000), we define *identity* as a sense of self. To enhance this sense of self, one has to engage in certain behaviours (e.g., good Samaritans help people, Harley Davidson bikers prefer beer over hot milk, and environmentalists sort their waste).

[10]As observed early on by Adam Smith, one does not need to be scrutinised by others to fear being inconsistent: *"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).

Context: The context in which people make decisions influences the relative importance of each motive. Educative pieces of information can induce intrinsically motivated pro-environmental deeds, for instance. Conversely, situations where rewards are contingent on acting pro-environmentally can induce extrinsically motivated actions. Thus, how our past choices influence our current ones depends on the context in which we made these choices. In this model, we assume that individuals perfectly remember the context in which they make decisions.

Decision utility: Let us define the function describing individuals' decision processes. We consider a simple two-period model. We assume that decision-makers only consider the current period when making a choice. This simplifying assumption, made without loss of generality for the results pertinent to this paper, allows us to mirror the conditions of our experiment presented in Section IV.[11] Their decision utility at period 2 is of the form:

$$U_2 \equiv u(x_2|I_1, E_1, \eta_2, \epsilon_2) - c_2 \cdot x_2 \tag{II.1}$$

x_2 is the amount of effort spent acting pro-environmentally in period 2. In our experiment, x_2 corresponds to donations to pro-environmental charities. x_1 corresponds to the efforts made for the environment when choosing food. η_2 denotes the propensity of individuals to act out of intrinsic motivation. ϵ_2 denotes the propensity of individuals to act out of extrinsic motivation. Decision utility (II.1) is increasing and concave in x_2, η_2 and ϵ_2. The marginal utility of doing pro-environmental deeds x_2 increases in η_2 and ϵ_2. c_2 is the cost of exerting one unit of pro-environmental effort. It captures the difficulty of acting pro-environmentally (e.g., the number of steps before making an online donation to a charity). The context in which decisions are made alters parameters c_2, η_2 and ϵ_2.

Functions I_1 and E_1 capture the influence of choices of period 1 on choices of period 2, such that $I_1 : \{x_1, \eta_1\} \mapsto \mathbb{R}^+$ and $E_1 : \{x_1, \epsilon_1\} \mapsto \mathbb{R}^+$. They represent how individuals remember being

[11] The implications of relaxing this assumption are explored in detail in Picard (2023).

intrinsically or extrinsically motivated, respectively. Both functions are increasing and concave in their arguments with positive cross-derivatives. Positive cross-derivatives imply that the higher the pro-environmental effort x_1 and the higher η_1 (ϵ_1), the more individuals remember they were intrinsically (extrinsically) motivated. The decision utility of period one is defined similarly, with I_0 and E_0 given. In what follows, we focus on the effect of pro-environmental decisions of period 1 on period 2. We ignore the effect of the context of period two (i.e., c_2, η_2 and ϵ_2) on period two choices. Let us define two competing mechanisms influencing the utility of x_2: *consistency* and *moral licensing*.

Definition 1. *Consistency describes an increase in pro-environmental effort following a first pro-environmental deed. It occurs when remembering that one was intrinsically motivated in period* 1 *increases the utility of doing* x_2 *(*$\partial_{I_1} U_2 > 0$ *and* $\partial_{x_2 I_1} U_2 > 0$*).*

Definition 2. *Moral licensing describes a decrease in pro-environmental effort following a first pro-environmental deed. It occurs when remembering that one was extrinsically motivated in period* 1 *reduces the utility of doing* x_2 *(*$\partial_{E_1} U_2 < 0$ *and* $\partial_{x_2 E_1} U_2 < 0$*).*

Main effect of behavioural policies We assume a social planner seeking to increase individuals' pro-environmental efforts in period 1. To do this, she designs a policy that either increases individuals' motivations to act pro-environmentally in period 1 or reduces the difficulty of acting pro-environmentally. Following Löfgren and Nordblom (2020)'s typology, we refer to policies altering motivations as *preference* nudges. *Preference* nudges can take the form of communication campaigns. We refer to policies altering the difficulty of acting pro-environmentally as *choice architecture* nudges. Such nudges can simply consist of making pro-environmental options more salient. In the model, *preference* nudges increase parameters η_1 or ϵ_1. We denote by θ_1 the parameter altered by the policy, such that $\theta_1 \in \{\eta_1, \epsilon_1\}$. *Choice architecture* nudges decrease parameter c_1.

Lemma 1. *In period one, behavioural policies increase pro-environmental efforts when increasing* $\theta_1 \in$

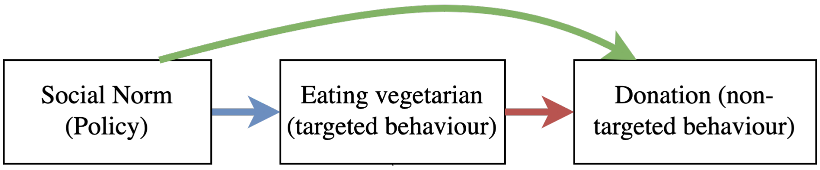


Figure 1: Side effects a policy

Note: Causal mechanisms of the effects of a policy on non-targeted decisions. The red arrow represents behavioural spillovers, whilst the green arrow represents crowding-out/in effects.

$\{\eta_1, \epsilon_1\}$ *or reducing* c_1.

$$\frac{\partial x_1^*}{\partial \theta_1} = \frac{\partial_{x_1\theta_1} U_1}{-\partial_{x_1x_1} U_1} > 0 \qquad \frac{\partial x_1^*}{\partial c_1} = \frac{1}{\partial_{x_1x_1} U_1} < 0 \tag{II.2}$$

See the proof in Appendix A. x_1^* is the amount of pro-environmental that maximises individuals' decision utility.

Side effects of policies: *Preference* nudges affect the decision of period 2 through two channels as described by equation (II.3). Figure 1 illustrates these channels in the context of our experiment. The first channel is through the effect of the behaviour targeted by the policy (x_1) on the non-targeted behaviour (x_2). We refer to this effect as a "behavioural spillover". The red arrow represents this "behavioural spillover" in Figure 1. The second channel is through the effect of the policy on individuals' motivations. The green arrow represents this "crowding-in/out" effect" in Figure 1.

Lemma 2. *Preference nudges* $\theta_1 \in \{\eta_1, \epsilon_1\}$ *alter decisions of period two through two channels captured by the following equation:*

$$\underbrace{\frac{dx_2^*}{d\theta_1}}_{\textit{Side effect}} = \frac{1}{-\partial_{x_2x_2} U_2}\Bigg(\underbrace{\partial_{x_1x_2} U_2}_{\textit{Behavioural spillover}} \times \underbrace{\frac{\partial x_1^*}{\partial \theta_1}}_{\textit{Main effect}} + \underbrace{\partial_{x_2\theta_1} U_2}_{\textit{Crowding-in/out effect}} \Bigg) \tag{II.3}$$

Choice architecture nudges only affect decisions of period 2 through a behavioural spillover effect.

Lemma 3. *Choice architecture nudges only alter decisions of period 2 through one channel captured by the following equation:*

$$\underbrace{\frac{dx_2^*}{dc_1}}_{\textit{Side effect}} = \underbrace{\frac{\partial_{x_1x_2}U_2}{-\partial_{x_2x_2}U_2}}_{\textit{Behavioural spillover}} \times \underbrace{\frac{\partial x_1^*}{\partial c_1}}_{\textit{Main effect}} \tag{II.4}$$

See Appendix A for proofs of lemma 2 and 3. The sign of the behavioural spillover effect indicates whether individuals are more intrinsically motivated (and therefore more consistent) than they are extrinsically motivated (prone to engage in moral accounting):

$$\partial_{x_1x_2}U_2 = \underbrace{\partial_{x_2I_1}U_2 \cdot \partial_{x_1}I_1}_{\text{Consistence } (>0)} + \underbrace{\partial_{x_2E_1}U_2 \cdot \partial_{x_1}E_1}_{\text{Moral licensing } (<0)} \tag{II.5}$$

It captures the degree of complementarity between x_1 and x_2. x_1 and x_2 are complementary (substitutable) when the behavioural spillover is positive (negative). In other words, doing x_1 increases (decreases) the marginal utility of doing x_2. The crowding-in/out effect captures the effect of the policy on period two choices through its impact on motivations:

$$\partial_{x_2\eta_1}U_2 = \underbrace{\partial_{x_2I_1}U_2 \cdot \partial_{\eta_1}I_1}_{\text{Crowding-in}}, \quad \partial_{x_2\varepsilon_1}U_2 = \underbrace{\partial_{x_2E_1}U_2 \cdot \partial_{\varepsilon_1}E_1}_{\text{Crowding-out}} \tag{II.6}$$

When playing on intrinsic (extrinsic) motivations, policies increase (decrease) the marginal utility of doing x_2.

Proposition 1. *Behavioural spillovers and crowding-in/out effects can have opposite signs.*

The sign of the behavioural spillover effect does not depend on the policy. Conversely, the crowding-in/out effect is positive when policies affect individuals' intrinsic motivations. Therefore, policies that increase intrinsic motivations strengthen people's willingness to do extra pro-environmental actions.

Proposition 2. *Playing on intrinsic motivations maximises the side effects of policies.*

In the next section, we develop an experimental design which allows us to estimate these two effects.

III Empirical Strategy

Estimating behavioural spillovers allows us to find behaviours leading to other pro-environmental actions. Crowding-in/out effects tell us if policies weaken or reinforce behavioural spillovers. Crowding-in/out effects also inform us if policies affect intrinsic or extrinsic motivations. However, estimating behavioural spillovers and crowding-in/out effects 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 play on intrinsic motivations for some people and extrinsic motivations for others. Thus, crowding-in/out effects 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 $\boldsymbol{x}_2$.:

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

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

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 side effects 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 we showed in Section II, the side effect of policies is composed of a behavioural spillover effect and a crowding-in/out effect. 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 spillovers from crowding-in/out effects consists of fitting the following linear model:

$$x_{2i} = \tilde{\alpha} + \tilde{\beta}^{BS} x_{1i} + \tilde{\beta}^{C} \theta_{1i} + \tilde{\varepsilon}_i \tag{III.3}$$

$\hat{\tilde{\beta}}^{BS}$ is a naive estimate of the behavioural spillover. $\hat{\tilde{\beta}}^{C}$ is the naive estimate of the crowding-in/out effect. 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. 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 spillovers and crowding-in/out effects with two-stage least squares:

$$\begin{aligned} &\text{Stage 1: } x_{1i} = \alpha + \beta_1 c_{1i} + \beta_2 \theta_{1i} + \varepsilon_i \\ &\text{Stage 2: } x_{2i} = \alpha' + \beta^{BS} \hat{x}_{1i} + \beta^{C} \theta_{1i} + \varepsilon'_i \end{aligned} \tag{III.4}$$

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 spillovers and crowding-in/out effects are unbiased when this is the case. Furthermore, one can derive the following proposition:

Proposition 3. *Estimates of models* (III.2) *and* (III.4) *are such that:*

$$\underbrace{\hat{\beta}^{SE}}_{\textit{Side effect}} = \underbrace{\hat{\beta}^{BS}}_{\textit{Behavioural spillover}} \times \underbrace{\hat{\beta}^{ME}}_{\textit{Main effect}} + \underbrace{\hat{\beta}^{C}}_{\textit{Crowding in/out effect}} \tag{III.5}$$

See Appendix A for the proof. Proposition 3 shows that we can interpret estimates of model (III.2) and (III.4) in the same way as equation (II.3) derived in Section II.

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. We estimate the conditional average treatment effects of policy θ_1 in three steps. First, estimate the function $y_i = f(W_i')$ such that:

$$\hat{f} \in \arg\min_{f} L(y_i, f(W_i')) \tag{III.6}$$

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

$$\hat{y}_i = \hat{f}(W_i) \tag{III.7}$$

Finally, estimate treatment effects for each type.

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

IV Application

In the rest of this article, we study the side effects of a social norm nudge promoting vegetarian diets. We do this in an online experiment delivered to 2,775 English respondents. 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 crowds out or crowds in this spillover effect. In subsection IV.A, we detail the design of this experiment. Then, subsection IV.B presents our empirical strategy. Results are presented in subsection IV.C. Finally, we explore heterogeneity in subsection IV.D.

IV.A Experimental Design and Data Collection

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 2.[12] We use the *training* sample to look at the heterogeneity in our treatment effects as part of an exploratory analysis (see subsection IV.D).

Policy: The policy of interest in this experiment is a social norm nudge. More precisely, we consider the following dynamic social norm message[13]:

[12]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.

[13]We construct it using the study of Stewart, Piernas, Cook, and Jebb (2021) analysing UK meat consumption trends using data from the National Diet and Nutrition Survey.

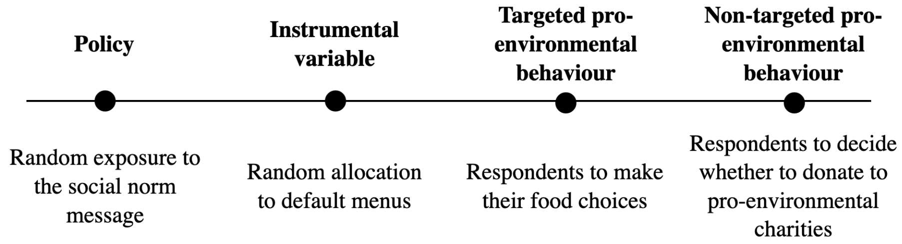


Figure 2: Timeline of the experiment

> *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, Attwood, Vennard, and Mayneris (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 III, 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 7 and 6 in Appendix 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 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-

Table 1: Sample sizes of treatment groups

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

environmental decision is whether participants choose vegetarian food.[14] 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 5 in Appendix D). As such, all participants have the same information on the environmental consequences of their choices. Table 22 in Appendix 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.[15] 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.

[14]To mitigate any biases stemming from this choice being hypothetical, we ask two questions inspired by the literature on willingness-to-pay estimation (Andor, Frondel, & Vance, 2017; Champ, Moore, & Bishop, 2009; Mohammed, 2012; Ready, Champ, & Lawton, 2010). Namely, participants can revise their choices before continuing the survey:
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*]
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 the respondents have asked to revise their choices, and only 15.56% of them would either somewhat disagree or strongly disagree with going to a restaurant serving the same menus.

[15]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 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 for the difference in age between the treatment and the control group. We use a Chi-square test for gender differences. We use trend tests for the differences in education and income between the two groups.

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 2 shows descriptive statistics per treatment group. The median respondent is 35 years old, 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 9 in Appendix D). Randomisation was successful. No significant differences exist across the treatment groups for age, gender, income and education. About 98.28% of the *main* sample has passed an attention check placed at the beginning of the survey.[16] From these 98.28%, 99.75% passed a focus check we placed after the pre-treatment questionnaire.[17] Furthermore, 81.69% of the participants passed a manipulation check between the food choice and the donation task.[18] This suggests that respondents were attentive when taking the survey.

[16] After consenting to participate in the survey experiment, respondents are screened based on whether they provide the correct answer to an attention check. Namely, 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'."*

[17] 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."*

[18] 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.B Estimation Strategy

To estimate the effects of the social norm nudge on donations, we fit a linear model analogous to specifications (III.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.[19]

As we showed in Section II, the effect of the nudge on donations is composed of a behavioural spillover effect and a crowding in/out effect. A naive approach to disentangle these two effects consists of fitting an OLS model analogous to specification (III.3):

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

$FoodChoice_i$ 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 crowding in/out effect of the nudge. To tackle potential omitted variable biases, we instrument respondents' food choices by $Menu_i$, the dummy equal to 1 if vegetarian items are salient, 0 otherwise. We use a specification analogous to model (III.4), estimated with two-stage least squares

[19] Estimate $\hat{\beta}^{SE}$ corresponds to an intention-to-treat effect. In Appendix 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.

(2SLS):[20]

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

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.[21] We also add lasso-selected controls to increase the precision of our estimates (see Appendix C, Belloni, Chernozhukov, and Hansen 2014). We report standard p-values corrected for the false discovery rate (Benjamini & 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, McCrary, Moreira, and Porter (2022)'s procedure.

IV.C Results

Main effect: Table 3 presents the effect of the social norm nudge on food choices.[22] The nudge increases intentions to choose vegetarian food by 6.7 percentage points and reduces the carbon footprint of food choices by 11.8%. Results are robust to non-linear probit specifications (see Table 9 in Appendix C).

Side effects: Table 4 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 total side effects of the nudge. In both panels, these

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

[21]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).

[22]Analyses were conducted on R using the package *estimatr* (Blair, Cooper, Coppock, Humphreys, & Sonnet, 2022).

Table 3: 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<0.001	q=0.003
Vegetarian salient	0.115***	−7.875***
	(0.016)	(0.928)
	q<0.001	q<0.001
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).

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 crowding in/out 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 crowding in/out 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

10 and 11 in Appendix 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 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, Jaeger, & Baker, 1995; Staiger & 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; Ortmann, Ryvkin, Wilkening, & Zhang, 2023; Van Gestel, Adriaanse, & De Ridder, 2020). This confirms priors in the literature (e.g., see Hansen and 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 systematically choose meat-based options when vegetarian items are salient and vice-versa. Nonconformist participants are more likely to behave as defiers: they always oppose what is suggested. This behaviour is likely to be orthogonal to motivations driving pro-environmental decisions. 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, Imbens, and Rubin (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.

Table 4: Total side effects, behavioural spillovers and crowding-in/out 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<0.001		q<0.001
R^2	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=0.002		q=0.002
R^2	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 overall side effect of the social norm nudge on donations. The other columns show estimates of behavioural spillovers and the crowding-in/out effect. The second and the fourth columns show estimates of the social norm nudge and food choices on donations. The third and the fifth columns display the same estimates where, this time, we instrument food choices. 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).

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 10 in Appendix D). Another caveat regards the hypothetical nature of food choices. An experimenter's demand effect can inflate 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 12 in Appendix C). Besides, our results align with evidence from field experiments finding positive behavioural spillovers between pro-environmental actions (Alacevich et al., 2021; Comin & Rode, 2023). Finally, our empirical strategy relied on Assumption 1 (see Section III). 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 13 in Appendix C).

As highlighted in Section III, average treatment effects can hide heterogeneity. We explore the heterogeneity of our causal effects in subsection IV.D.

IV.D 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 plays on different cognitive 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:[23]

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 machine learning classifier (GBM) to predict respondents' answers based on attitudinal measures and social-demographic characteristics. Then, we use the algorithm

[23] 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.

to predict the types of respondents in the *main* sample.[24] 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 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.[25] We estimate the average performance of GBM using nested 10×10 folds cross-validation. Overall, GBM performs twice better than chance. Appendix 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 11, 12 and 13 in Appendix D).

Profile of predicted types: Table 24 in Appendix 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 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

[24]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.

[25]The reader can refer to Gareth, Daniela, Trevor, and Robert (2013) for more information on how these algorithms work.

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:[26]

$$FoodChoice_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + u_i \tag{IV.4}$$

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

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

And:

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

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:

$$Donation_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + u_i \tag{IV.5}$$

Here again, β_k is the average side-effect of the social norm nudge conditional on being predicted to be of type k. To investigate heterogeneity in the crowding-in/out effect, we fit the following model:

$$Donation_i = \alpha + \sum_{k \in \Omega_-} \mathbf{1}_k \delta_k + \sum_{k \in \Omega} \mathbf{1}_k \beta_k Norm_i + \beta_2 \widehat{FoodChoice}_i + \varepsilon_i \tag{IV.6}$$

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

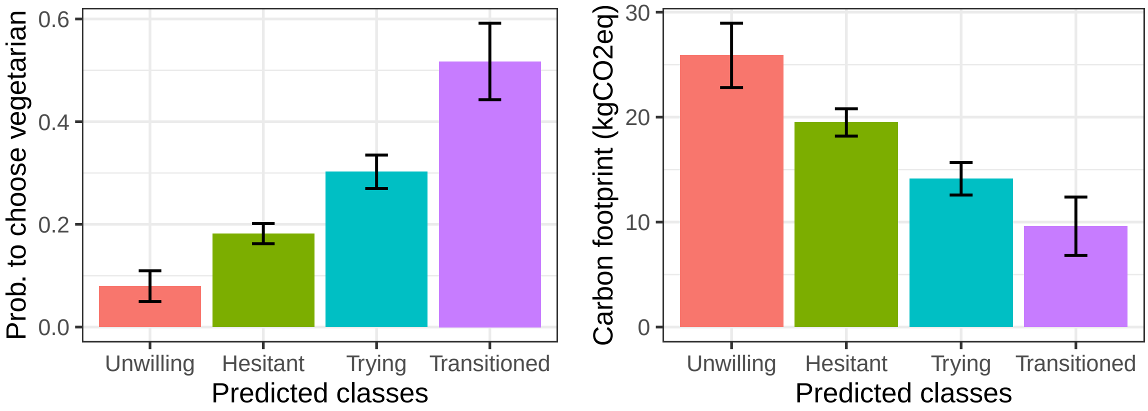


Figure 3: Food choices of each predicted type

Here, the coefficient β_k is the crowding in/out 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, 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 5 displays the results obtained by fitting equation (IV.4). 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 pp). These results are robust (see Tables 14 and 15 in Appendix 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 6 shows the results of regression (IV.5) in the first two columns and regression (IV.6) in the last four columns. Although not significant after p-value correction, the social norm nudge

Table 5: 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=0.006
Social norm × Trying	0.108***	−2.261
	(0.033)	(1.577)
	q=0.001	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 resampling are displayed last (q).

negatively affects the amount given by the predicted *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 crowding out effect is globally robust (see Tables 18, 19, 20 and 21 in Appendix C). Again, p-values of re-randomisation tests confirm that leverage is not driving statistical significance. We also observe suggestive evidence of a crowding-in 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 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 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 & 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. Overall, this heterogeneity suggests that experimenter demand is unlikely to drive our results. Indeed, not all types behave in the direction expected by this bias.

We find robust evidence that the social norm message crowds out donations of the predicted *trying* type. Our model in Section II suggests the *trying* respondents may have treated the social norm message as an extrinsic pressure to choose vegetarian food. This would have induced

Table 6: Crowding-in/out 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=0.008	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 overall effect of the social norm nudge on donations. The crowding-in/out 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).

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 crowding-in effect of the social norm message on the predicted *unwilling*. Although not statistically significant, this would explain why the average crowding-in/out 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, Iliev, & Medin, 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.

V Conclusion

In Section II, we model the side effects of behavioural 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. We label the second a *crowding-in/out* 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 spillovers from the crowding-in/out effects of the social norm nudge. 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 reductions 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. Although the effect of this nudge is heterogeneous. We find the social norm nudge to work for people who we predict as trying to change their diets to follow the norm and those hesitating 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 crowding-out effect outweighs the positive behavioural spillover effect. We do not observe a similar crowding-out effect on the respondents 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. We have laid out a theoretical and empirical framework to assess this question. We hope this paper will provide the methodological foundations for further research.

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.

Competing interests

The authors have no competing interests to declare.

Code and Data Availability

The code and data for the analysis are available upon request.

References

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

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

Alacevich, C., Bonev, P., & 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., & 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., & 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., & 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., & 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., & 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., & 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., & Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, *39*(5), 1070–1084.

Banerjee, S., Galizzi, M. M., John, P., & Mourato, S. (2022). What works best in promoting climate citizenship? a randomised, systematic evaluation of nudge, think, boost and nudge+. *A Randomised, Systematic Evaluation of Nudge, Think, Boost and Nudge+(April 26, 2022). Research Square pre-prints*.

Banerjee, S., & 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*, 102024.

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

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

Benjamini, Y., & 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., & Sonnet, L. (2022). estimatr: Fast estimators for design-based inference [Computer software manual]. (https://declaredesign.org/r/estimatr/, https://github.com/DeclareDesign/estimatr)

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

Bonev, P. (2023). Behavioral spillovers.

Bound, J., Jaeger, D. A., & 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., & 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., & 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., & 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., & Bishop, R. C. (2009). A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review*, *38*(2), 166–180.

Cialdini, R. B., & 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., & 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., & Rode, J. (2023). *Do green users become green voters?* (Tech. Rep.). National Bureau of Economic Research.

Costa, D. L., & 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.

Dolan, P., & Galizzi, M. M. (2014). Because i'm worth it: a lab-field experiment on the spillover effects of incentives in health.

Dolan, P., & 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.

Evans, W. N., & 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., & 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., & 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. (1962). Cognitive dissonance. *Scientific American*, *207*(4), 93–106.

Fornara, F., Carrus, G., Passafaro, P., & 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*, 1189–1232.

Gareth, J., Daniela, W., Trevor, H., & 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., & 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., & Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science*, *58*(1), 179–187.

Goetz, A., Mayr, H., & 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., & 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., ... 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., & 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., & 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.

Hartmann, P., Marcos, A., & Barrutia, J. M. (2023). Carbon tax salience counteracts price effects through moral licensing. *Global Environmental Change*, *78*, 102635.

Jessoe, K., Lade, G. E., Loge, F., & 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.

Kristofferson, K., White, K., & 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., & 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.

Lapinski, M. K., Rimal, R. N., DeVries, R., & 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., & Porter, J. (2022). Valid t-ratio inference for iv. *American Economic Review*, *112*(10), 3260–3290.

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

Löfgren, Å., & Nordblom, K. (2020). A theoretical framework of decision making explaining the mechanisms of nudging. *Journal of Economic Behavior & Organization*, *174*, 1–12.

Lokhorst, A. M., Werner, C., Staats, H., van Dijk, E., & Gale, J. L. (2013). Commitment and behavior change: A meta-analysis and critical review of commitment-making strategies in environmental research. *Environment and behavior*, *45*(1), 3–34.

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

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

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

Mazar, N., & 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*.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass*, *4*(5), 344–357.

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

Nigbur, D., Lyons, E., & 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., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and social psychology bulletin*, *34*(7), 913–923.

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

Picard, J. (2023). Double-edged nudges? micro-foundations to behavioural interventions and their spillover effects. *Available at SSRN 3784009*.

Ready, R. C., Champ, P. A., & 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., & 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., & 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., … others (2022). Mitigation pathways compatible with long-term goals. *IPCC, 2022: Cli-*

mate 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., & Klöckner, C. A. (2018). A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability*, *10*(8), 2848.

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

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

Scarborough, P., Appleby, P. N., Mizdrak, A., Briggs, A. D., Travis, R. C., Bradbury, K. E., & 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., & 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, . . . 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*.

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

Sparkman, G., & 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., & Walton, G. M. (2020). Developing a scalable dynamic norm menu-based intervention to reduce meat consumption. *Sustainability*, *12*(6), 2453.

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

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

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

Stewart, C., Piernas, C., Cook, B., & 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., & 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., & 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.

Tiefenbeck, V., Staake, T., Roth, K., & 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., & 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 der Werff, E., Steg, L., & Keizer, K. (2013). It is a moral issue: The relationship between environmental self-identity, obligation-based intrinsic motivation and pro-environmental behaviour. *Global environmental change*, *23*(5), 1258–1265.

Van der Werff, E., Steg, L., & Keizer, K. (2014). I am what i am, by looking past the present: the influence of biospheric values and past behavior on environmental self-identity. *Environment and Behavior*, *46*(5), 626–657.

Van Gestel, L., Adriaanse, M., & 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*, 1–21.

Van Rookhuijzen, M., De Vet, E., & 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., & Gruchmann, T. (2018). Consumer preferences for local food: Testing an extended norm taxonomy. *Sustainability*, *10*(5), 1313.

Wolstenholme, E., Poortinga, W., & 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.

Xu, L., Zhang, X., & 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.

A Appendix: Proofs

Proof. (Main effect of policies) The effect of a policy altering $\theta_1 \in \{\eta_1, \epsilon_1\}$ on choices of period 1 as described by equation (II.2) is derived as follows. Individuals maximise their period one utility. By assumption, they only consider period 1 when choosing period 1 pro-environmental effort. As such, they solve the following:

$$\partial_{x_1} U_1 - c_1 = 0 \tag{A.1}$$

Where x_1^* is the solution to equation (A.1). It is a function of I_0, E_0, x_0, η_1, ϵ_1, and c_1. Using the implicit function theorem, we differentiate (A.1) with respect to $\theta_1 \in \{\eta_1, \epsilon_1\}$:

$$\partial_{x_1 x_1} U_1 \frac{\partial x_1^*}{\partial \theta_1} + \partial_{x_1 \theta_1} U_1 = 0 \Leftrightarrow \frac{\partial x_1^*}{\partial \theta_1} = \frac{\partial_{x_1 \theta_1} U_1}{-\partial_{x_1 x_1} U_1} > 0$$

The same reasoning applies when $\theta_1 = c_1$. ■

Proof. (Spillovers of policies) The optimal level of pro-environmental efforts at period 2 is a function of the choices of period one. Expressing the side effects of a policy at period one on choices of period two as in equation (II.3) amounts to using the implicit function theorem and differentiating the first order conditions of the maximisation programme at period 2 with respect to $\theta_1 \in \{\eta_1, \epsilon_1\}$:

$$\begin{aligned} & \partial_{x_2 x_2} U_2 \tfrac{d x_2^*}{d \theta_1} + \partial_{x_2 x_1} U_2 \tfrac{\partial x_1^*}{\partial \theta_1} + \partial_{x_2 \theta_1} U_2 = 0 \\ \Leftrightarrow \quad & \tfrac{d x_2^*}{d \theta_1} = \tfrac{1}{-\partial_{x_2 x_2} U_2} \left[\partial_{x_2 x_1} U_2 \tfrac{\partial x_1^*}{\partial \theta_1} + \partial_{x_2 \theta_1} U_2 \right] \end{aligned}$$

The same reasoning applies when $\theta_1 = c_1$. ■

Proof. (Proposition 3) First, model (III.4) can be rewritten in a reduced form as below:

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

Where:

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

This implies that:

$$\beta^{BS} = \frac{\overline{\beta}_1}{\beta_1} \quad \beta^{C} = \overline{\beta}_2 - \frac{\overline{\beta}_1}{\beta_1}\beta_2 \tag{A.4}$$

Using ordinary least square, we can show that the coefficients of model (A.2) are equal to:

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

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 (III.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 (III.4) 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

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

■

B Appendix: 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 oversampling 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$^{\text{th}}$ 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.[27] Here, GBM performs better than chance for each type and, on average, 1.9 times better than chance across all types (see Table 7).

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 performing better than chance.[28] 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 7).

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

[27] 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 $\frac{\frac{n_i}{4}}{\frac{n_i}{4}+\frac{n-n_i}{4}} = \frac{n_i}{n}$.

[28] 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 $\frac{\frac{n_i}{4}}{\frac{n_i}{4}+\frac{3n_i}{4}} = \frac{1}{4}$.

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 7.[29] Results in Table 7 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 7 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, here again, the difference is not statistically significant (see Table 7).

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.

[29]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: $2 \times \frac{\frac{n_i}{n} \times \frac{1}{4}}{\frac{n_i}{n} + \frac{1}{4}} = \frac{2 \times n_i}{n_i \times 4 + n}$.

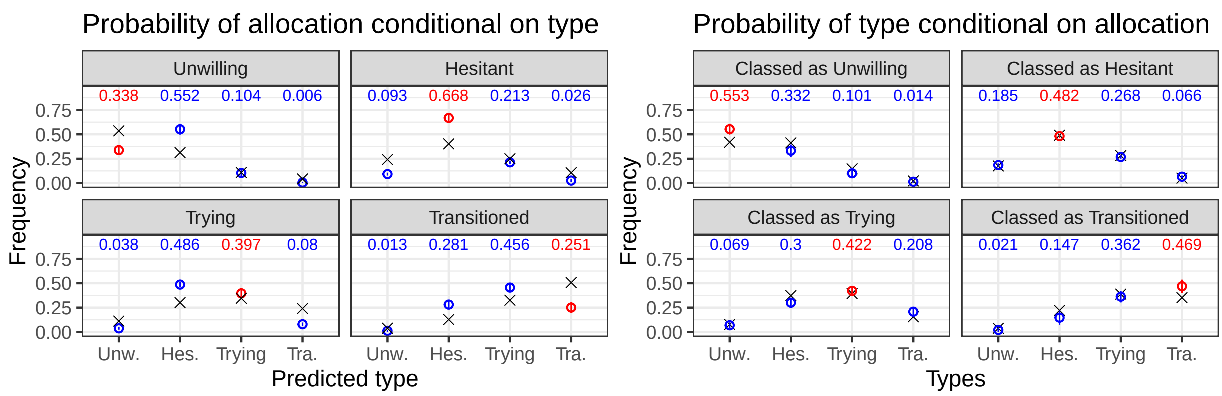
Table 7: Estimated performance of GBM

	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.

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 4 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 4 indicates that over-sampling has increased the ability of the algorithm to correctly identify 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 4. In other words, over-sampling seems to make GBM better at detecting the



(a) Miss-classification by type

(b) Miss-classification by predicted type

Figure 4: 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.

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 8. 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'

Table 8: 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

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 14, 15, and 16 in Appendix D.

C Appendix: Robustness Checks

Table 9: 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 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 10: Robustness checks of the behavioural spillovers and crowding-in/out effects of the social norm message 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 spillovers and crowding-in/out 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 11: Robustness checks of the behavioural spillovers and crowding-in/out effects of the social norm message 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 spillovers and crowding-in/out 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 12: 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 13: 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 that we interact with the social norm nudge. Robust standard errors are displayed in parentheses.

Table 14: 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=0.005	q=0.002	q=0.001	q<0.001
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=0.001	q=0.004	q<0.001	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 15: 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=0.008	q=0.005	q=0.007	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 oversampling 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 resampling are displayed last (q).

Table 16: 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=0.003	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 17: 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=0.007	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 18: Robustness checks of the crowding-out/in effect of the social norm message 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=0.009	q=0.001	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=0.001	q=0.005	q=0.005	q=0.005	q=0.008	q=0.008
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

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

Note: Crowding-out/in 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 19: Robustness checks of the crowding-out/in effect of the social norm message 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=0.003	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=0.001	q=0.004	q=0.002	q=0.005	q=0.008	q=0.005
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

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

Note: Crowding-out/in 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 20: Robustness checks of the crowding-out/in effect of the social norm message 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=0.008	q=0.001	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=0.003	q=0.000	q=0.001	q=0.000	q=0.002	q=0.002	q=0.001
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

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

Note: Crowding-out/in 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).

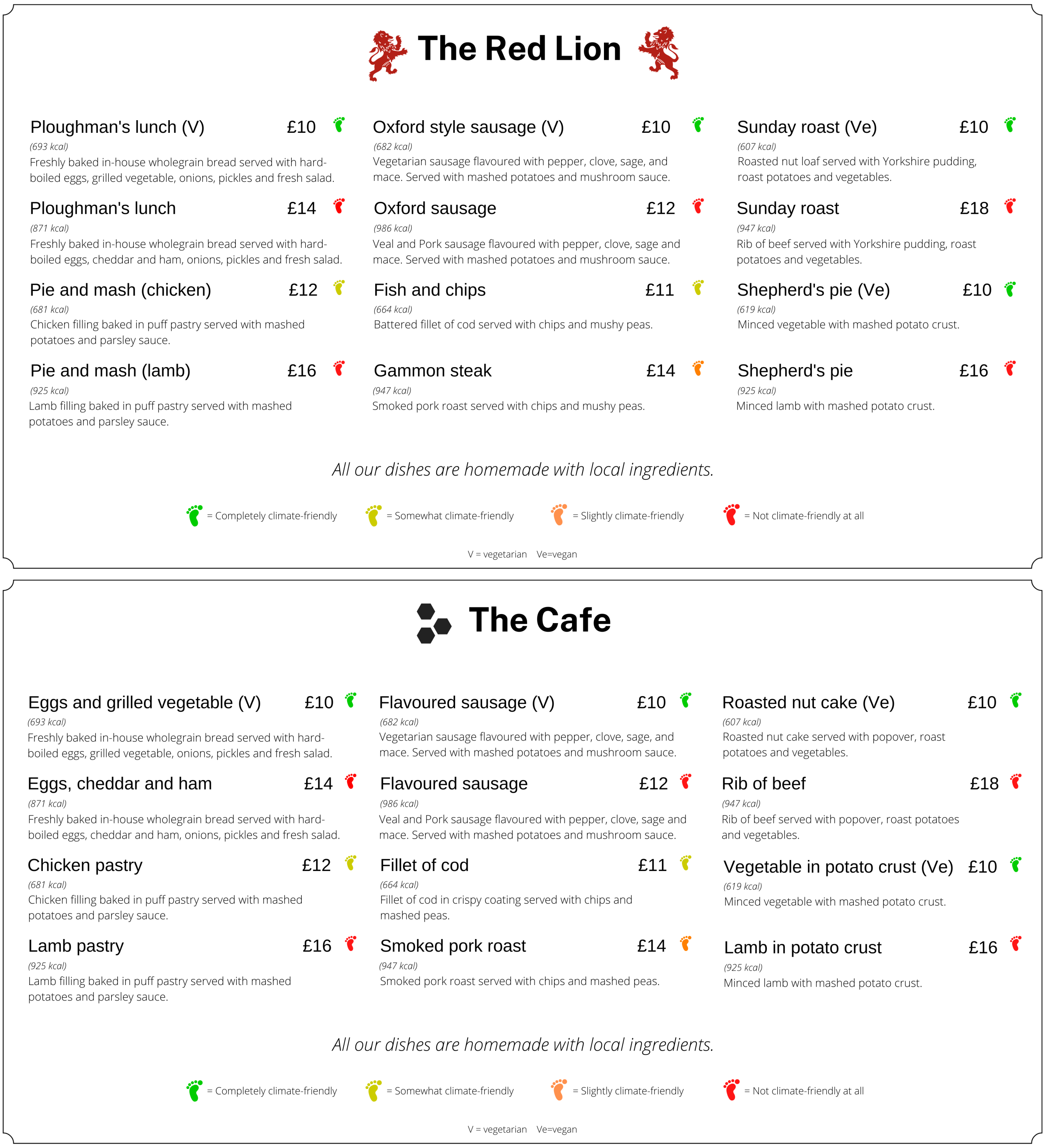
Table 21: Robustness checks of the crowding-out/in effect of the social norm message conditional on predicted classes IV

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=0.005	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=0.003	q=0.000	q=0.001	q=0.001	q=0.001	q=0.001	q=0.001
Num.Obs.	2730	2431	2730	2730	2730	2730	2730

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

Note: Crowding-out/in 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).

D Appendix: Supplementary Tables and Figures



The Red Lion

Ploughman's lunch (V) £10
(693 kcal)
Freshly baked in-house wholegrain bread served with hard-boiled eggs, grilled vegetable, onions, pickles and fresh salad.

Ploughman's lunch £14
(871 kcal)
Freshly baked in-house wholegrain bread served with hard-boiled eggs, cheddar and ham, onions, pickles and fresh salad.

Pie and mash (chicken) £12
(681 kcal)
Chicken filling baked in puff pastry served with mashed potatoes and parsley sauce.

Pie and mash (lamb) £16
(925 kcal)
Lamb filling baked in puff pastry served with mashed potatoes and parsley sauce.

Oxford style sausage (V) £10
(682 kcal)
Vegetarian sausage flavoured with pepper, clove, sage, and mace. Served with mashed potatoes and mushroom sauce.

Oxford sausage £12
(986 kcal)
Veal and Pork sausage flavoured with pepper, clove, sage and mace. Served with mashed potatoes and mushroom sauce.

Fish and chips £11
(664 kcal)
Battered fillet of cod served with chips and mushy peas.

Gammon steak £14
(947 kcal)
Smoked pork roast served with chips and mushy peas.

Sunday roast (Ve) £10
(607 kcal)
Roasted nut loaf served with Yorkshire pudding, roast potatoes and vegetables.

Sunday roast £18
(947 kcal)
Rib of beef served with Yorkshire pudding, roast potatoes and vegetables.

Shepherd's pie (Ve) £10
(619 kcal)
Minced vegetable with mashed potato crust.

Shepherd's pie £16
(925 kcal)
Minced lamb with mashed potato crust.

All our dishes are homemade with local ingredients.

= Completely climate-friendly = Somewhat climate-friendly = Slightly climate-friendly = Not climate-friendly at all

V = vegetarian Ve=vegan

The Cafe

Eggs and grilled vegetable (V) £10
(693 kcal)
Freshly baked in-house wholegrain bread served with hard-boiled eggs, grilled vegetable, onions, pickles and fresh salad.

Eggs, cheddar and ham £14
(871 kcal)
Freshly baked in-house wholegrain bread served with hard-boiled eggs, cheddar and ham, onions, pickles and fresh salad.

Chicken pastry £12
(681 kcal)
Chicken filling baked in puff pastry served with mashed potatoes and parsley sauce.

Lamb pastry £16
(925 kcal)
Lamb filling baked in puff pastry served with mashed potatoes and parsley sauce.

Flavoured sausage (V) £10
(682 kcal)
Vegetarian sausage flavoured with pepper, clove, sage, and mace. Served with mashed potatoes and mushroom sauce.

Flavoured sausage £12
(986 kcal)
Veal and Pork sausage flavoured with pepper, clove, sage and mace. Served with mashed potatoes and mushroom sauce.

Fillet of cod £11
(664 kcal)
Fillet of cod in crispy coating served with chips and mashed peas.

Smoked pork roast £14
(947 kcal)
Smoked pork roast served with chips and mashed peas.

Roasted nut cake (Ve) £10
(607 kcal)
Roasted nut cake served with popover, roast potatoes and vegetables.

Rib of beef £18
(947 kcal)
Rib of beef served with popover, roast potatoes and vegetables.

Vegetable in potato crust (Ve) £10
(619 kcal)
Minced vegetable with mashed potato crust.

Lamb in potato crust £16
(925 kcal)
Minced lamb with mashed potato crust.

All our dishes are homemade with local ingredients.

= Completely climate-friendly = Somewhat climate-friendly = Slightly climate-friendly = Not climate-friendly at all

V = vegetarian Ve=vegan

Figure 5: 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 22: 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.

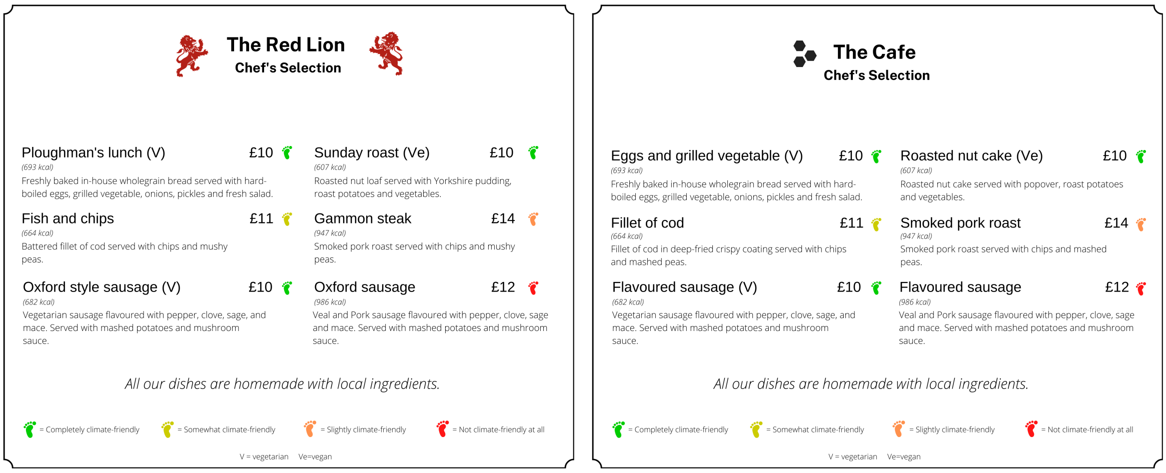


Figure 6: Plant-intensive default menus

Note: these are the two versions of the plant-intensive default menus shown to participants.

Figure 7: Meat-intensive default menus

Note: These are the two versions of the meat-intensive default menus shown to participants.

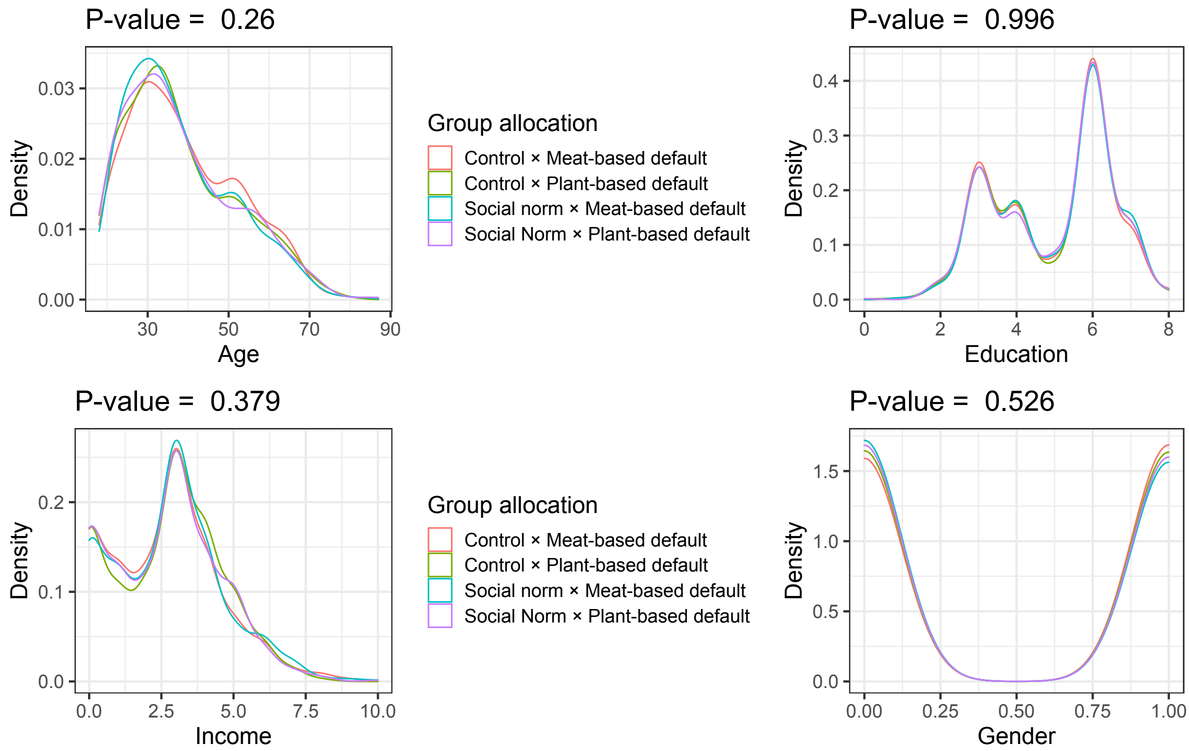


Figure 8: Distribution of the main covariates by treatment group

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 23: 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.

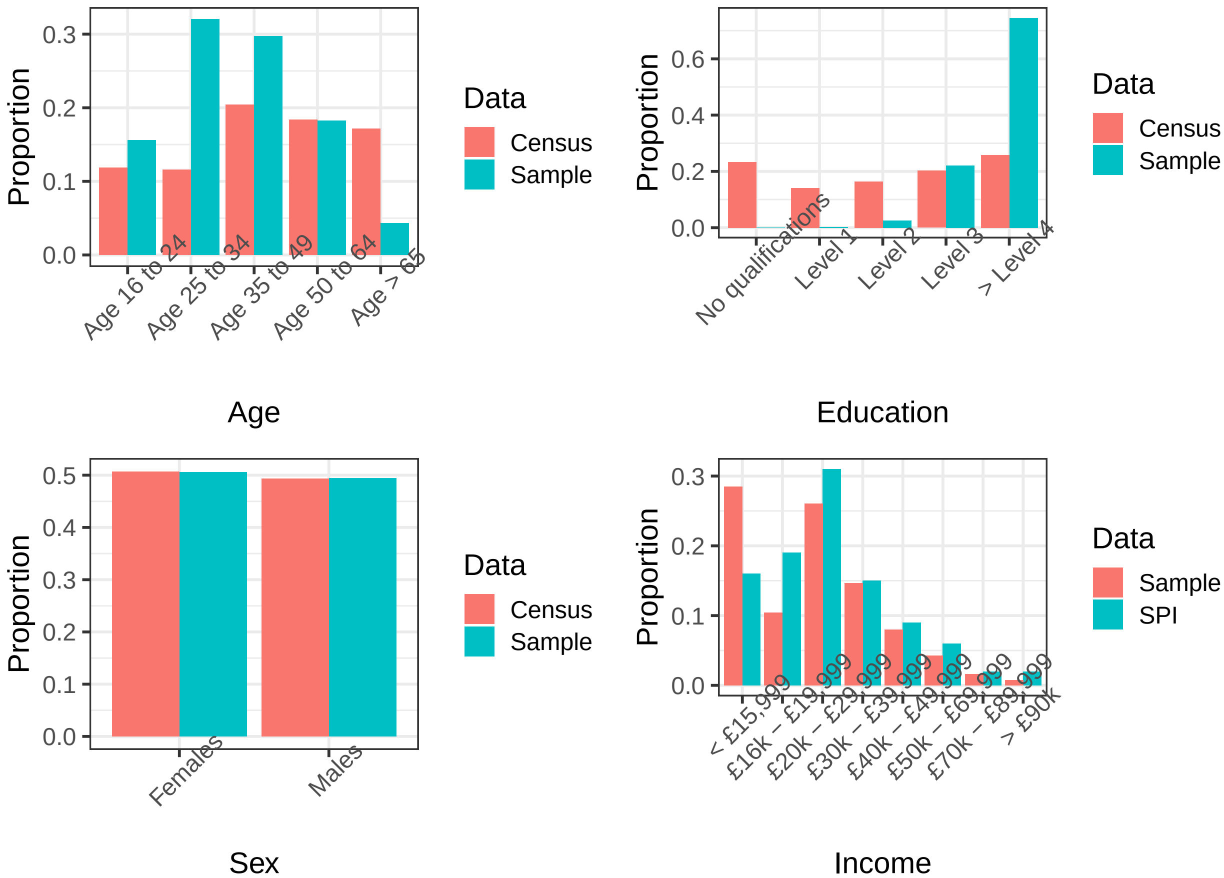


Figure 9: Comparison with UK population

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.

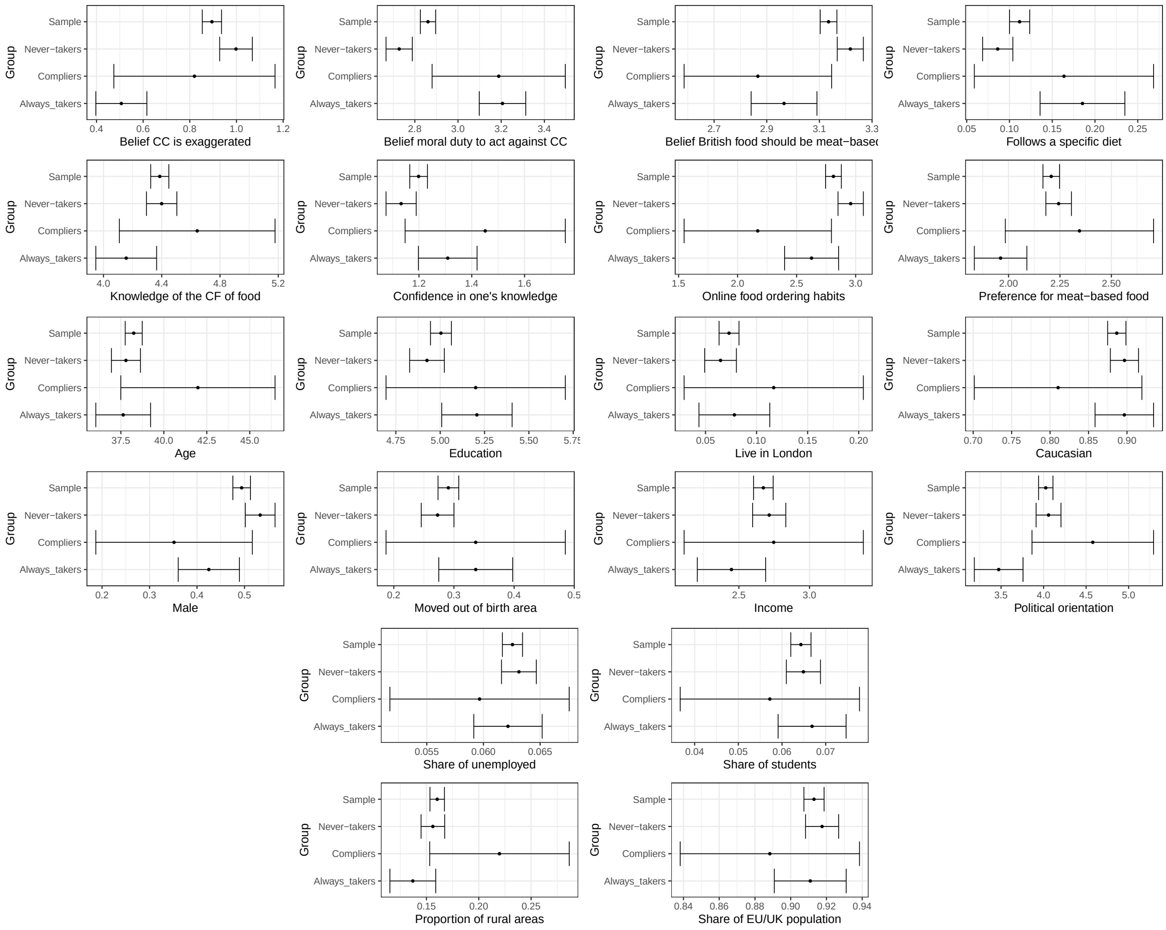


Figure 10: Profile of compliers

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).

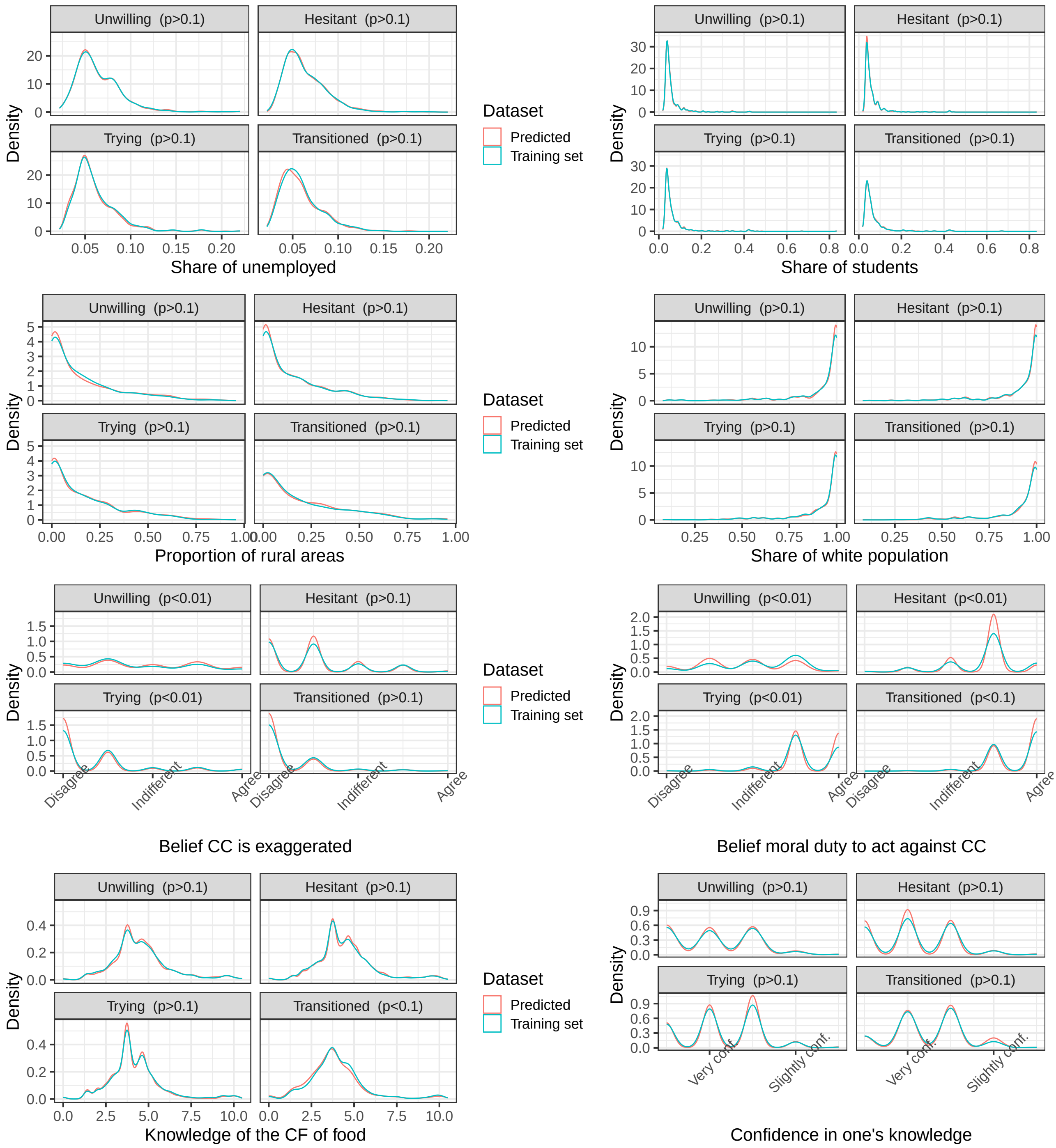


Figure 11: Distribution of the predictors by type I

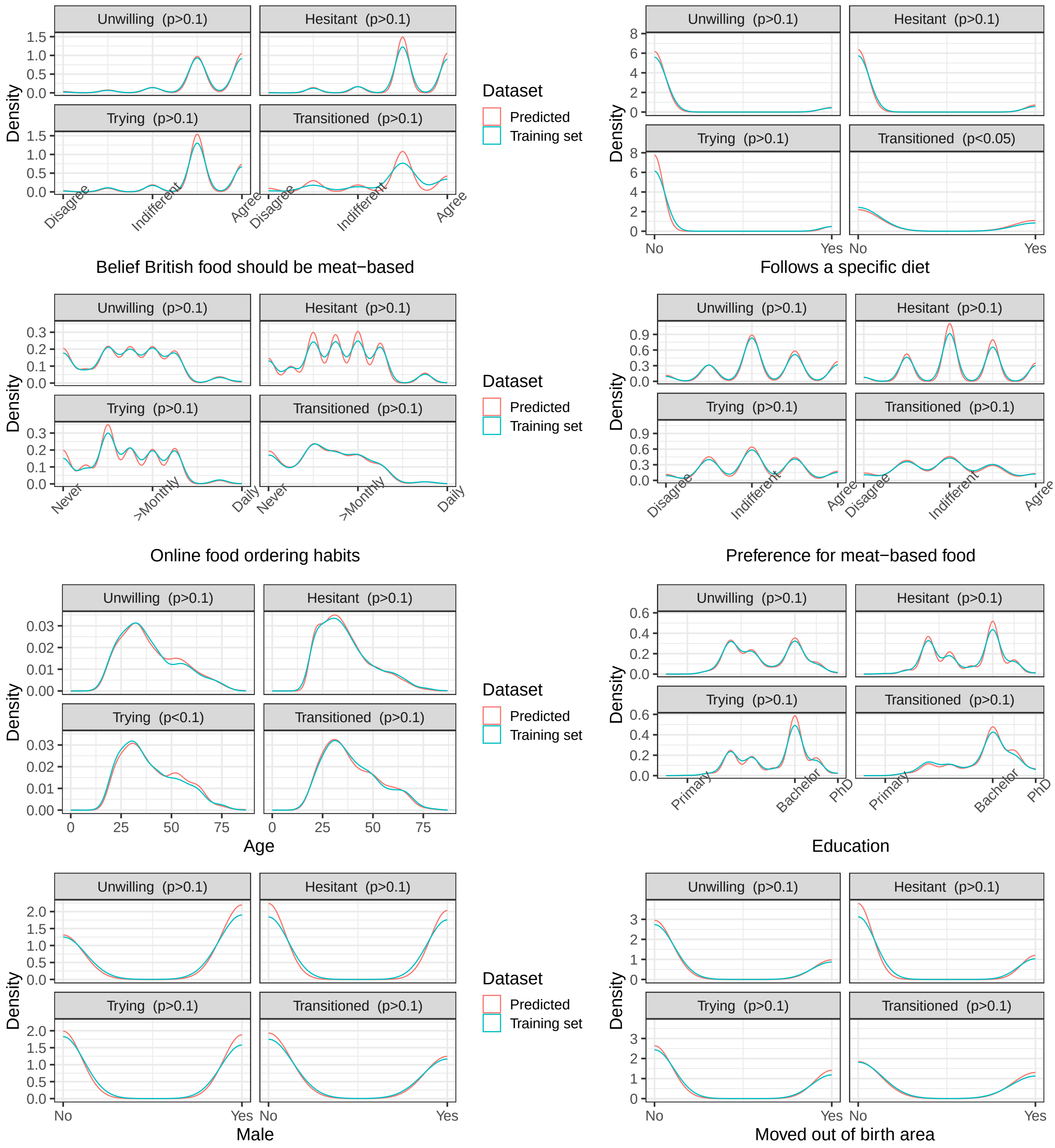


Figure 12: Distribution of the predictors by type II

Table 24: 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.

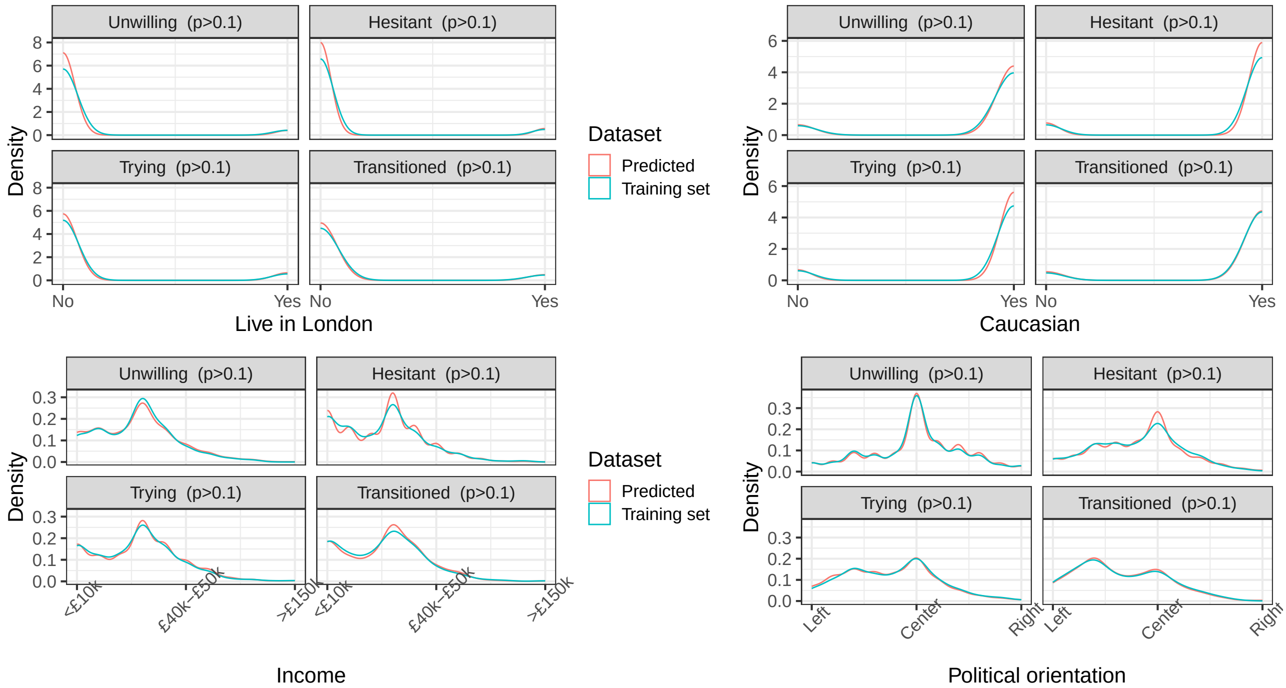


Figure 13: Distribution of the predictors by type III

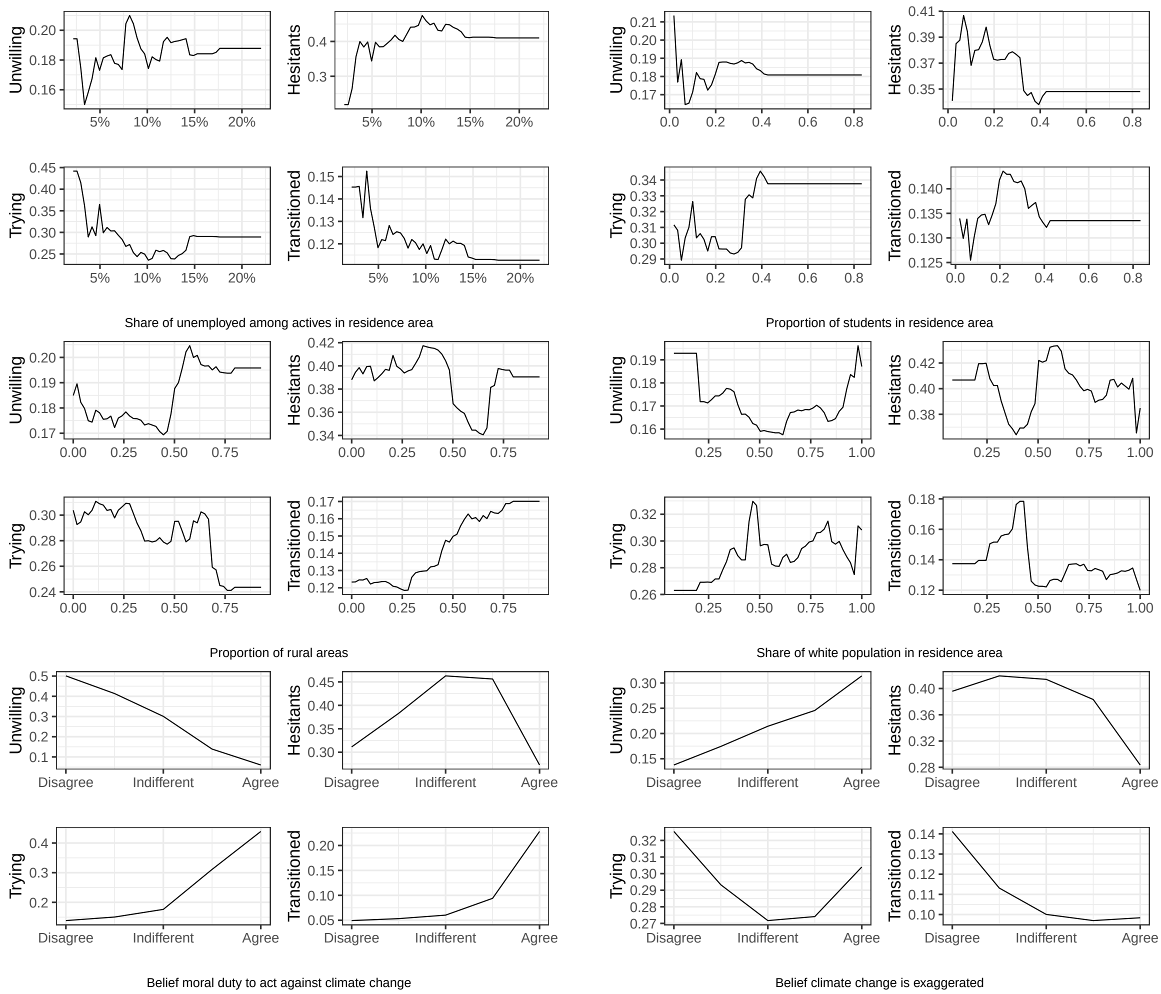


Figure 14: Partial dependence plots of the GBM algorithm I

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

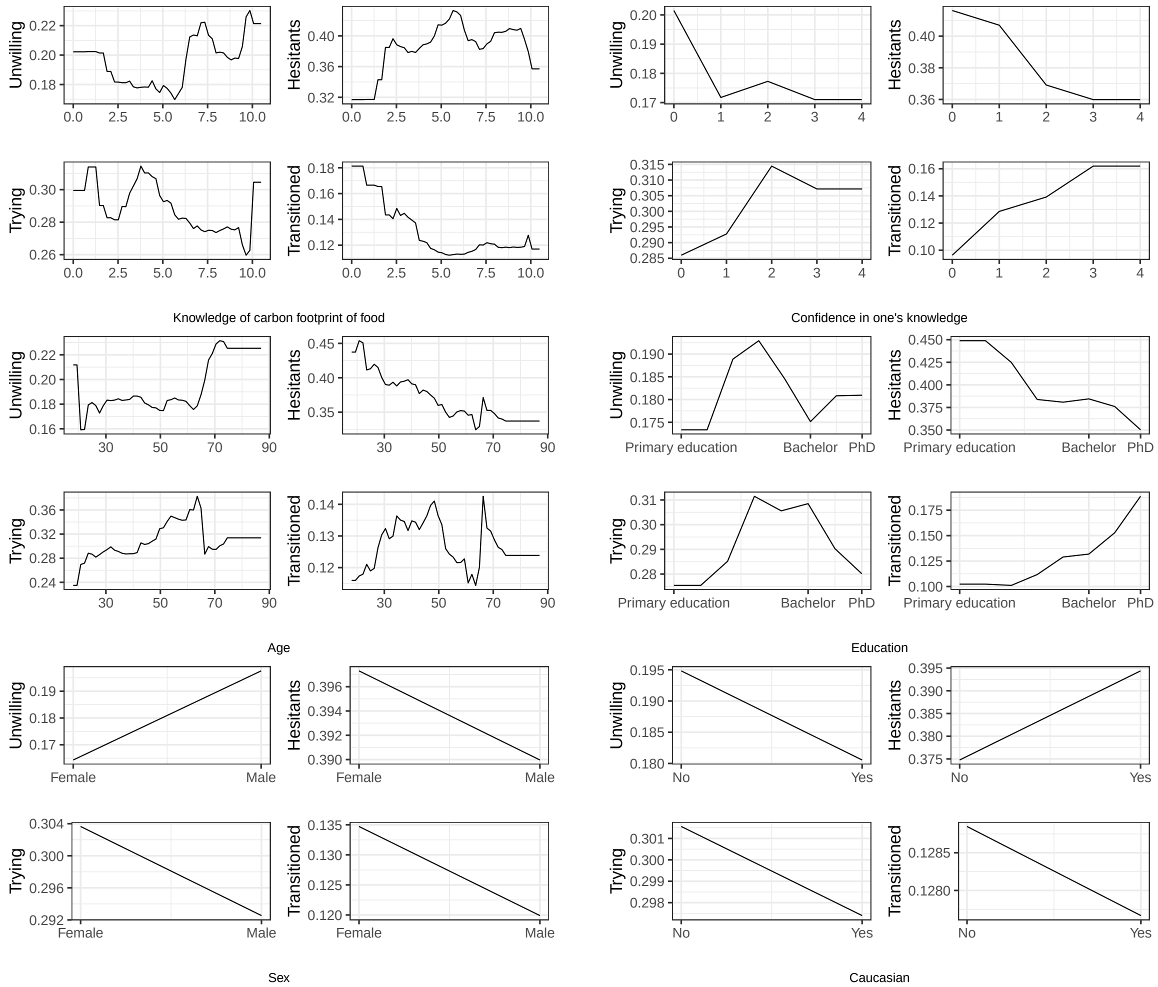


Figure 15: Partial dependence plots of the GBM algorithm II

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

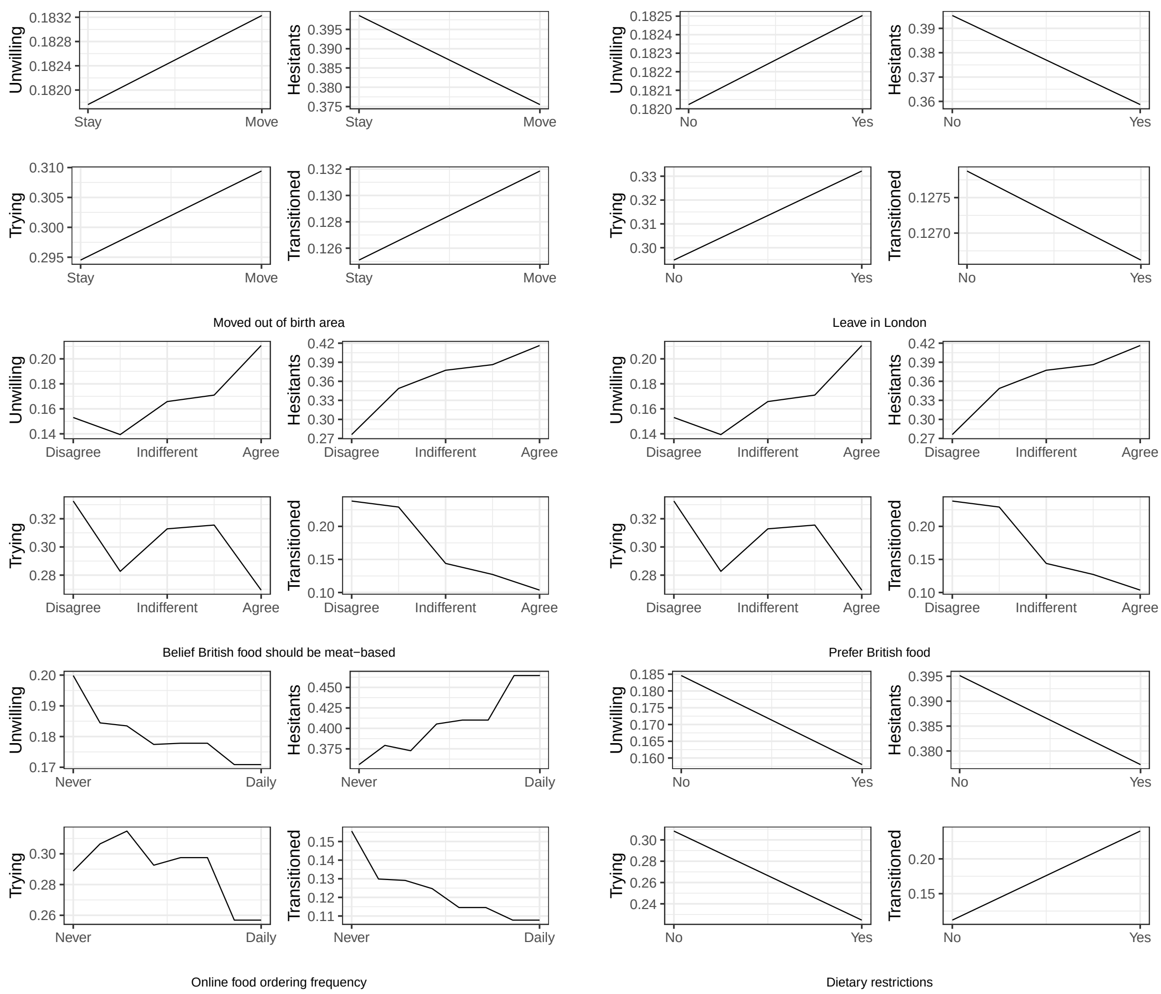


Figure 16: Partial dependence plots of the GBM algorithm III

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