ANALYSIS PLAN

With this experimental design, and based on the existing literature, we aim at testing the following hypotheses:

Hypotheses related to motivated reasoning:
H1: Individuals hold overly optimistic beliefs regarding the environmental performance of their own meal boxes;
H2: Individuals update their beliefs asymmetrically, assigning greater weight to favourable information than to unfavourable information regarding the environmental performance of their meal boxes;

Hypotheses related to behavior:
H3: Individuals exhibiting stronger motivated reasoning mechanisms make smaller downward revisions to the carbon footprint of their meal boxes following unfavourable information;

H4a: Informational intervention is less effective among individuals exhibiting stronger motivated belief mechanisms;

H4b: Motivated reasoning mechanisms moderate the effectiveness of informational interventions more strongly than the effectiveness of carbon taxation.

Indeed, taxes alter incentives directly and therefore leave less room for motivated reinterpretation than informational interventions.

Tests of the different hypotheses:

To test H1, we will compare:

(i) Subjects' stated probability that their meal box has a lower carbon footprint than the average carbon footprint of meal boxes selected during the baseline sessions, denoted as p_i, and

(ii) the objective realization of this event, i.e., a dummy variable equal to 1 if the actual carbon footprint of the participant's meal box is below the baseline average and 0 otherwise, denoted as y_i.

We will then construct the variable $d_i = p_i - y_i$, where positive values indicate overly optimistic beliefs regarding the environmental performance of one's own meal box.
Our primary analysis tests H0: $E[d_i] = 0$ vs. H1: $E[d_i] > 0$ using a one-tailed one-sample t-test.
Since elicited beliefs may depart from normality and are bounded between 0 and 1, we will additionally report results from a non-parametric Wilcoxon signed-rank test as a robustness check.

To test H2, we will follow an approach similar to Drobner and Georg (2024). We will transform beliefs using a logit transformation, on which Bayesian updating reduces to adding or subtracting a fixed amount from prior beliefs, depending on whether the signal is favourable or unfavourable. We will estimate this model using OLS, with standard errors clustered at the individual level, since each participant contributes two belief updates.
Formally, let $\hat{\mu}_{it}$ denote individual i's belief at round t that their box has a lower carbon footprint than the reference average. Following Drobner and Georg (2024), we will estimate:

$$logit(\hat{\mu}_{it}) = \delta logit(\hat{\mu}_{it-1}) + \beta_P \mathbb{I}(s_{it} = P) \log\left(\frac{q}{1-q}\right) + \beta_N \mathbb{I}(s_{it} = N) \log\left(\frac{q}{1-q}\right) + \epsilon_{it}$$

where $s_{it} \in \{Positive, Negative\}$ is the signal received in round t and $q = 2/3$ is the known signal accuracy. Under pure Bayesian updating, $\delta = \beta_P = \beta_N = 1$.
Our primary test compares the average response to favourable and unfavourable signals using a Wald test on the coefficient equality.

In addition, we will report descriptive statistics on the distribution of prior beliefs and belief revisions, and assess whether optimistic asymmetry varies across individual characteristics, in particular environmental attitudes and environmental concern.

To test H3, H4a and H4b, we will construct variables c_1, c_2 and c_3 that report the carbon footprint corresponding to the first, second and third meal box selected by participants. Then, we will measure $\Delta_i^{2-1} = c_2 - c_1$ and $\Delta_i^{3-2} = c_3 - c_2$.

To study whether motivated beliefs affect subsequent food choices and policy effectiveness, we construct an individual-level measure of motivated reasoning based on subjects' reactions to favorable and unfavorable information.
Our preferred measure of motivated reasonings is a continuous quantification of updating asymmetry. Following the approach of Drobner and Georg (2024), we estimate the extent to which each participant updates their beliefs following favorable and unfavorable signals. We then construct an index measuring the relative weight assigned to favorable versus unfavorable information, with higher values indicating a stronger tendency to discount unfavorable information and therefore stronger motivated reasoning.

For ease of interpretation in some specifications, we will also construct a binary variable, *MR*, equal to one for participants exhibiting above-median levels of the updating asymmetry index and zero otherwise.
As robustness checks, we will consider alternative definitions based on different classification thresholds and on alternative continuous measures of updating asymmetry.

Formally, to test H3 we will estimate the determinants of Δ_i^{2-1} with a linear regression using the subsample of subjects who received at least one negative signal and we will test the significance of the variable *MR* defined above.
Formally, let $\Delta_i^{2-1} = c_i^2 - c_i^1$ be the change in carbon footprint between the initial box and the box revised after the signals. We estimate:

$$\Delta_i^{2-1} = \alpha + \gamma MR_i + X_i'\beta + \epsilon_i$$

on the subsample of subjects who received at least one negative signal, where MR_i is a dummy equal to 1 if subject *i* employed motivated belief mechanisms, and X_i is a vector of controls (initial carbon footprint c^1_i, demographics).

To test H4a and H4b, we will estimate the determinants of Δ_i^{3-2} with a linear regression. We will include the variable MR defined above, as well as a dummy variable “Tax”, taking the value 1 for participants in the tax condition and 0 otherwise, and a dummy variable “Label”, taking the value 1 for participants in the label and boost condition and 0 otherwise. Our main focus will be on the interactions “MR x Tax” and “MR x Label”. This regression will be estimated on the subsample of subjects who received at least one negative signal during phase 1.

Formally, let $\Delta_i^{3-2} = c_i^3 - c_i^2$ be the change in carbon footprint between the revised box and the box chosen after the policy. We estimate:

$$\Delta_i^{3-2} = \alpha + \gamma_1 MR_i + \gamma_2 Tax_i + \gamma_3 Label_i + \gamma_4 (MR_i.Tax_i) + \gamma_5 (MR_i.Label_i) + X_i'\beta + \epsilon_i$$

estimated on the subsample of subjects who received at least one negative signal during Phase 1. Tax_i and $Label_i$ are dummies for the carbon tax condition and the label and boost condition, respectively.

Sample selection and data cleaning:

Observations with boundary beliefs ($\hat{\mu}_{it}$= 0 or $\hat{\mu}_{it}$= 1) will be excluded from the logit regressions used to test H2, as the logit transformation is not defined at these values. We will report as a robustness check results obtained by replacing these boundary values with 0.01 and 0.99 respectively.

For H3, H4a and H4b, given that carbon footprint values are naturally bounded by the range of available recipes, we do not anticipate extreme outliers. Should any implausible values arise (e.g., due to data entry errors), we will exclude them and report our reasoning transparently.