Pre-Analysis Plan: The Determinants of Conspiracy Theory Labeling

Charles Louis-Sidois
WU Vienna

Version Revised Before the Start of the Pilot: July 30, 2026

1 Background and Research Question

Misinformation has become a central topic in both public debate and academic research. Describing this multifaceted phenomenon requires a wide range of labels, including "fake news", "conspiracy theories", "alternative realities", and "lies". Despite the prominence of these labels, little attention has been paid to how they are used and perceived. This paper investigates why individuals label certain pieces of information in a given way, with a particular focus on labels associated with misinformation.

Conspiracy theories constitute our leading example; the experiment also examines the labeling of claims as rumors and misinformation. While there is broad consensus on the definition of conspiracy theories — explanations of events that attribute causal power to secret plots orchestrated by powerful actors — there is substantial disagreement regarding which specific narratives qualify as such. Labeling decisions may partly reflect perceived characteristics of the claim itself, such as its perceived veracity or the severity of the action described, which relate directly to standard definitions of conspiracy theories. Beyond these content-based factors, partisan considerations may also play a role: a respondent may be more likely to label a claim as a conspiracy theory if it is attributed to a politician they dislike, or less likely to do so if the alleged actor is an entity they support. Disentangling these two channels — epistemic and partisan — is one of the central goals of this study.

2 Hypotheses

My main hypotheses relate to the labeling decision as conspiracy theories:

H1 — Perceived veracity. Respondents are more likely to label a claim as a conspiracy theory when they perceive it as implausible (negative relationship between perceived plausibility and labeling).

H2 — Perceived harm. Respondents are more likely to label a claim as a conspiracy theory when they perceive the described action as harmful.

H3 — Affinity with the actor. Respondents are less likely to label a claim as a conspiracy

theory when they have high affinity with the alleged actor, in order to discredit accusations against a group they support.

H4 — Affinity with the source. Respondents are more likely to label a claim as a conspiracy theory when they have low affinity with the political source of the claim, in order to discredit a political figure they do not like.

H5 — Naming an actor or source. Naming a specific actor or source affects the probability of labeling a claim as a conspiracy theory, independently of affinity.

Moreover, I will also test whether affinities with the actor and the source directly affect labeling, or whether the effect operates through a shift in perceived veracity and perceived harm, in the spirit of the motivated reasoning literature.

Turning to the other labeling decisions, I expect the rumor label to depend positively on perceived familiarity — as rumors are typically understood as unconfirmed information that has widely circulated — and negatively on perceived veracity. I have no hypothesis on perceived harm. For the misinformation label, I expect a negative relationship with perceived plausibility, but no clear hypothesis on perceived harm or familiarity. Both labels may also be affected by partisan considerations, as they can be used to discredit claims that contradict a respondent's political views. However, since these terms carry less negative connotations than "conspiracy theory," I expect partisan effects to be weaker for these labels.

3 Experimental Design

The study is a survey experiment conducted on a representative sample of French adults recruited via Prolific. Participants first answer background questions covering socio-demographic characteristics (age, gender, education), political preferences (vote in the first round of the 2022 French presidential election, interest in news and politics), and support for all actors and public figures relevant to the experimental stimuli.

Participants are then exposed to four short claims, each structured around a common conspiracy theory template in which an actor secretly engages in a potentially harmful action. For each claim, one of three versions is randomly and independently assigned. Two claims vary the identity of the alleged actor; two vary the political source attributed to the claim. The order of claims is randomized across respondents. The four claims and their versions are presented in Table 1.

After each claim, respondents answer two sets of questions. The first concerns labeling decisions and determines the key dependent variables of the study: respondents indicate whether they would classify the claim as a conspiracy theory, as misinformation, as a rumor, or as credible information, answering yes or no for each label. The second set elicits respondents' assessments of the claim along three dimensions: the likelihood that the claim is true (1–5 scale), the severity of the described action if confirmed (1–5 scale), and whether they had previously encountered the claim. An attention check is included at the end of each claim block, asking a simple factual question about the content of the claim. Most respondents answer the labeling questions first; a randomly selected subset answers the assessment questions first, as a robustness check. The order of the four labeling categories is randomized across respondents.

At the end of the survey, respondents complete an incentivized belief elicitation task: for each of the four claims, they estimate the share of other respondents who labeled it as a conspiracy theory.

The 10% of respondents whose estimates are closest to the actual shares receive a bonus payment. Respondents are then shown a debriefing statement.

The study is conducted in two waves: a pilot study and a main study, the size of which will be determined based on the pilot results and a formal power calculation.

Versions
Foreign interference: Germany / Russia / Foreign powers discreetly interfere in the affairs of France in order to weaken it and promote their own interests.
Venomous species release: Pharmaceutical companies / Environmental activist groups / People have secretly released venomous species in the French countryside.
Election manipulation: According to Raphaël Glucksmann / Jordan Bardella / [no source], foreign powers secretly influence French elections.
Civil war preparation: According to Jean-Luc Mélenchon / Éric Zemmour / [no source], public authorities are covertly preparing civil war in France.

Table 1: Claims used in the experiment

4 Empirical Specifications and Outcome Variables

4.1 Empirical specification

The first main specification is a pooled regression estimated on the answers to all four claims simultaneously:

$$\begin{aligned} Y_{ic} = \alpha &+ \beta_1 \text{Plausible}_{ic} + \beta_2 \text{Harm}_{ic} + \beta_3 \text{Familiarity}_{ic} \\ &+ \text{ActorAffin}_{ic} + \text{SourceAffin}_{ic} \\ &+ \lambda_c + \mathbf{X}_i'\theta + \varepsilon_{ic} \end{aligned} \tag{1}$$

The dependent variable Y_{ic} is a labeling decision by respondent i for claim c; the full list of outcomes is described in the next subsection.

The variables Plausible_{ic}, Harm_{ic}, and Familiarity_{ic} capture respondents' perceptions of the claim: perceived likelihood that the claim is true, perceived severity of the described action, and prior exposure to the claim, respectively. Their coefficients reveal whether the epistemic channel — the perceived characteristics of the claim — affects labeling decisions.

ActorAffin_{ic} and SourceAffin_{ic} are vectors of dummy variables, one for each level of affinity (from 1 to 5) with the actor or source named in claim c. For claims where no actor or source is named, all dummies are set to 0. These coefficients capture the effect of each affinity level relative to the case where no actor or source is named, and reveal whether the partisan channel affects labeling decisions. Finally, λ_c denotes claim fixed effects and $\mathbf{X}_i$ is a vector of individual controls including age, gender, education, and political interest.

The second main specification consists of separate regressions for each of the four claims:

$$Y_i = \alpha_c + \beta_1^c \text{Plausible}_i + \beta_2^c \text{Harm}_i + \beta_3^c \text{Familiarity}_i + \text{Affin}_i + \mathbf{X}_i' \theta^c + \varepsilon_i \tag{2}$$

where Affin_i refers to affinity with the named actor for actor variation claims, and affinity with the named source for source variation claims. This specification allows all coefficients to vary freely across claims, revealing potential heterogeneity that the pooled specification would otherwise obscure.

4.2 Outcome variables

The specifications above are estimated for the following outcome variables:

- LabelComp_{ic} **[main outcome]**: binary variable equal to 1 if respondent i labels claim c as a conspiracy theory.
- LabelDesin_{ic}: binary variable equal to 1 if respondent i labels claim c as misinformation.
- LabelRum_{ic}: binary variable equal to 1 if respondent i labels claim c as a rumor.
- LabelCred_{ic}: binary variable equal to 1 if respondent i labels claim c as credible information.
- IncentVar_{ic}: respondent i's belief about the share of other respondents labeling claim c as a conspiracy theory.

5 Decomposition of Epistemic and Partisan Channels

The specifications above control for perceived plausibility, harm, and familiarity, isolating the direct effect of affinity on labeling decisions. However, affinity may also affect labeling indirectly by shaping how respondents perceive the claim — for instance, a claim may appear more plausible when attributed to a distrusted actor. This indirect effect is not captured by the first specifications. To disentangle the direct from the indirect effect, I will estimate specifications with perceived plausibility, harm, and familiarity as dependent variables, examining how they depend on affinity with the source and actor. These results will inform the relative importance of the epistemic and partisan channels, and may serve as the basis for a more formal decomposition.

6 Robustness Checks

The following robustness checks are pre-specified:

1. **Exclusion.** Exclude responses from respondents who failed attention checks.
2. **Familiarity restriction.** Restrict the sample to respondents who had no prior familiarity with the claim.
3. **Label order.** Control for the order in which the four labeling categories were displayed to the respondent.
4. **Order effects.** Test whether presenting labeling questions before or after the plausibility and harm questions affects answers.

7 Power Analysis

A power calculation will be conducted after the pilot study and before launching the main study to determine the sample size.

8 Deviations from Pre-Analysis Plan

Any deviations from this pre-analysis plan will be clearly documented and justified in the paper. The specific claims used in the main study may be adjusted relative to those used in the pilot if the pilot results indicate that some claims are perceived as too implausible or too credible by all respondents, leaving insufficient variation to exploit.