Artificial Intelligence, Misinformation, and Ideological Bias: Evidence From a News Classification Experiment.

Riccardo Manghi

LUISS Guido Carli University

Abstract

This paper investigates how ideological bias and classification errors vary across AI-mediated information environments. We conduct an incentivized laboratory experiment in which participants classify news items as true or false across three within-subject phases: a human-written baseline, an AI-generated misinformation phase in which the role of AI is not made explicit, and an AI-detection phase in which participants receive an explicit AI probability signal.

Average classification accuracy does not differ significantly across phases. The central result is instead that errors are strongly structured by ideological congruence: participants are more likely to accept congruent false news as true and to reject incongruent true news as false. The AI-detection phase suggests greater skepticism toward information, reflected in a higher false-negative rate. This skepticism seems to be selectively applied: the relation between congruence and classification errors becomes stronger when AI is introduced explicitly. The analysis of detector responsiveness points to non-selective partial mistrust of AI signals. Overall, AI-based verification does not mechanically correct ideologically structured classification errors.

Keywords: Misinformation; Artificial Intelligence; Ideological Bias; Political Polarization; Experimental Economics

1 Introduction

Artificial intelligence is transforming the information environment by simultaneously enabling content creation and automated verification. Generative models have reduced the cost of producing fluent and plausible false content, potentially increasing the scale and persuasiveness of misinformation. At the same time, AI systems are increasingly used to assess the credibility of claims, flag misleading content, and provide evaluations of news veracity (Pilati and Venturini, 2025). AI therefore enters the information environment through two distinct margins: as a generator and as a detector of misinformation.

This paper examines how information processing changes across different information environments in which artificial intelligence plays distinct roles. We conduct an incentivized laboratory experiment in which participants classify true and false news items across three within-subject phases that vary the role of AI: a human-written baseline, an environment with AI-generated misinformation, and an environment with explicit AI detection signals.

We ask three main questions. First, how do classification errors and their composition vary across these environments? Second, does ideological congruence shape classification errors, and does this relationship vary with the role of AI? Third, how do individuals interpret and respond to explicit AI credibility signals?

A large literature documents that individuals are more likely to accept information consistent with their attitudes and to discount or reject information that conflicts with them (Taber and Lodge, 2006; Kahan, 2013). In the present setting, this generates two complementary forms of congruence-related error: the acceptance of ideologically congruent false news and the rejection of ideologically incongruent true news. Such patterns are often interpreted through motivated reasoning, but they need not uniquely reflect a motivation to defend preferred beliefs. They may also arise from prior-dependent interpretation of ambiguous evidence, confirmatory misperception, asymmetric attention, or differential scrutiny (Pennycook and Rand, 2019).

These patterns can be understood through models of belief-dependent interpretation of

information. Individuals may interpret ambiguous evidence in ways consistent with their prior beliefs, effectively updating as if they had received a different signal than the one observed (Rabin and Schrag, 1999; Fryer et al., 2019). We build on this mechanism to model how participants form veracity beliefs both from news content and from AI detection signals.

These classification patterns may also represent a first stage of political polarization. If individuals with different ideological priors assign different levels of credibility to the same information, common exposure need not produce convergence. Existing evidence on selective exposure and partisan information processing supports this possibility (Stroud, 2010; Gentzkow and Shapiro, 2011; Iyengar et al., 2012). This paper focuses on the first stage of this mechanism.

This article relates to four strands of literature. The first examines the production, diffusion, and political consequences of misinformation (Allcott and Gentzkow, 2017; Grinberg et al., 2019) and individual heterogeneity in truth discernment (Pennycook and Rand, 2019; Nyhan, 2021). The second investigates how prior beliefs shape information processing and formalizes these patterns in models of belief-dependent interpretation of ambiguous evidence (Taber and Lodge, 2006; Kahan, 2013; Rabin and Schrag, 1999; Fryer et al., 2019). The third studies the effectiveness of corrective interventions such as fact-checking (Nyhan and Reifler, 2010a; Wood and Porter, 2019). The fourth, more recent strand examines the implications of generative artificial intelligence for misinformation production, verification tools, and responsiveness to algorithmic advice (Park, 2025; Kuznetsova et al., 2025; Dietvorst et al., 2015; Greevink et al., 2024).

These literatures have largely examined misinformation production, politically structured information processing, and corrective interventions separately. Less is known about how ideological classification errors evolve when AI operates on both sides of the information environment — first as a producer of misinformation and then as an explicit source of verification — and how individuals respond to AI signals under different degrees of trust or mistrust.

These questions are addressed through an incentivized laboratory experiment with 76 participants each completing 30 classification decisions across three within-subject phases: a human baseline, an AI-generated misinformation phase, and an explicit AI-detection phase. Prior to the task, we elicited participants' positions on four policy domains and coded each news item by its ideological leaning, allowing us to construct a continuous congruence-bias index.

Three main results emerge. First, average classification accuracy remains broadly stable across environments, but the composition of errors shifts in the AI-detection phase, with an increase in false negatives consistent with greater skepticism. Second, classification errors are strongly predicted by ideological congruence, and this relationship becomes stronger when AI is introduced explicitly. Third, participants do not incorporate the AI detector uniformly. The evidence is most consistent with non-selective partial mistrust: weak detector signals are disregarded, independently from their congruence with participant beliefs.

This paper makes three contributions. First, it studies artificial intelligence as a structural feature of the information environment by distinguishing its effects as a generator of misinformation from its effects as a provider of verification signals within the same design. Second, it provides experimental evidence on how ideological congruence shapes classification errors and shows that this relationship varies with the role of AI. Third, it develops and tests a theoretical framework that distinguishes different regimes of trust and mistrust toward AI signals, linking observed behavior to models of belief-dependent information processing.

The remainder of the paper proceeds as follows. Section 2 describes the experimental design. Section 3 defines the main variables and empirical strategy. Section 4 develops the theoretical framework. Section 5 presents the results. Section 6 discusses the implications and limitations.

2 Experimental Design

2.1 Experimental Structure and Classification Task

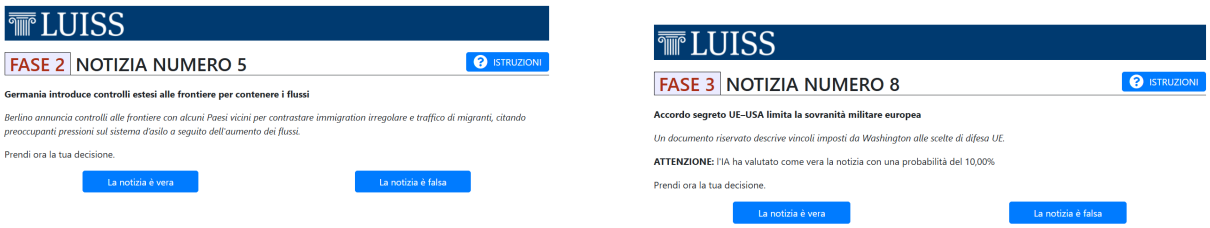
The experiment is designed to study how individuals classify information as true or false across environments in which both the production of misinformation and the availability of AI-based verification vary. Participants complete an incentivized classification task in which they evaluate a sequence of short news items. Each item consists of a headline and a brief description, and participants classify it as either true or false.

The experiment follows a within-subject design with three consecutive phases. In Phase 1, participants evaluate human-written news items, including both true and false content. This phase provides a baseline information environment in which artificial intelligence plays no explicit role. In Phase 2, false items are generated using artificial intelligence, but participants are not informed about their origin. The interface and the decision task remain identical to Phase 1. In Phase 3, AI becomes an explicit feature of the information environment: each news item is accompanied by an AI-generated probability that the item is true, after which participants make the same binary classification decision.

All participants complete the phases in the same order. This sequence preserves the non-salience of AI in Phase 2. Introducing the detector before the AI-generation phase would have made participants aware of the presence of AI in the experiment and could have altered how they evaluated subsequent news content.

In each phase, participants classify exactly 10 news items. Each phase draws from a fixed pool of 16 news items, with a balanced ideological composition, consisting of 10 true and 6 false items. For each participant and phase, the program randomly selects without replacement 7 true items and 3 false items. This design ensures that every participant faces the same truth-state composition (7 true and 3 false items) while the specific items and their presentation order vary across participants. After submitting a choice, they proceed to the following item without receiving feedback about the accuracy of their classification.

Panels (a) and (b) of Figure 1 compare the classification interface used before and after the explicit introduction of AI verification.



(a) Interface without AI verification (b) Interface with AI verification

Notes: Panel (a): Phases 1–2. Panel (b): Phase 3 with AI probability signal.

2.2 Political-Belief Elicitation and Ideological Direction

Participants report their general political orientation and their positions on four policy domains. Each response is elicited on an 11-point scale ranging from 0 to 10. The general political-orientation question ranges from "Extreme left" to "Extreme right." The issue-specific questions concern foreign policy and EU–NATO involvement, immigration, economic policy, and LGBT rights and gender-identity policy.

Table 1 reports the substantive endpoints presented in the experimental questionnaire. The issue-specific measures, rather than the general left–right scale, are used to match each participant's political position to the corresponding topic of each news item.

Let $R_{ik} \in \{0, \ldots, 10\}$ denote participant i's response in policy domain k. The response is normalized as

$$B_{ik} = \frac{R_{ik} - 5}{5}, \tag{1}$$

where $B_{ik} \in [-1, 1]$. Negative values represent positions closer to the left or progressive endpoint of the relevant scale, whereas positive values represent positions closer to the right or conservative endpoint.

Each news item is assigned to one of the four policy domains and coded according to the

Table 1: Political-belief scales

Political domain	Endpoint 0	Endpoint 10
General political orientation	Extreme left	Extreme right
Foreign policy / EU–NATO	Strongly favorable to EU/NATO involvement	Strongly opposed to EU/NATO involvement
Immigration policy	Immigration policy should be more open	Immigration should be much more restricted
Economic policy	Greater state intervention and redistribution	More market-oriented policy with less state intervention
LGBT rights / gender-identity policy	More progressive policies	More conservative policies

Notes: Participants select an integer from 0 to 10. The endpoint descriptions reproduce the substantive alternatives presented in the experimental questionnaire.

position of the scale that would be supported if the claim were true. In the original dataset,

$$NewsSide_j = \begin{cases} 1, & \text{if item } j \text{ supports beliefs associated with lower values in the corresponding topic beli} \\ 0, & \text{if item } j \text{ supports beliefs associated with higher values in the corresponding topic bel} \end{cases} \tag{2}$$

For the congruence analysis, this indicator is transformed into

$$L_j = 1 - 2NewsSide_j, \tag{3}$$

so that $L_j = -1$ for items supporting the left/progressive endpoint and $L_j = +1$ for items supporting the right/conservative endpoint. This coding aligns the direction of the news with the normalization of participants' political positions.

The coding concerns the political implication of believing the claim, rather than the positive or negative tone of its wording. For example, the news item in Figure 1a highlights excessive migration pressures and therefore provides support for more restrictive immigration policies. This supports the conservative endpoint of the immigration scale and is coded as

$L_j = +1$. Likewise, the (false) news item in Figure 1b criticizes U.S. and NATO involvement in Europe. This narrative supports opposition to EU–NATO involvement and is therefore coded as $L_j = +1$ on the foreign policy scale.

The resulting participant-item match is used to construct ideological congruence and the congruence-bias index. Their formal definitions are provided in Section 3.

2.3 News Dataset

The experimental news dataset covers four politically salient domains: foreign policy and EU-NATO involvement, immigration, economic policy, and LGBT rights and gender-identity policy. Each news item consists of a headline and a brief description and is classified according to its objective veracity, policy topic, ideological direction, and production method.

The true items are based on verifiable factual claims and are accompanied in the news dataset by links to the corresponding sources. Human-written false items consist of claims that had circulated publicly and were subsequently debunked; the dataset similarly reports links documenting their circulation and falsity. This procedure distinguishes false content that had appeared in the real information environment from false items generated specifically for the experiment using artificial intelligence.

All items are presented in a standardized headline-plus-description format. Source names, hyperlinks, logos, and other visual indicators of origin are not displayed during the classification task. Participants therefore base their decisions on the textual content and, in Phase 3, on the additional AI-verification signal.

The complete item-level dataset is made available together with the experimental data. It reports the headline and description of every news item, the relevant source or debunking link, objective veracity, policy topic, directional coding, AI-generation status, phase assignment, and Phase 3 detector probability.

2.4 AI-Generated Misinformation and AI-Based Verification Signal

In Phases 2 and 3, false content is generated using ChatGPT from the OpenAI GPT-4 family. The generation procedure begins by providing the model with the definitions of the four issue-specific political scales and the two substantive endpoints associated with each domain. The model is also instructed to follow the standardized structure used in the experiment, consisting of a headline and a brief description.

For each AI-generated item, the prompt asks the model to create a false news item concerning a specified policy domain and supporting one of the two endpoints of the corresponding political scale. The general structure of the request is:

> *Create a false news item on [policy domain], consisting of a headline and a brief description, whose content supports [specified endpoint of the issue-specific scale].*

The prompt therefore determines the topic and intended political direction of the item while allowing the model to generate its specific content and wording. The resulting news items are incorporated into the experimental pools using the same headline-plus-description format as the human-sourced items. Participants are not informed that false items are AI-generated and receive no source information that would reveal their production method.

Phase 3 additionally introduces an AI-based verification signal generated using Grok, developed by xAI. Grok is deployed on the X platform as an AI assistant with access to publicly available online information and web-search functions, and it is increasingly used in social-media interactions to evaluate contested factual claims (Mei et al., 2026). Using different systems for generation and verification separates the two technological margins and avoids having the same model evaluate its own output.

For each Phase 3 item, the detector provides a numerical assessment of the probability that the news is true. Let

$$A_j \in [0, 100] \tag{4}$$

denote the probability assigned by the detector to item j. The interface displays the following information:

The AI evaluated this news item as true with a probability of A_j percent.

The probability concerns the veracity of the news item, not the probability that it was AI-generated. Participants are informed only that the assessment is provided by an AI system. They are not told that the underlying system is Grok and receive no information about its expected accuracy, technical characteristics, or relationship to the model used to generate false content.

The detector value is fixed at the item level, so all participants who evaluate the same item observe the same probability. The realized detector outputs range from 5 to 95 percent. The empirical analysis transforms these raw probabilities into a measure of confidence in the objectively correct state and into indicators identifying intermediate or weak outputs. These variables are formally defined in Section 3.

2.5 Procedures, Incentives, and Ethical Approval

The experiment was programmed and administered using oTree. Participants were recruited through ORSEE from the subject pool of the CESARE Laboratory at LUISS Guido Carli University, and the study was conducted at the CESARE Lab in Rome.

The final sample contains 76 participants. Each participant could provide up to 30 classification decisions, corresponding to 10 items in each phase. Three Phase 3 observations are missing because the participants concerned did not submit a classification for those items. The final dataset therefore contains 2,277 participant-item-phase decisions: 760 in Phase 1, 760 in Phase 2, and 757 in Phase 3.

Participants receive a fixed participation payment of EUR 3 and a performance-based bonus. At the end of the experiment, one decision is randomly selected from each phase. Participants receive EUR 5 for each selected news item that they classified correctly and zero otherwise. Total earnings can therefore range from EUR 3 to EUR 18. Average payment in the experiment was EUR 15.20.

The complete experimental instructions are reported in the Appendix. The experimental data, news corpus and replication materials are publicly available in the data repository linked in the Data Availability Section.

The experimental protocol received ethical approval from the LUISS Research Committee. Participation was voluntary and based on informed consent. Participants were informed about compensation, data processing, and their right to withdraw without penalty. Responses were anonymized and used exclusively for research purposes. A debriefing at the end of the experiment clarified the presence of misinformation and the purpose of the study.

3 Methodology

The empirical analysis is organized into three blocks. First, we compare classification performance across the three information environments, considering both average accuracy and the composition of classification errors. Second, we examine whether ideological congruence predicts misclassification and whether this relationship changes across phases. Third, we use the Phase 3 detector outputs to assess how participants respond to AI-generated veracity signals and which interpretation regime is most consistent with the observed behavior.

Because all participants complete the phases in the same order and the news pools are phase-specific, the estimates are interpreted as within-participant phase contrasts rather than as fully isolated treatment effects. Participant fixed effects absorb time-invariant individual heterogeneity, while standard errors are clustered at the participant level to account for correlation among repeated decisions made by the same individual.

3.1 Outcomes and Phase Comparisons

The unit of observation is the participant-item-phase decision. Let i index participants, j news items, and $t \in \{1, 2, 3\}$ experimental phases. Let

$$Y_{ijt} \in \{0, 1\}$$

denote participant i's classification of item j in phase t, where $Y_{ijt} = 1$ if the item is classified as true and $Y_{ijt} = 0$ if it is classified as false. Objective veracity is denoted by

$$D_{jt} \in \{0, 1\},$$

where $D_{jt} = 1$ for true items and $D_{jt} = 0$ for false items. We also define

$$\text{Fake}_{jt} = 1 - D_{jt}.$$

Classification accuracy and error are defined as

$$\text{Accuracy}_{ijt} = \mathbf{1}\{Y_{ijt} = D_{jt}\}, \tag{5}$$

and

$$\text{Error}_{ijt} = 1 - \text{Accuracy}_{ijt}. \tag{6}$$

To study the composition of classification errors, we define the false-positive and false-negative components as

$$\text{FalsePositive}_{ijt} = \mathbf{1}\{D_{jt} = 0,\ Y_{ijt} = 1\}, \tag{7}$$

and

$$\text{FalseNegative}_{ijt} = \mathbf{1}\{D_{jt} = 1,\ Y_{ijt} = 0\}. \tag{8}$$

These are unconditional indicators defined over all classification decisions. Their sample means therefore measure the shares of all decisions that result in false-positive and false-negative errors, respectively, and satisfy

$$\text{Error}_{ijt} = \text{FalsePositive}_{ijt} + \text{FalseNegative}_{ijt}.$$

For each outcome $Z_{ijt} \in \{\text{Accuracy}, \text{Error}, \text{FalsePositive}, \text{FalseNegative}\}$, we estimate pairwise phase comparisons using

$$Z_{ijt} = \alpha + \tau^{ab}\mathbf{1}\{t = b\} + \mu_i + \varepsilon_{ijt}, \qquad t \in \{a, b\}, \tag{9}$$

where μ_i denotes participant fixed effects. We consider three contrasts:

$$(a, b) \in \{(1, 2), (2, 3), (1, 3)\}.$$

The Phase 2–Phase 1 contrast describes the change associated with moving from the human-written baseline to the environment containing undisclosed AI-generated misinformation. The Phase 3–Phase 2 contrast describes the change associated with making AI explicit through the detector. The Phase 3–Phase 1 contrast summarizes the overall difference between the baseline and the environment in which AI operates both as a generator and as a provider of verification signals.

3.2 Ideological Congruence and Classification Error

Let $k(j)$ denote the policy domain associated with item j. As described in Section 2, let

$$B_{i,k(j)} \in [-1, 1]$$

denote participant i's normalized position on the relevant policy issue, and let

$$L_{jt} \in \{-1, +1\}$$

denote the ideological direction of the news item. Ideological congruence is defined as

$$\text{Congruence}_{ijt} = B_{i,k(j)} L_{jt}. \tag{10}$$

Positive values indicate that the political implications of believing the news are aligned with the participant's issue-specific position, whereas negative values indicate ideological incongruence.

Congruence can affect classification differently depending on objective veracity. Congruence pushes participants toward error when a congruent false item is accepted as true, while incongruence pushes them toward error when a true item is rejected as false. To capture both cases in a single continuous measure, we define

$$\text{CongruenceBias}_{ijt} = \text{Congruence}_{ijt} \left(2\text{Fake}_{jt} - 1\right). \tag{11}$$

For false news, $\text{CongruenceBias}_{ijt} = \text{Congruence}_{ijt}$, so the index is positive for congruent false items. For true news, $\text{CongruenceBias}_{ijt} = -\text{Congruence}_{ijt}$, so the index is positive for incongruent true items. Higher values therefore indicate that the combination of political position, news direction, and objective veracity creates stronger pressure toward misclassification.

We first estimate the average relationship between congruence bias and classification error using

$$\text{Error}_{ijt} = \alpha + \theta_2 \text{Ph2}_t + \theta_3 \text{Ph3}_t + \delta \text{CongruenceBias}_{ijt} + \mu_i + \varepsilon_{ijt}, \tag{12}$$

where Phase 1 is the omitted category. The coefficient δ captures the average association

between congruence bias and classification error across the three phases.

We then allow this relationship to vary across information environments:

$$\begin{aligned}\text{Error}_{ijt} = \alpha + \theta_2 \text{Ph2}_t + \theta_3 \text{Ph3}_t + \delta \text{CongruenceBias}_{ijt} \\ + \lambda_2 \left(\text{CongruenceBias}_{ijt} \times \text{Ph2}_t\right) + \lambda_3 \left(\text{CongruenceBias}_{ijt} \times \text{Ph3}_t\right) + \mu_i + \varepsilon_{ijt}.\end{aligned} \tag{13}$$

In this specification, δ measures the relationship between congruence bias and error in Phase 1. The coefficients λ_2 and λ_3 measure how this relationship changes in the AI-misinformation and AI-detection environments, respectively.

3.3 Responsiveness to the AI Verification Signal

The Phase 3 analysis examines how classification errors vary with the strength and informativeness of the AI detector output. Let

$$A_j \in [0, 100]$$

denote the probability displayed to participants that item j is true. Because a high value is corrective for true news while a low value is corrective for false news, we first transform the raw detector probability into confidence in the objectively correct state:

$$\text{CorrectiveSignal}_j^{\text{raw}} = D_{j3} A_j + (1 - D_{j3})(100 - A_j). \tag{14}$$

Thus,

$$\text{CorrectiveSignal}_j^{\text{raw}} = \begin{cases} A_j, & \text{if } D_{j3} = 1, \\ 100 - A_j, & \text{if } D_{j3} = 0. \end{cases}$$

Higher values always represent a more directional detector signal in favor of the objectively correct classification. For estimation, the variable is centered at its Phase 3 sample mean and divided by ten:

$$\text{CorrectiveSignal}_j = \frac{\text{CorrectiveSignal}_j^{\text{raw}} - \overline{\text{CorrectiveSignal}^{\text{raw}}}}{10}. \tag{15}$$

The coefficient can therefore be interpreted as the change in error associated with a ten-percentage-point increase in detector confidence in the correct state, evaluated around the sample mean.

We estimate

$$\begin{aligned}\text{Error}_{ij3} = \gamma_0 + \gamma_1 \text{CorrectiveSignal}_j + \gamma_2 \text{CongruenceBias}_{ij3} \\ + \gamma_3 \left(\text{CorrectiveSignal}_j \times \text{CongruenceBias}_{ij3}\right) + \mu_i + \varepsilon_{ij3}.\end{aligned} \tag{16}$$

A negative γ_1 would indicate that stronger corrective signals are associated with fewer classification errors. With error as the dependent variable, a positive γ_3 indicates that this error-reducing relationship becomes weaker as congruence bias increases. Because the detector probability is fixed at the item level rather than experimentally randomized, these coefficients are interpreted as associations between detector strength and classification behavior, not as isolated causal effects of signal intensity.

We complement the continuous specification with a weak-signal analysis. The benchmark indicator is defined directly from the raw probability displayed to participants:

$$\text{WeakSignal}_j^{35\text{–}65} = \mathbf{1}\{35 \leq A_j \leq 65\}. \tag{17}$$

This indicator captures detector outputs that are close to the midpoint of the probability scale and therefore provide relatively weak directional evidence in favor of either a true or a

false classification. We estimate

$$\begin{aligned}\text{Error}_{ij3} = \beta_0 + \beta_1 \text{WeakSignal}_j + \beta_2 \text{CongruenceBias}_{ij3} \\ + \beta_3 \left(\text{WeakSignal}_j \times \text{CongruenceBias}_{ij3}\right) + \mu_i + \varepsilon_{ij3}. \end{aligned} \tag{18}$$

A positive β_1 is consistent with weak AI outputs being less effective than strongly directional outputs. Under non-selective partial mistrust, weak signals are discounted independently of ideological congruence, so no additional sign restriction is imposed on β_3. Selective partial mistrust instead predicts $\beta_3 > 0$, because weak corrective outputs should be particularly ineffective when ideological congruence favors the erroneous classification.

As a broader sensitivity analysis, reported in the Appendix, we define

$$\text{WeakSignal}_j^{20\text{–}80} = \mathbf{1}\{20 \leq A_j \leq 80\}. \tag{19}$$

This alternative definition includes a wider range of intermediate detector probabilities and is used to assess whether the results depend on the narrower benchmark classification of weak signals.

3.4 Estimation and Robustness

All main specifications are estimated as linear probability models with participant fixed effects. Standard errors are clustered at the participant level throughout.

The robustness analysis augments the congruence-bias and Phase 3 signal models with observable item characteristics. The full set of item controls includes objective veracity, AI-generation status, news direction, and topic indicators for foreign policy, immigration, and LGBT rights, with economic policy as the omitted topic category.

For the phase-interaction model, we additionally control for the linear and quadratic position of each item within its phase. This evaluates whether the Phase 3 interaction is

related to learning, fatigue, or other systematic within-phase order effects. Participant-cluster bootstrap inference is also reported for the main congruence-bias coefficient and its Phase 3 interaction.

The empirical comparison of detector regimes is based on the joint pattern of the continuous and weak-signal specifications. These regressions do not uniquely identify trust or mistrust as psychological mechanisms. Rather, they indicate which theoretical interpretation regime is most consistent with the observed relationship between detector outputs, ideological congruence, and classification error.

4 Theoretical Framework

This section develops a basic theoretical framework to explain how ideological congruence may shape news-classification errors and how individuals interpret the AI-based detection signal. The framework distinguishes between two conceptually separate objects. Individuals hold prior beliefs about an underlying political or social state of the world, while separately forming a subjective belief about whether a particular news item is true. Political beliefs may therefore affect the perceived credibility of news.

The framework builds on the literature on belief-dependent interpretation of information. In Fryer et al. (2019), ambiguous evidence may be interpreted differently depending on prior beliefs, so that individuals exposed to the same information can sometimes extract different effective signals and then update rationally conditional on those interpretations. A related microfoundation is provided by Rabin and Schrag (1999), where evidence contradicting a favored hypothesis may be misperceived as supporting it. In both cases, the distortion occurs before updating: individuals may update consistently given the signal they perceive, even though that perceived signal differs from the objective evidence.

Belief-dependent classification need not, however, arise exclusively from a motivation to defend preferred political beliefs. The same pattern may result from asymmetric attention,

differential scrutiny, lower cognitive effort devoted to congruent claims, or different weights assigned to confirming and disconfirming evidence (Kunda, 1990; Taber and Lodge, 2006; Kahan, 2013; Pennycook and Rand, 2019). The interpretation mechanism introduced below should therefore be understood as a reduced-form representation of several possible cognitive foundations. The framework uses this mechanism to study how political beliefs affect the perceived veracity of news, how an AI detector provides a second signal, how mistrust in the information environment can potentially alter prior credibility or the threshold for classifying a news item as true, and how congruence-dependent assessments of veracity may ultimately contribute to greater dispersion in political beliefs and thus higher polarization.

4.1 Political Beliefs, News Veracity, and Ideological Congruence

Let the underlying political or social state be

$$\theta \in \{A, B\},$$

and let

$$\pi_i = P_i(\theta = A) \tag{20}$$

denote participant i's prior political belief. Accordingly, $1 - \pi_i = P_i(\theta = B)$. The belief π_i concerns the underlying state of the world and is conceptually distinct from the participant's belief about whether a particular news item is true.

Each news item j has objective veracity

$$D_j \in \{0, 1\},$$

where $D_j = 1$ denotes a true item and $D_j = 0$ denotes a false item. Before processing its

content, participant i assigns the item a prior probability of being true:

$$q_{ij}^{(0)} = P_i(D_j = 1). \tag{21}$$

After observing and interpreting the content of the news, this prior is updated to

$$q_{ij}^{(1)} = P_i(D_j = 1 \mid \widetilde{s}_{ij}), \tag{22}$$

where $\widetilde{s}_{ij}$ denotes the subjectively interpreted signal contained in the news. The interpretation mechanism generating $\widetilde{s}_{ij}$ is specified in the following subsection.

A news item also has a political direction

$$L_j \in \{-1, +1\}.$$

We set $L_j = +1$ when the implications of the news, if the item is true, provide evidence in favor of state A, and $L_j = -1$ when they provide evidence in favor of state B.

Once the participant has formed the posterior veracity belief $q_{ij}^{(1)}$, the perceived credibility of the news may update the political belief:

$$\pi_{ij}^{(1)} = P_i\Big(\theta = A \mid q_{ij}^{(1)}, L_j\Big). \tag{23}$$

The direction of this updating depends on the political implications of the news. In particular,

$$\frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} > 0 \qquad \text{if } L_j = +1, \tag{24}$$

whereas

$$\frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} < 0 \qquad \text{if } L_j = -1. \tag{25}$$

Thus, an increase in the perceived probability that an A-oriented item is true raises belief in state A, while an increase in the perceived probability that a B-oriented item is true lowers

belief in state A and correspondingly raises belief in state B.

As defined in the empirical analysis, let $B_i \in [-1, 1]$ denote participant i's normalized political position, with $B_i > 0$ indicating greater support for state A and $B_i < 0$ indicating greater support for state B. Ideological congruence between participant i and news item j is

$$C_{ij} = B_i L_j. \tag{26}$$

The item is ideologically congruent when $C_{ij} > 0$ and incongruent when $C_{ij} < 0$.

This definition has a direct belief-updating interpretation. A news item is congruent when an increase in its perceived veracity moves the political posterior in the direction of the participant's initial political position. Formally,

$$B_i \frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} > 0 \qquad \Longleftrightarrow \qquad C_{ij} > 0. \tag{27}$$

Conversely, a news item is incongruent when greater perceived veracity moves the political posterior against the participant's initial position:

$$B_i \frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} < 0 \qquad \Longleftrightarrow \qquad C_{ij} < 0. \tag{28}$$

For example, when $B_i > 0$, participant i initially favors state A. An A-oriented news item is then congruent, and

$$\frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} > 0 :$$

the more credible the news is perceived to be, the stronger the participant's belief in state A. A B-oriented item is instead incongruent, and

$$\frac{\partial \pi_{ij}^{(1)}}{\partial q_{ij}^{(1)}} < 0 :$$

greater perceived credibility weakens the participant's initial belief. The relationships are

reversed for participants with $B_i < 0$.

Congruence therefore concerns the political implications of believing the news, not its objective veracity. A false item may be congruent with the participant's political position, while a true item may be incongruent.

The content of news item j provides a signal about its objective veracity. Adapting the signal structure of Fryer et al. (2019) to news classification, let

$$s_j \in \{s^T, s^F, s^U\},$$

where s^T is evidence supporting the hypothesis that the news is true, s^F is evidence supporting the hypothesis that it is false, and s^U is evidence that is open to interpretation.

The participant does not update directly on the raw signal s_j, but on an interpreted signal

$$\widetilde{s}_{ij} = J(s_j, C_{ij}). \tag{29}$$

Following the benchmark interpretation rule in Fryer et al. (2019), unambiguous signals retain their objective meaning:

$$J(s^T, C_{ij}) = s^T, \qquad J(s^F, C_{ij}) = s^F. \tag{30}$$

Ambiguous evidence is instead resolved in the direction consistent with the participant's existing political belief:

$$J(s^U, C_{ij}) = \begin{cases} s^T, & \text{if } C_{ij} > 0, \\ s^F, & \text{if } C_{ij} < 0. \end{cases} \tag{31}$$

Thus, when a news item is politically congruent, ambiguous elements of its content are interpreted as evidence that the item is true. When the same content is politically incongruent, ambiguous elements are interpreted as evidence that the item is false. The individual may therefore update consistently given the signal that is subjectively perceived,

even though the interpretation of that signal depends on the initial political belief.

Because s^T supports truth and s^F supports falsity,

$$P_i(D_j = 1 \mid s^T) > q_{ij}^{(0)} > P_i(D_j = 1 \mid s^F). \tag{32}$$

Combining Equations (31) and (32) yields

$$C_{ij} > 0, \quad s_j = s^U \quad \Longrightarrow \quad q_{ij}^{(1)} = P_i(D_j = 1 \mid s^T) > q_{ij}^{(0)}, \tag{33}$$

whereas

$$C_{ij} < 0, \quad s_j = s^U \quad \Longrightarrow \quad q_{ij}^{(1)} = P_i(D_j = 1 \mid s^F) < q_{ij}^{(0)}. \tag{34}$$

The Fryer-type interpretation mechanism therefore produces the directional relationship

$$\frac{\partial q_{ij}^{(1)}}{\partial C_{ij}} > 0 \tag{35}$$

for evidence that is open to interpretation. Greater political congruence raises the perceived probability that the news is true, whereas greater incongruence lowers it.

The participant classifies the news item as true whenever the posterior veracity belief exceeds an individual threshold:

$$Y_{ij} = 1 \quad \Longleftrightarrow \quad q_{ij}^{(1)} \geq \tau_i. \tag{36}$$

The congruence-dependent distortion in Equation (35) therefore generates two complementary forms of classification error. For a false but congruent item,

$$D_j = 0, \quad C_{ij} > 0 \quad \Longrightarrow \quad P(Y_{ij} = 1) \uparrow, \tag{37}$$

so that the probability of a false positive increases. For a true but incongruent item,

$$D_j = 1, \quad C_{ij} < 0 \quad \Longrightarrow \quad P(Y_{ij} = 0) \uparrow, \tag{38}$$

so that the probability of a false negative increases.

Using the congruence-bias measure introduced in the empirical section,

$$CB_{ij} = C_{ij}(2Fake_j - 1), \tag{39}$$

these two cases can be summarized by the central prediction

$$\frac{\partial P(Error_{ij} = 1)}{\partial CB_{ij}} > 0. \tag{40}$$

Congruence bias therefore captures the extent to which political alignment pushes perceived veracity, and hence classification, in the direction favored by the participant's current political belief.

4.2 AI Detection as a Second Signal

In Phase 3, participant i receives an additional signal about the veracity of news item j from an AI detector. Let

$$a_j \in [0, 1]$$

denote the probability reported by the detector that the news item is true. The AI signal is observed after the participant has processed the news content and formed the intermediate veracity belief

$$q_{ij}^{(1)} = P_i(D_j = 1 \mid \widetilde{s}_{ij}) \, .$$

The detector provides a second source of information about the same latent state D_j.

After observing and interpreting the AI output, the participant forms the final veracity belief

$$q_{ij}^{(2)} = P_i(D_j = 1 \mid \widetilde{s}_{ij}, \widetilde{a}_{ij}) \,, \tag{41}$$

where

$$\widetilde{a}_{ij} = J(a_j) \tag{42}$$

denotes the informational content of the detector output that is subjectively used in the second update. After processing the AI signal, participant i forms the final veracity belief $q_{ij}^{(2)}$. Classification follows the threshold rule

$$Y_{ij} = 1 \quad \Longleftrightarrow \quad q_{ij}^{(2)} \geq \tau_i, \tag{43}$$

where $Y_{ij} = 1$ denotes a classification of the news item as true. This preserves the same general architecture introduced for news content:

$$\text{raw signal} \longrightarrow \text{interpreted signal} \longrightarrow \text{Bayesian posterior}.$$

In both stages, participants update on the signal that they perceive as informative rather than mechanically on the raw information supplied to them. The substantive meaning of ambiguity, however, differs across the two sources. Textual content may be open to competing interpretations, so that an ambiguous content signal can be resolved in the direction of the participant's political belief. By contrast, the detector reports an explicit numerical probability. Ambiguity in this second stage therefore concerns the perceived reliability or informativeness of the AI source, rather than the semantic interpretation of the numerical value itself.

Four main cases are considered.

Full trust. The participant regards the reported probability as valid and informative. The signal is incorporated according to its numerical value:

$$J(a_j) = a_j. \tag{44}$$

The final belief is therefore

$$q_{ij}^{(2)} = P_i\Big(D_j = 1 \mid q_{ij}^{(1)}, a_j\Big). \tag{45}$$

If the detector provides evidence in favor of truth, then

$$q_{ij}^{(2)} > q_{ij}^{(1)}, \tag{46}$$

whereas evidence in favor of falsity implies

$$q_{ij}^{(2)} < q_{ij}^{(1)}. \tag{47}$$

Under full trust, the numerical strength of the corrective signal should therefore affect classification.

Partial mistrust. The participant may regard strongly directional detector outputs as informative while treating intermediate probabilities as insufficiently reliable or informative. To preserve the three-signal structure used for news content, let the detector output be mapped into

$$a_j \in \{a^T, a^F, a^U\},$$

where a^T supports the hypothesis that the news is true, a^F supports the hypothesis that it is false, and a^U denotes an intermediate output treated as ambiguous.

For two cutoffs satisfying

$$0 < \underline{a} < \frac{1}{2} < \overline{a} < 1,$$

the perceived signal category is

$$a_j = \begin{cases} a^F, & a_j < \underline{a}, \\ a^U, & \underline{a} \le a_j \le \overline{a}, \\ a^T, & a_j > \overline{a}. \end{cases} \tag{48}$$

Signals perceived as sufficiently directional retain their informational content:

$$J(a^T, C_{ij}) = a^T, \qquad J(a^F, C_{ij}) = a^F. \tag{49}$$

Let a^0 denote a null signal that provides no additional usable information and therefore leaves the intermediate veracity belief unchanged:

$$P_i\Big(D_j = 1 \mid q_{ij}^{(1)}, a^0\Big) = q_{ij}^{(1)}. \tag{50}$$

We further distinguish between two forms of partial mistrust.

Non-selective partial mistrust. Under non-selective partial mistrust, every intermediate detector output is treated as a null signal, independently of the political implications of the item:

$$J(a^U, C_{ij}) = a^0 \qquad \text{for all } C_{ij}. \tag{51}$$

It follows that

$$a_j = a^U \quad \Longrightarrow \quad q_{ij}^{(2)} = q_{ij}^{(1)}. \tag{52}$$

In this case, weak AI signals do not reverse or distort the participant's previous assessment. They simply fail to displace the belief already formed from the news content. The primary empirical implication is that intermediate outputs may be associated with more

classification errors than strongly directional outputs:

$$\frac{\partial P(Error_{ij} = 1)}{\partial WeakSignal_j} > 0. \tag{53}$$

Non-selective partial mistrust does not, by itself, imply that this effect varies with ideological congruence.

Selective partial mistrust. Under selective partial mistrust, the treatment of an intermediate detector output depends on whether its direction conflicts with the assessment favored by ideological congruence. A weak corrective signal that challenges the ideologically favored classification is more likely to be neutralized, whereas an equally weak signal that is consistent with that classification may still be incorporated.

Let $R_{ij} = 1$ indicate that the AI signal is corrective relative to the classification favored by congruence bias. The interpretation rule can be written as

$$J(a^U, C_{ij}, R_{ij}) = \begin{cases} a^0, & R_{ij} = 1, \\ a^U, & R_{ij} = 0. \end{cases} \tag{54}$$

Thus, when a weak AI signal conflicts with the conclusion favored by the participant's political beliefs,

$$q_{ij}^{(2)} = q_{ij}^{(1)}, \tag{55}$$

whereas a weak but congruent signal may still affect the final veracity belief.

Selective partial mistrust generates a stronger empirical prediction than its non-selective counterpart. In addition to a possible direct effect of weak signals, it predicts that weak signals are especially ineffective when congruence bias pushes the participant toward an incorrect classification:

$$\frac{\partial^2 P(Error_{ij} = 1)}{\partial WeakSignal_j \, \partial CB_{ij}} > 0. \tag{56}$$

The distinction is empirically relevant. A positive weak-signal effect without a corresponding interaction with congruence bias is consistent with general difficulty in using intermediate AI outputs. A positive interaction would instead indicate that weak signals are discounted selectively when they challenge ideologically favored conclusions.

Full mistrust. The participant does not regard the AI detector as a credible source of information. In this case, every output is assigned to the ambiguous category, irrespective of its numerical strength:

$$J(a_j) = a^0 \qquad \text{for all } a_j \in [0,1]. \tag{57}$$

It follows that

$$q_{ij}^{(2)} = q_{ij}^{(1)} \qquad \text{for all } a_j. \tag{58}$$

The detector then has no independent effect on the final veracity belief, and classification continues to reflect the belief already formed from the news content.

The regimes differ in the set of AI outputs that are transformed into the null signal a^0. Under full trust, this set is empty. Under partial mistrust, it contains only intermediate outputs perceived as insufficiently informative. Under full mistrust, it contains the entire support of the detector signal.

This formulation retains the same theoretical structure used for news content while distinguishing two forms of ambiguity. In the first stage, an ambiguous textual signal is open to interpretation and may be resolved in the direction of ideological congruence. In the second stage, an ambiguous AI signal reflects uncertainty about the reliability of the source and is therefore neutralized. In both cases, however, the participant updates Bayesianly only after the raw signal has been transformed into the information that is subjectively regarded as usable.

The empirical analysis of Phase 3 uses the estimated effect of the corrective signal and its interaction with congruence-related classification bias to assess which of these regimes is most consistent with observed behavior. A separate specification examining weak detector signals

evaluates whether the evidence is more consistent with partial rather than full mistrust.

4.3 Classification Thresholds and Skepticism

After forming the final veracity belief $q_{ij}^{(2)}$, participant i classifies news item j as true whenever this belief exceeds an individual decision threshold:

$$Y_{ij} = 1 \quad \Longleftrightarrow \quad q_{ij}^{(2)} \geq \tau_i, \tag{59}$$

where $\tau_i \in (0, 1)$ represents the minimum subjective probability required for the participant to accept a claim as true.

The explicit introduction of AI verification may affect classification even when participants respond only weakly to the detector. In Phase 2, participants are not informed that some misinformation has been generated by AI. In Phase 3, by contrast, the presence of AI in the information environment becomes explicit through the detector. This change may affect participants' perceptions not only of the detector itself, but also of the broader reliability of the information environment.

Such skepticism can operate through two conceptually distinct channels. First, the explicit presence of AI may reduce the prior probability that participants assign to the veracity of a generic news item. Let $q_{ij}^{(0)}$ denote the prior belief that item j is true before its content is processed. Greater mistrust in the information environment may imply

$$q_{ij}^{(0),AI} = q_{ij}^{(0)} - \Delta q_i, \qquad \Delta q_i > 0. \tag{60}$$

A lower prior shifts the entire updating process toward the hypothesis that the news is false.

Second, explicit AI verification may increase the amount of evidence required to classify a news item as true:

$$\tau_i^{AI} = \tau_i^0 + \Delta\tau_i, \qquad \Delta\tau_i > 0. \tag{61}$$

An increase in the threshold leaves the posterior veracity belief unchanged but makes it more difficult for that belief to generate a "true" classification.

Although the two channels operate at different stages, they generate similar directional implications for observed classifications. Both

$$\Delta q_i > 0$$

and

$$\Delta \tau_i > 0$$

reduce the probability that an item is classified as true:

$$\frac{\partial P(Y_{ij} = 1)}{\partial \Delta q_i} < 0, \qquad \frac{\partial P(Y_{ij} = 1)}{\partial \Delta \tau_i} < 0. \tag{62}$$

Consequently, greater skepticism increases the probability that true news is rejected:

$$D_j = 1 \quad \Longrightarrow \quad P(Y_{ij} = 0) \uparrow . \tag{63}$$

It may simultaneously reduce the acceptance of false news. Whether average accuracy improves therefore depends on the relative change in false negatives and false positives. A sufficiently large increase in false negatives can offset, or exceed, any reduction in false positives.

These environment-level channels are conceptually distinct from the belief-dependent interpretation of the detector developed in the previous subsection. A participant may interpret the numerical AI signal as informative while nevertheless becoming more skeptical about the information environment as a whole. Conversely, a participant may leave the prior and classification threshold unchanged while treating the detector output as ambiguous or unreliable. The interpretation of the individual AI signal and the general response to the explicit presence of AI therefore need not coincide.

Skepticism may also be applied selectively. Existing political beliefs may determine which claims receive the greatest scrutiny or face the highest effective burden of proof. A reduced-form representation is

$$\tau_{ij}^{AI} = \tau_i^0 + \Delta\tau_i - \kappa C_{ij}, \qquad \kappa > 0, \tag{64}$$

where C_{ij} denotes ideological congruence. Since $C_{ij} < 0$ for incongruent news, Equation (64) implies

$$\frac{\partial \tau_{ij}^{AI}}{\partial C_{ij}} = -\kappa < 0. \tag{65}$$

Thus, incongruent news faces a higher effective classification threshold, whereas congruent news faces a lower one.

For true news, selective skepticism therefore implies

$$D_j = 1, \quad C_{ij} < 0 \quad \Longrightarrow \quad P(Y_{ij} = 0) \uparrow, \tag{66}$$

so that the increase in false negatives is expected to be stronger for politically incongruent claims. In this sense, explicit AI verification may generate greater general caution while prior political beliefs shape where that caution is applied.

The experiment does not directly elicit participants' prior beliefs about the general veracity of news, their individual classification thresholds, or their perceived trust in the detector. It therefore cannot separately identify whether the observed change in classifications is generated by a lower prior $q_{ij}^{(0)}$, a higher threshold τ_i, or a combination of the two. Nor can it establish selective skepticism as a unique mechanism. These channels should instead be interpreted as theoretical explanations consistent with the joint behavior of average accuracy, false positives, false negatives, and congruence-related classification errors across phases.

4.4 Implications for Classification Error and Polarization

The framework generates direct implications for news classification. As established above, belief-dependent interpretation of ambiguous news content implies that ideological congruence increases perceived veracity:

$$\frac{\partial q_{ij}^{(1)}}{\partial C_{ij}} > 0. \tag{67}$$

Consequently, congruence shifts classification in the direction favored by the participant's political position. For a false but politically congruent item,

$$D_j = 0, \qquad C_{ij} > 0 \quad \Longrightarrow \quad P(Y_{ij} = 1) \uparrow, \tag{68}$$

which increases the probability of a false positive. For a true but politically incongruent item,

$$D_j = 1, \qquad C_{ij} < 0 \quad \Longrightarrow \quad P(Y_{ij} = 0) \uparrow, \tag{69}$$

which increases the probability of a false negative.

Using the congruence-bias measure defined in the empirical analysis, these two cases imply

$$\frac{\partial P(Error_{ij} = 1)}{\partial CB_{ij}} > 0. \tag{70}$$

This is the central empirical prediction of the framework: classification errors should be systematically related to the interaction between the participant's political position, the political direction of the news, and its objective veracity.

The explicit introduction of AI verification may affect these errors through two distinct mechanisms. First, the AI signal may be incorporated fully, partially, or not at all, as described in Section 4.2. Second, the explicit presence of AI may alter prior credibility or the classification threshold, as discussed in Section 4.3. The first mechanism concerns the use of the item-specific detector output; the second concerns skepticism toward the broader information environment.

If AI verification is fully trusted, stronger corrective signals should reduce classification errors:

$$\frac{\partial P(Error_{ij} = 1)}{\partial CorrectiveSignal_j} < 0. \tag{71}$$

Under partial mistrust, intermediate detector outputs are treated as providing less usable information than strongly directional signals. We distinguish between two cases. Under non-selective partial mistrust, weak outputs are neutralized independently of ideological congruence, so that

$$q_{ij}^{(2)} = q_{ij}^{(1)}.$$

This regime predicts that weak signals are associated with more classification errors:

$$\frac{\partial P(Error_{ij} = 1)}{\partial WeakSignal_j} > 0,$$

but imposes no sign restriction on their interaction with congruence bias.

Under selective partial mistrust, weak corrective signals are more likely to be neutralized when congruence bias favors the incorrect classification. This stronger regime additionally predicts

$$\frac{\partial^2 P(Error_{ij} = 1)}{\partial WeakSignal_j \, \partial CB_{ij}} > 0.$$

Under full mistrust, the detector should not affect the final veracity belief:

$$q_{ij}^{(2)} = q_{ij}^{(1)}, \tag{72}$$

and therefore

$$\frac{\partial P(Error_{ij} = 1)}{\partial CorrectiveSignal_j} = 0. \tag{73}$$

Environment-level skepticism generates a different prediction. A lower prior probability that news is true or a higher classification threshold reduces the probability of a "true"

response. Hence,

$$\Delta q_i > 0 \quad \text{or} \quad \Delta \tau_i > 0 \quad \Longrightarrow \quad P(Y_{ij} = 1) \downarrow . \tag{74}$$

For true news, this implies

$$P(FalseNegative_{ij} = 1) \uparrow . \tag{75}$$

Whether average accuracy rises or falls depends on whether the increase in false negatives is offset by a reduction in false positives.

If skepticism is selectively applied according to ideological congruence, the increase in false negatives should be disproportionately concentrated among true news items that are politically incongruent with participants' prior beliefs. Evidence of a stronger congruence-related increase in false negatives during Phase 3 would therefore be consistent with selective skepticism. However, the analysis cannot distinguish whether the pattern is generated by a lower prior belief in news veracity, a higher classification threshold, greater scrutiny of incongruent claims, or another form of belief-dependent processing. Selective skepticism is therefore an interpretation of the observed error composition, but not a uniquely identified cognitive mechanism.

The framework also provides a potential link between classification and political polarization. Once participant i forms the final veracity belief $q_{ij}^{(2)}$, this belief may be used to update the underlying political belief:

$$\pi_{ij}^{(2)} = P_i\Big(\theta = A \mid q_{ij}^{(2)}, L_j\Big) . \tag{76}$$

The direction of updating satisfies

$$\frac{\partial \pi_{ij}^{(2)}}{\partial q_{ij}^{(2)}} > 0 \qquad \text{if } L_j = +1, \tag{77}$$

and

$$\frac{\partial \pi_{ij}^{(2)}}{\partial q_{ij}^{(2)}} < 0 \qquad \text{if } L_j = -1. \tag{78}$$

Consider two participants i and h with different political priors. Because the interpretation of ambiguous information depends on congruence, they may assign different veracity beliefs to the same news item:

$$q_{ij}^{(2)} \neq q_{hj}^{(2)}. \tag{79}$$

If perceived veracity determines how strongly the news updates political beliefs, the two participants may move further apart. Under sufficiently strong belief-dependent interpretation,

$$\left|\pi_{ij}^{(2)} - \pi_{hj}^{(2)}\right| > \left|\pi_i - \pi_h\right|. \tag{80}$$

Thus, common exposure to the same information need not produce belief convergence. When individuals disagree about the credibility of politically relevant news, the same information environment can instead reinforce or amplify initial disagreement.

The experiment directly observes classification decisions, but it does not re-elicit political beliefs after exposure to the news environment. It therefore identifies the first stage of this potential polarization mechanism-the congruence-dependent assessment of veracity-rather than the consequent divergence of political beliefs. A natural extension would elicit political positions both before and after the classification task, together with intermediate veracity beliefs, perceived trust in the detector, and individual classification thresholds.

5 Results

5.1 Data and Descriptive Statistics

The final dataset contains 2,277 participant-item-phase decisions from 76 participants. Each participant could contribute up to 30 classification decisions, corresponding to 10 news items in each of the three experimental phases. The dataset is nearly balanced across phases: it contains 760 observations in Phase 1, 760 observations in Phase 2, and 757 observations in Phase 3. The small imbalance is due to three missing item-level decisions in Phase 3. Since

participants are observed across the three phases, the empirical analysis exploits within-subject variation in the informational environment.

Table 2 reports descriptive statistics for the political-belief variables used to construct ideological congruence. All belief variables are normalized on a $[-1, 1]$ scale from raw responses on the original 0-10 scale for each issue. Under this transformation, zero corresponds to the midpoint of the original scale, positive values correspond to higher raw-scale values, and negative values correspond to lower raw-scale values. The average normalized left-right position is close to zero, indicating that the sample is not strongly tilted toward either side of the general ideological scale. At the same time, the standard deviation is sizeable, and the extremism index shows meaningful dispersion around the ideological midpoint. Issue-specific beliefs also display substantial variation. Immigration and economic-policy attitudes are centered and close to zero, while foreign-policy and civil-rights/LGBT attitudes are more shifted towards the progressive side. This variation is important for the empirical design because ideological congruence is constructed at the topic level: each news item is matched to the participant's position on the corresponding policy domain.

Table 2: Sample descriptives: normalized political beliefs

Measure	N	Mean	SD
Panel A: General ideology			
Left-right position	76	0.02	0.455
Extremism	76	0.374	0.256
Panel B: Issue-specific beliefs			
Foreign policy	76	-0.353	0.417
Immigration	76	0.002	0.515
Economy	76	-0.034	0.479
Civil rights/LGBT	76	-0.334	0.606

Notes: Political positions are normalized to the interval $[-1, 1]$, with zero corresponding to the midpoint of the original 0-10 scale. The extremism index is the absolute distance of the general ideological position from the midpoint. The direction of each issue-specific scale follows the endpoint coding described in the experimental design.

The ideological composition of the sample is relatively balanced. Participants are classi-

fied as left-leaning if their normalized left-right score is between -1 and -0.4, centrist if it is between -0.4 and 0.4, and right-leaning if it is between 0.4 and 1. Based on this rule, 31 participants are classified as centrist, 24 as left-leaning, and 21 as right-leaning, corresponding to 40.8 percent, 31.6 percent, and 27.6 percent of the sample, respectively.

5.2 Part I: Average Effects of AI on Classification Performance

We first examine whether artificial intelligence changes average classification performance across the three experimental phases. Phase 1 provides the human-written baseline, Phase 2 introduces AI-generated misinformation, and Phase 3 adds an AI-based detection signal. The analysis distinguishes between overall performance, measured by accuracy and error rates, and the composition of errors, measured by false positives and false negatives.

Table 3 reports descriptive rates by phase.

Table 3: Classification performance by experimental phase

Phase	Obs.	Accuracy	Error	FP	FN
P1: Baseline	760	0.818	0.182	0.086	0.097
P2: AI misinformation	760	0.838	0.162	0.079	0.083
P3: AI detection	757	0.804	0.196	0.075	0.123

Notes: The unit of observation is the participant–item–phase decision. FP denotes false positives, defined as fake news items classified as true. FN denotes false negatives, defined as true news items classified as false.

The descriptive evidence shows that average accuracy remains relatively high across all phases. The more pronounced descriptive change concerns the composition of errors. False positives remain essentially stable across phases, while false negatives increase in Phase 3. Thus, the AI detection phase is not associated with an increase in accuracy but rather with a change in the composition of errors.

Figures 2 and 3 summarize these descriptive observations.

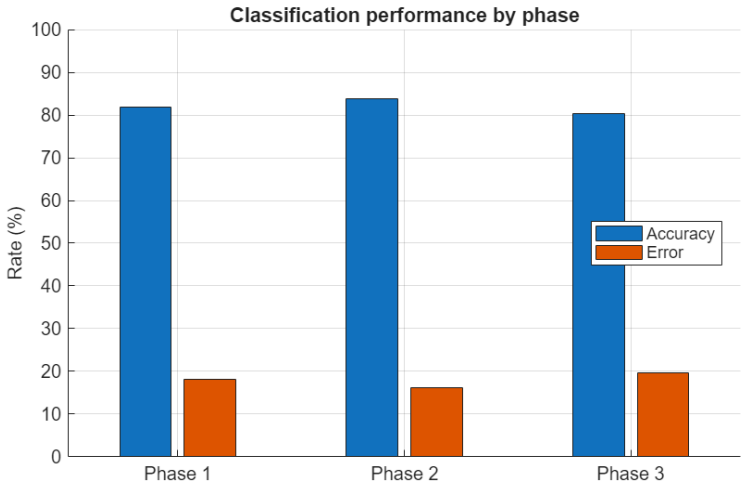


Figure 2: Accuracy and error rates by experimental phase

Notes: Bars report average rates by experimental phase.

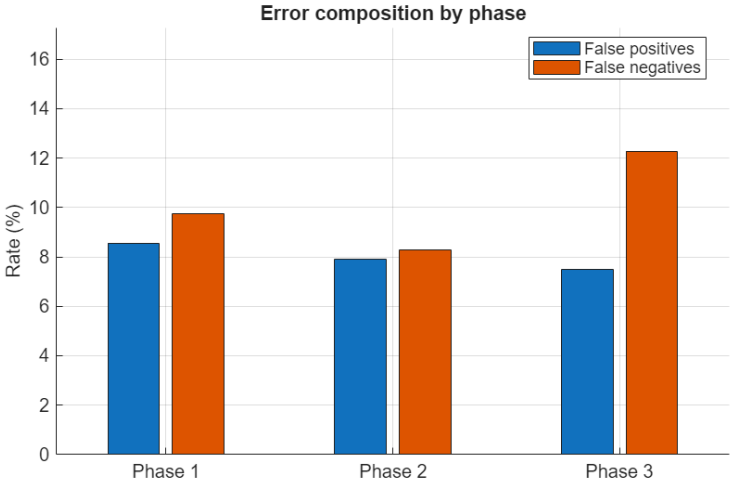


Figure 3: False positive and false negative rates by experimental phase

Notes: FP denotes false positives, defined as fake news items classified as true. FN denotes false negatives, defined as true news items classified as false.

Table 4 reports the corresponding within-subject regression comparisons. The comparison between Phase 2 and Phase 1 identifies the effect of AI-generated misinformation. The comparison between Phase 3 and Phase 2 identifies the effect of introducing the AI-based detection signal. The comparison between Phase 3 and Phase 1 reports the net effect of AI operating on both margins.

Table 4: Average effects of AI on classification performance

Outcome	Coefficient	p-value
Panel A: P2–P1, AI-generated misinformation		
Error	-0.020	0.362
	(0.022)	
Accuracy	0.020	0.362
	(0.022)	
False positive	-0.007	0.638
	(0.014)	
False negative	-0.014	0.304
	(0.014)	
Panel B: P3–P2, AI detection environment		
Error	0.035	0.188
	(0.026)	
Accuracy	-0.035	0.188
	(0.026)	
False positive	-0.005	0.773
	(0.016)	
False negative	0.041*	0.055
	(0.021)	
Panel C: P3–P1, net AI effect		
Error	0.015	0.507
	(0.023)	
Accuracy	-0.015	0.507
	(0.023)	
False positive	-0.011	0.503
	(0.017)	
False negative	0.027	0.160
	(0.019)	

Notes: Each coefficient is estimated from a separate linear probability model with participant fixed effects. Standard errors clustered at the participant level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The regression results show that neither AI-generated misinformation nor AI-based detection significantly alters average classification performance relative to the baseline. Likewise, Phase 3 does not produce a statistically significant change in average accuracy relative to Phase 2. The more informative pattern concerns the composition of errors: false positives remain essentially unchanged, whereas false negatives increase by 4.1 percentage points, with the estimate marginally significant at the ten-percent level.

This Phase 3 pattern is consistent with a skepticism-based interpretation of the explicit introduction of AI into the information environment. In Phases 1 and 2, participants classify news without being directly exposed to an external verification technology. In Phase 3, by contrast, artificial intelligence becomes salient because participants are explicitly informed that an AI system is evaluating the truthfulness of each item. This may affect not only the information available for each decision, but also participants' perceptions of the broader information environment. They may infer that misinformation is more widespread, that unaided judgment is less reliable, or that claims require greater caution before being accepted as true.

As formalized in the theoretical framework, such skepticism may operate through a lower prior belief in the general veracity of news, a higher evidentiary threshold for classifying an item as true, or both. Each channel reduces the propensity to provide a "true" response and may therefore increase false negatives. The phase-level evidence is consistent with this mechanism: the AI-detection environment is associated with more true news being rejected as false, without a corresponding reduction in false positives. However, because the change in false negatives is only marginally significant in one panel and the experiment does not directly elicit prior credibility or decision thresholds, the results should be interpreted as suggestive rather than conclusive evidence of increased skepticism.

This interpretation relates to several strands of previous research. Corrective information does not necessarily eliminate politically relevant misperceptions (Nyhan and Reifler, 2010b; Nyhan, 2020). Performance in fake-news classification may depend on analytical engage-

ment and cognitive effort, rather than exclusively on motivated reasoning (Pennycook and Rand, 2019). Moreover, research on algorithm aversion shows that individuals may reduce their reliance on algorithmic advice when algorithmic judgment is made salient or perceived as fallible (Dietvorst et al., 2015). In the present setting, making AI verification explicit may therefore increase caution toward the information environment without generating a corresponding improvement in average truth discernment.

Overall, Part I provides no evidence that AI-generated misinformation worsens average classification accuracy and no evidence that AI-based verification improves it. The clearest Phase 3 pattern is instead suggestive evidence of a shift in error composition toward false negatives relative to the AI-misinformation phase. This pattern is consistent with greater caution in accepting information as true, but it does not by itself distinguish between lower prior credibility, a higher classification threshold, reduced trust in the detector, or other changes in information processing. The net comparison between Phase 3 and the original baseline is not statistically significant, so the evidence should not be interpreted as showing that AI detection generally worsens classification performance.

5.3 Part II: Ideological Congruence and Classification Error

We next examine whether classification errors are systematically related to the alignment between participants' political beliefs and the news content. As defined in the empirical strategy, the congruence-bias index is positive when ideological congruence pushes toward classification error: when a politically congruent false item is accepted as true or when a politically incongruent true item is rejected as false.

Table 5 presents the main estimates. Column (1) measures the average association between congruence bias and classification error across the three phases, controlling for phase indicators. Column (2) allows this relationship to vary in the different information environments.

Table 5: Ideological congruence and classification errors

	Classification error	
	(1)	(2)
Phase 2	-0.017	-0.019
	(0.022)	(0.022)
	[0.441]	[0.382]
Phase 3	0.016	0.016
	(0.023)	(0.023)
	[0.489]	[0.483]
Congruence bias	0.078***	0.063**
	(0.015)	(0.029)
	[< 0.001]	[0.036]
Congruence bias × Phase 2		-0.050
		(0.043)
		[0.249]
Congruence bias × Phase 3		0.089**
		(0.038)
		[0.022]
Participant fixed effects	Yes	Yes
Observations	2,277	2,277

Notes: Linear probability models. The dependent variable equals one when the news item is incorrectly classified. Phase 1 is the omitted category. Standard errors clustered at the participant level are reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Congruence bias is a strong predictor of classification error. In the pooled specification reported in Column (1), a one-unit increase in the index is associated with a 7.8-percentage-point increase in the probability of misclassification. The estimate is statistically significant at the one-percent level. The relationship remains positive when its interactions with the experimental phases are introduced in Column (2).

This result is consistent with the central prediction of the theoretical framework. Participants do not assess the veracity of political news independently of its implications for their existing beliefs. Congruent false news is more likely to be accepted as true, whereas incongruent true news is more likely to be rejected as false. Ideological congruence therefore systematically shifts perceived veracity in the direction of the participant's current political position.

The result should not, however, be interpreted as uniquely identifying motivated reasoning. As discussed in the theoretical framework, the same classification pattern may arise from prior-dependent interpretation of ambiguous content, confirmatory signal misperception, asymmetric attention, differential scrutiny of congruent and incongruent claims, or lower cognitive effort devoted to information that appears consistent with prior beliefs. The experiment identifies the resulting congruence-related distortion in classification, but it does not separately distinguish among these cognitive foundations.

The interaction between congruence bias and Phase 2 is negative but statistically insignificant. There is therefore no evidence that the introduction of AI-generated misinformation, when the role of AI is not made explicit to participants, systematically changes the relationship between ideological congruence and classification error.

By contrast, the interaction between congruence bias and Phase 3 is positive and statistically significant, indicating that making AI verification explicit does not eliminate ideologically structured classification errors. Instead, the relationship between congruence and error becomes stronger in the environment in which participants receive an AI-based assessment of veracity. This amplification of ideologically driven errors is stable when controlling for objective veracity, AI generation, news direction, topic composition, and within-phase presentation order. Participant-cluster bootstrap inference also confirms the result. Accordingly, the stronger Phase 3 relationship is not explained by the observable composition or ordering of the news items. These robustness checks are reported in the appendix.

One possible interpretation is selective skepticism. As discussed in Part I, the introduction of the detector is associated with a shift in the composition of errors toward false negatives. If the explicit presence of AI increases general caution, while political beliefs determine where that caution is applied, participants may become especially reluctant to accept true news that conflicts with their prior views. Under this interpretation, skepticism toward the information environment is filtered through ideological congruence.

The interaction coefficient does not, by itself, uniquely identify this mechanism. A

stronger Phase 3 association between congruence and error could also reflect greater reliance on prior evaluations when the detector is distrusted, heterogeneous use of AI advice, or changes in attention and scrutiny induced by the presence of verification technology.

Part III may provide more insights by studying participants' responsiveness to the strength of the AI signal and by testing whether weaker detector outputs leave greater scope for congruence-related error.

Taken together, Part II establishes two findings. First, ideological congruence is a strong and meaningful predictor of classification error across the experiment. Second, this relationship becomes stronger when AI verification is made explicit. The first result directly supports the framework's central prediction that political beliefs shape perceived news veracity. The second shows that providing an AI-based detection signal does not neutralize congruence-related distortions and may alter the informational environment in ways that make them more consequential. Selective skepticism is a plausible interpretation consistent with these results.

5.4 Part III: AI-Signal Responsiveness and Interpretation Regimes

We finally examine how participants incorporate the signal of the AI detector. This analysis is restricted to Phase 3, in which participants observe both the news content and the detector's estimated probability that the observed news item is true. As described in the theoretical framework, the empirical patterns can be compared with four interpretation regimes: full trust, selective and non selective partial mistrust of intermediate outputs, and full mistrust of the detector.

We first study responsiveness to the strength of the corrective signal. The corrective-signal measure is constructed so that larger values indicate greater confidence of the detector in the objectively correct classification. For a true news item, it corresponds to the probability assigned to truth; for a false news item, it corresponds to one minus that probability. The measure is centered and divided by ten, so that its coefficient represents the effect of a

ten-percentage-point increase in corrective confidence.

Table 6 relates classification error to corrective-signal strength, congruence bias, and their interaction.

Table 6: AI-signal responsiveness and ideological congruence bias

	Classification error
Corrective signal	-0.010
	(0.027)
	[0.719]
Congruence bias	0.070
	(0.047)
	[0.143]
Corrective signal × congruence bias	0.015*
	(0.008)
	[0.054]
Participant fixed effects	Yes
Observations	757

Notes: The sample is restricted to Phase 3. The dependent variable equals one when the news item is incorrectly classified. Standard errors clustered at the participant level are reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The strength of the corrective signal does not significantly reduce classification error on average. Its coefficient is negative, as expected under effective correction, but small and statistically insignificant. This result provides no support for the full-trust benchmark, under which participants should respond systematically to the direction and strength of the detector output.

The interaction between corrective-signal strength and congruence bias is positive and marginally significant. Since the dependent variable is classification error, the positive coefficient implies that corrective information becomes less effective as congruence bias increases. The estimate however is not robust to the introduction of a set of observable item controls. The results therefore provide only suggestive and control-sensitive evidence of ideological discounting of the detector signal.

We next test the threshold-based implications of partial mistrust. A weak AI signal is defined as an output between 35 and 65 percent, inclusive. This interval captures proba-

bilities no more than 15 percentage points away from the uninformative benchmark of 50 percent. Given the discrete support of the detector outputs, the indicator identifies the least directional output observed in the experiment. The broader 20–80 definition is examined as a sensitivity check in Appendix 6.

Table 7 reports the estimates.

Table 7: Weak AI signals and ideological congruence bias

	Classification error	
	(1)	(2)
Weak AI signal	0.264^{***}	0.158^{*}
	(0.083)	(0.095)
	[0.002]	[0.099]
Congruence bias	0.150^{***}	0.155^{***}
	(0.028)	(0.029)
	[< 0.001]	[< 0.001]
Weak AI signal × congruence bias	-0.078	-0.094
	(0.153)	(0.152)
	[0.612]	[0.539]
Participant fixed effects	Yes	Yes
Item controls	No	Yes
Observations	757	757
Participants	76	76

Notes: The sample is restricted to Phase 3. Weak AI signal equals one when the probability reported by the detector lies between 35 and 65 percent, inclusive, and zero otherwise. Given the discrete support of the detector probabilities, this indicator identifies the least directional output observed in the experiment. Weak signal and congruence bias are mean-centered. Item controls include objective veracity, AI generation, news direction, and topic indicators. Standard errors clustered at the participant level are reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Weak detector outputs are associated with substantially more classification errors. Without item controls, the least directional output increases the estimated probability of error by 26.4 percentage points. After observable item characteristics are included, the estimate remains positive at 15.8 percentage points and is marginally significant. The broader 20-80 definition produces the same qualitative conclusion, although the magnitude and precision of the coefficient vary across specifications. The evidence is therefore consistent with inter-

mediate outputs providing less useful guidance than strongly directional signals.

The interaction between weak signals and congruence bias is negative, small relative to its standard error, and statistically insignificant in both columns. The same null interaction is obtained under the broader definition reported in the appendix. The threshold-based evidence therefore supports non-selective rather than selective partial mistrust: weak outputs appear difficult to use in general, but not disproportionately so when ideological congruence favors an incorrect classification.

At the same time, congruence bias remains a strong predictor of error within Phase 3. Its coefficient is approximately 0.15 and highly significant in both specifications. The belief formed from the news content therefore remains strongly associated with classification even after participants receive an external AI signal.

The discrete weak-signal results and the continuous-signal interaction provide complementary evidence. The former are most consistent with non-selective partial mistrust: insufficiently directional outputs often fail to improve classification regardless of their ideological implications. The latter provide borderline evidence that responsiveness to corrective information may also decline gradually as congruence bias increases. Thus, any selective attenuation does not appear to be confined to a sharply defined weak-signal region.

Taken together, the results provide little support for full and uniform trust in the detector. Literal full mistrust is also too strong, since classification differs between the least directional and stronger outputs. The evidence is most consistent with heterogeneous partial reliance: participants make limited use of weak outputs, while the effectiveness of corrective information may also be attenuated when congruence-related pressure toward error is stronger.

This interpretation connects the three parts of the analysis. Making AI verification explicit does not improve average accuracy and is associated with suggestive evidence of a shift toward false negatives. At the same time, congruence-related errors become stronger in Phase 3. The detector therefore supplies additional information without consistently

displacing the politically structured assessment already formed from the news content.

The central limitation of AI-based verification may consequently lie not in a uniform rejection or reversal of algorithmic advice, but in its limited capacity to override prior assessments. Intermediate outputs often provide insufficient guidance, while even stronger corrective information may become less effective as congruence bias increases.

6 Discussion and Conclusion

This paper examines how ideological congruence shapes news classification across different AI-mediated information environments. More specifically, it studies whether congruence-related classification errors change when AI enters the information environment as a generator of misinformation and when it is introduced explicitly as a source of verification signals. The paper also investigates how individuals interpret and respond to AI-generated assessments of news veracity.

Three main findings emerge.

First, ideological congruence is a powerful predictor of classification error. Participants are systematically more likely to make mistakes when those mistakes are ideologically favorable, either by accepting congruent false news as true or by rejecting incongruent true news as false. This relationship is economically meaningful, statistically robust, and remains present across all information environments considered in the experiment.

Second, the relationship between congruence bias and classification error becomes stronger when AI verification is made explicit. Although average classification accuracy remains broadly stable across phases, the AI-detection environment is associated with a stronger link between ideological congruence and misclassification. In addition, false negatives increase relative to the preceding phase, providing suggestive evidence that explicit AI verification may induce greater caution or skepticism toward information. Such skepticism may operate through lower prior beliefs about news veracity, higher acceptance thresholds, or a combi-

nation of both. The present design cannot distinguish these channels directly, and therefore this interpretation should be viewed as suggestive rather than causal evidence of a specific psychological mechanism.

Third, the evidence suggests that individuals do not uniformly incorporate AI-generated verification signals. The theoretical framework distinguishes four possible interpretation regimes: full trust, non-selective partial mistrust, selective partial mistrust, and full mistrust. The empirical results provide little support for full trust, since stronger corrective signals do not systematically improve classification performance. At the same time, literal full mistrust appears too strong, because participants do respond differently to weak and strong detector outputs. The weak-signal analysis is most consistent with non-selective partial mistrust: intermediate detector outputs are associated with higher error rates, but this difficulty does not increase with ideological congruence. Thus, weak AI signals often fail to overturn the judgment formed from the news content itself, regardless of whether the news is ideologically congruent or incongruent.

These findings have two broader implications. First, the role of AI verification cannot be evaluated only by asking whether the detector is technically informative. What matters is also how users interpret and incorporate the signal. A probabilistic AI output may be too weak, ambiguous, or insufficiently credible to displace the judgment already formed from the news content. Second, AI verification changes the informational context in which classification takes place. Making AI explicit may increase caution toward news, but this caution does not necessarily improve accuracy and may leave ideologically structured errors intact.

Several limitations should be acknowledged. First, the experimental phases are presented in a fixed order. This structure avoids contaminating earlier phases with explicit exposure to AI verification, but it also means that phase comparisons may partly reflect sequence, learning, or fatigue effects. Second, the detector signal is not experimentally randomized across items. Weak AI outputs may therefore capture both limited reliance on the detector

and intrinsic item difficulty. Third, the experiment observes classification decisions but does not directly elicit prior veracity beliefs, intermediate beliefs after reading the news, final beliefs after observing the detector, trust in the AI system, or individual classification thresholds. As a result, the design cannot separately identify whether the observed patterns arise from prior shifts, threshold shifts, selective attention, differential scrutiny, or trust in the detector. Fourth, the sample consists of university participants and the task uses short news items, so external validity to broader populations and richer information environments should be assessed in future work.

Future research can build directly on these limitations. A first extension would elicit beliefs at multiple stages: the prior probability that a news item is true, the perceived veracity of the item after reading its content, and the final belief after observing the detector. This would make it possible to distinguish content interpretation, AI-signal interpretation, and classification thresholds. A second extension would measure political beliefs both before and after exposure to the news environment. The present paper identifies the first stage of a potential polarization mechanism, namely congruence-dependent classification of news veracity; future work could test whether these classification errors translate into actual divergence in political beliefs. A third extension would experimentally vary the format of the AI signal, comparing simple probability scores with explanations, source citations, confidence intervals, or counter-arguments. This would help identify whether limited reliance reflects mistrust, misunderstanding, or the perceived ambiguity of intermediate outputs.

Overall, the paper shows that AI verification does not mechanically remove ideological distortions from news classification. Across the environments studied here, classification errors remain strongly shaped by ideological congruence, and this relationship becomes stronger when AI verification is made explicit. At the same time, the detector is not incorporated in a fully uniform way: weak signals are associated with higher error rates, while the evidence does not indicate that this difficulty is selectively stronger when congruence bias is high. The resulting pattern is most consistent with partial and non-selective mistrust of weak AI

outputs. Improving AI verification therefore requires not only more accurate detectors, but also a better understanding of how individuals interpret probabilistic signals in politically charged information environments.

Acknowledgments

This work was supported by the PNRR M4C2, Investment 1.5, D.D. 3277/2021 - Programma di R&I Rome Technopole – Codice Progetto: ECS00000024 financed by the European Union – NextGenerationEU – Spoke 1 (CUP: I83C22000990001).

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The experimental dataset, MATLAB codes, and news dataset are available at Figshare: [https://doi.org/10.6084/m9.figshare.32716803]. The repository contains the dataset as exported from oTree together with all MATLAB scripts necessary to replicate the tables and figures presented in the paper.

Ethics Approval and Consent to Participate

The experiment was approved by the Ethics Committee of LUISS Guido Carli University. All participants provided informed consent before taking part in the study.

Declaration on the Use of Generative AI

During the preparation of this work the author used ChatGPT (OpenAI) in order to improve the clarity and readability of the text. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Fryer, R. G., Harms, P., and Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias. *Journal of Economic Theory*, 183:1–32.

Gentzkow, M. and Shapiro, J. (2011). Ideological segregation online and offline. *Quarterly Journal of Economics*, 126(4):1799–1839.

Greevink, I., Offerman, T., and Romagnoli, G. (2024). Ai-powered promises: The influence of chatgpt on trust and trustworthiness. Technical report, CREED, University of Amsterdam.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Iyengar, S., Sood, G., and Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, 76(3):405–431.

Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4):407–424.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498.

Kuznetsova, E., Vitulano, I., Makhortykh, M., Stolze, M., Nagy, T., and Vziatysheva, V. (2025). Fact-checking with generative ai: A systematic cross-topic examination of llms capacity to detect veracity of political information. *arXiv preprint*.

Mei, K. X., Wolfe, R., Weber, N., and Saveski, M. (2026). Grok in the wild: Characterizing the roles and uses of large language models on social media.

Nyhan, B. (2020). Facts and myths about misperceptions. *Journal of Economic Perspectives*, 34(3):220–236.

Nyhan, B. (2021). Facts and myths about misperceptions. *Journal of Economic Perspectives*, 35(3):220–236.

Nyhan, B. and Reifler, J. (2010a). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Nyhan, B. and Reifler, J. (2010b). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Park, S. (2025). Generative ai and misinformation: A scoping review of the role of large language models in generating and spreading misinformation. *AI & Society*.

Pennycook, G. and Rand, D. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning. *Cognition*, 188:39–50.

Pilati, F. and Venturini, T. (2025). The use of artificial intelligence in counter-disinformation: a world wide (web) mapping. *Frontiers in Political Science*, 7.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Stroud, N. (2010). Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576.

Taber, C. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.

Wood, T. and Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163.

Appendix A: Robustness Analysis

This appendix reports a parsimonious set of robustness checks for the main empirical results. We first document observable differences in the composition of the experimental phases. We then assess whether the relationship between ideological congruence bias and classification error is robust to controls for item characteristics, presentation order, and participant-cluster bootstrap inference. Finally, we examine the stability of the Phase 3 signal analysis after controlling for observable item characteristics.

Composition of the Experimental Phases

Table 8 reports the observable composition of the news items presented in each phase. The proportion of false items is by design the same across phases, whereas the topic distribution may change across phases. Moreover, ai generated content appears only in phase 2 and 3. These differences motivate the inclusion of item-level controls in the robustness specifications reported below.

Table 8: Observable composition of the experimental phases

	Phase 1	Phase 2	Phase 3
News side	0.434	0.525	0.410
Foreign policy	0.259	0.250	0.156
Immigration	0.313	0.363	0.182
Civil rights/LGBT	0.184	0.068	0.362

Notes: Entries report phase-specific sample means. Fake news, AI-generated content, news side, and topic categories are indicator variables. Economy is the omitted topic and can be recovered from the remaining topic shares.

Robustness of Congruence-Bias and Signal Results

This section evaluates the robustness of two results: the stronger association between congruence bias and classification error in Phase 3, and the relationship between congruence bias and responsiveness to the corrective AI signal.

Table 9 re-estimates the phase-interaction model under alternative specifications. Column (1) includes the full set of observable item controls: objective veracity, news side, and topic indicators. Column (2) controls for the linear and quadratic position of each item within the corresponding experimental phase.

Table 9: Robustness of ideological congruence effects

	Classification error	
	Full item controls (1)	Order controls (2)
Phase 2	0.019	-0.019
	(0.025)	(0.022)
	[0.444]	[0.382]
Phase 3	0.030	0.016
	(0.026)	(0.023)
	[0.248]	[0.483]
Congruence bias	0.063**	0.063**
	(0.030)	(0.030)
	[0.036]	[0.036]
Congruence bias × Phase 2	-0.054	-0.050
	(0.045)	(0.043)
	[0.237]	[0.245]
Congruence bias × Phase 3	0.087**	0.089**
	(0.038)	(0.038)
	[0.024]	[0.022]
Within-phase order		-0.001
		(0.003)
		[0.637]
Within-phase order squared		0.0004
		(0.0012)
		[0.741]
Participant fixed effects	Yes	Yes
Full item controls	Yes	No
Observations	2,277	2,277

Notes: Linear probability models. The dependent variable equals one when the item is incorrectly classified. Phase 1 is the omitted category. Full item controls include objective veracity, AI-generation status, news side, and topic indicators. Standard errors are clustered at the participant level and reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The congruence-bias result is highly stable. The coefficient on congruence bias remains positive and statistically significant after controlling for observable item characteristics and after accounting for within-phase presentation order. More importantly, the interaction

between congruence bias and Phase 3 remains close to the main estimate and statistically significant at the five-percent level in both specifications. The stronger association between congruence bias and classification error in the AI-detection environment is therefore not explained by observable item composition or by the order in which items are presented.

Participant-cluster bootstrap inference leads to the same conclusion. The bootstrap mean of the congruence-bias coefficient is 0.062, with a 95-percent confidence interval of $[0.005, 0.121]$ and $p = 0.036$. The bootstrap mean of the Phase 3 interaction is 0.089, with a 95-percent confidence interval of $[0.016, 0.166]$ and $p = 0.019$. Thus, the Phase 3 amplification of congruence-related classification error is robust both to alternative controls and to resampling-based inference.

Table 10 examines the robustness of the Phase 3 continuous-signal specification. Column (1) reports the baseline model, while Column (2) adds the full set of observable item controls.

“‘latex

The average association between corrective-signal strength and classification error remains small and statistically insignificant after the item controls are introduced. The interaction between the corrective signal and congruence bias is positive and marginally significant in the baseline specification, but becomes smaller and statistically insignificant with the full set of item controls. The continuous-signal interaction should therefore be interpreted as suggestive rather than robust.

Congruence bias itself remains positively associated with classification error and becomes statistically significant in the controlled specification. Among the item characteristics, false items are associated with fewer errors, while the coefficient on news direction indicates that items supporting the lower-valued endpoint of the relevant political scale are, on average, classified more accurately. These item-level differences are included to account for heterogeneous item difficulty and are not interpreted as central substantive results.

Table 10: Robustness of AI-signal responsiveness

	Classification error	
	Baseline (1)	Full item controls (2)
Corrective signal	-0.010	-0.014
	(0.027)	(0.027)
	[0.719]	[0.610]
Congruence bias	0.070	0.101**
	(0.047)	(0.048)
	[0.143]	[0.041]
Corrective signal × congruence bias	0.015*	0.010
	(0.008)	(0.008)
	[0.054]	[0.171]
False news		-0.189***
		(0.044)
		[< 0.001]
News side		-0.106***
		(0.040)
		[0.009]
Foreign-policy topic		0.118*
		(0.061)
		[0.058]
Immigration topic		0.096
		(0.065)
		[0.145]
LGBT topic		0.107**
		(0.051)
		[0.040]
Participant fixed effects	Yes	Yes
Observations	757	757

Notes: The sample is restricted to Phase 3. The dependent variable equals one when the item is incorrectly classified. The corrective signal is centered and divided by ten. The full item-control specification includes objective veracity, news direction, and topic indicators, with economic policy as the omitted topic category. Standard errors are clustered at the participant level and reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Sensitivity to Alternative Definitions of Weak AI Signals

The main specification defines a weak AI signal as a detector probability between 35 and 65 percent. Since the experimental detector outputs take only a limited set of discrete values, we assess the sensitivity of the results using a broader definition that classifies as weak outputs between 20 and 80 percent.

Table 11: Sensitivity to alternative definitions of weak AI signals

	Without item controls		With item controls	
	20-80	35-65	20-80	35-65
Weak AI signal	0.083	0.264***	0.150**	0.158*
	(0.057)	(0.083)	(0.062)	(0.095)
	[0.153]	[0.002]	[0.018]	[0.099]
Congruence bias	0.151***	0.150***	0.151***	0.155***
	(0.030)	(0.028)	(0.029)	(0.029)
	[< 0.001]	[< 0.001]	[< 0.001]	[< 0.001]
Weak AI signal × congruence bias	-0.081	-0.078	-0.058	-0.094
	(0.124)	(0.153)	(0.109)	(0.152)
	[0.515]	[0.612]	[0.599]	[0.539]
Participant fixed effects	Yes	Yes	Yes	Yes
Item controls	No	No	Yes	Yes
Observations	757	757	757	757

Notes: The sample is restricted to Phase 3. Each column uses a different interval to define the weak-signal indicator. Weak signal and congruence bias are mean-centered before constructing the interaction. Item controls include objective veracity, AI generation, news direction, and topic indicators. Standard errors clustered at the participant level are reported in parentheses; p-values are reported in square brackets. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The sensitivity analysis yields three main conclusions. First, the direct association between weak detector outputs and classification error is positive under both definitions. Under the 35-65 benchmark, the coefficient is larger and statistically significant without controls, while remaining positive and marginally significant after controls are introduced. Under the 20-80 broader specification, the coefficient is positive but statistically insignificant without controls and becomes significant after observable item characteristics are included. Second, the estimated effect of congruence bias is highly stable across all specifications. Third, the interaction between weak signals and congruence bias is statistically insignificant in every

specification. Its estimated sign is consistently negative, but the coefficients are imprecisely estimated and far from conventional significance levels. The results therefore provide no evidence that weak detector outputs are disregarded only depending on congruence bias.

Overall, the positive relationship between weak detector outputs and classification error is moderately robust, although its statistical precision depends on the chosen cutoff and on the inclusion of item controls. By contrast, the strong direct effect of congruence bias and the absence of a significant weak-signal interaction are highly stable across the two definitions.

Appendix B: Experimental Instructions

General Instructions

Welcome to this experiment.

During this fully computerized experiment, your decisions will determine your earnings.

For participating in the experiment, you will receive a show-up fee of **3.00 €**.

In addition, you can earn more money based on your decisions and chance. Both the show-up fee and any additional earnings will be paid at the end of the experiment via PayPal.

All earnings in the experiment are expressed in **ECU** (Experimental Currency Units). At the end of the experiment, your earnings in ECU will be converted at the rate of **1 ECU = 1 €**.

All participants in this experiment are reading the same instructions, which you can re-read at any time by pressing the **INSTRUCTIONS** button at the top of your screen. All participants are taking part in the experiment for the first time.

Please read the instructions carefully. If you have any questions, raise your hand and an experimenter will answer you privately.

From this point onward, any communication between participants is prohibited.

Experiment Instructions

In this study, you will complete a short task divided into **three phases**. In each phase, you will see a series of short news items (a headline + a brief summary). Your task is to decide whether each news item is **True** or **False**.

For each news item you will:

- Read the headline and the short summary.
- Choose one of the two options: **TRUE** or **FALSE**.
- After confirming your choice, you will move on to the next news item.

You will complete the task in **three phases**. The procedure is the same in each phase: you will evaluate news items and classify them as **TRUE** or **FALSE**.

Your payment consists of two parts:

- **Fixed payment**: 3.00 € (show-up fee).
- **Performance bonus**: At the end of the study, we will randomly select **3 news items**, one from each phase. For each selected news item:
  - If your classification is **correct**, you earn **5.00 €**.
  - If your classification is **incorrect**, you earn **0.00 €** for that news item.

Total earnings = 3.00 € (fixed payment) + bonus from the 3 randomly selected news items (up to 15.00 €).

Important notes:

- Work independently and do not communicate with other participants.
- Some news items may be difficult — do your best.
- There is no penalty for the time you take, but we ask you to stay focused throughout the task.

Good luck.