Towards Designing Content Labels under Detection Uncertainty: A Taxonomy and Experimental Research Agenda

Research-in-Progress Paper

René Theuerkauf
Martin Luther University Halle-Wittenberg
Halle (Saale), Germany
rene.theuerkauf@wiwi.uni-halle.de

Abstract

Platforms and regulators increasingly use labels to govern fake and AI-generated content. The EU AI Act requires provenance disclosure, and veracity warnings remain the main moderation tool. Because detection is unreliable, most problematic content stays unlabeled. Labels can then backfire, as users may take unlabeled misinformation as checked (implied truth effect). How this depends on label design is unknown. Studies disagree on whether users process what a label says or treat any label as a generic caution cue. We develop a preliminary taxonomy of user-facing content labels and select its two least understood dimensions, label target (veracity vs. provenance) and certainty framing (binary vs. probabilistic). We propose a preregistered feed experiment that isolates both, with a stimulus pool in which truth and origin vary independently. The agenda offers design guidance for transparency labels that regulation now requires. Research has not yet tested how their certainty framing affects belief and spillover.

Keywords: content labeling, misinformation, AI-generated content, implied truth effect, taxonomy

Introduction

Platforms can respond to fake and AI-generated content in two main ways. They can remove it or label it. Both depend on detection, and detection is unreliable. Adversarial attacks can reduce the accuracy of a state-of-the-art detector to almost zero (Rabhi et al., 2024). False negatives leave fake content unlabeled. Because warnings make unlabeled content look checked, selective labeling can lend credibility to the very misinformation it misses. This is the implied truth effect (Pennycook et al., 2020). False positives attach warnings to legitimate content and taint true news (Freeze et al., 2021).

Labeling is often seen as the more error-tolerant alternative to deletion. A wrongly labeled post stays visible and can be corrected, while a wrongly deleted post cannot. Regulators follow this logic. Article 50 of the EU AI Act, applicable since 2 August 2026, requires disclosure of certain AI-generated or manipulated content. Yet studies disagree on a basic design question, namely whether users process what a label says. Altay and Gilardi (2024) find that an "AI-generated" label has about one third of the effect of a "False" label on the same headlines, which suggests that users process its meaning. Pawelczyk et al. (2026) find no difference between process-based and harm-based labels, which suggests that users treat any label as a generic caution cue. AI labels also spill over onto unlabeled images (Pawelczyk et al., 2026), but to our knowledge no study has tested whether this spillover depends on what the label says.

Detection uncertainty reaches users in two ways. It shapes how certain a single label sounds, and it limits how much problematic content the labeling regime, the platform's labeling practice as a whole, can cover. Coverage decides what a missing label means. Our experiment isolates the two taxonomy dimensions with the least evidence, label target and certainty framing, in a 2×2 design plus control. Its stimulus pool varies truth and origin independently. The experiment manipulates how certain labels sound and measures what users infer about unlabeled content. A follow-up study manipulates coverage directly.

Background and Theoretical Framing

What labels do. Veracity warnings are among the best-evidenced interventions against misinformation. A meta-analysis across 21 experiments finds that fact-checker labels reliably reduce both belief in and sharing of false headlines, even among users who distrust fact-checkers (Martel & Rand, 2024). Provenance labels are more ambivalent. They reduce perceived accuracy regardless of whether a headline is true or false (Altay & Gilardi, 2024) and carry a small credibility penalty on average (Kim et al., 2026), yet C2PA-based disclosures can increase trust in the platform displaying them (Trattner et al., 2026).

What labels say. Across two experiments, every tested label lowered belief in misleading AI-generated images, but labels that only disclosed AI generation barely changed engagement intentions, which suggests that users respond to what a label asserts (Wittenberg et al., 2025). If labels are generic caution cues, regulators need only mandate a label. If users read them semantically, label target is a first-order design decision. No study has resolved this within one controlled design.

How certain labels sound. Detection is probabilistic. Classifiers produce confidence scores, and each label either passes this uncertainty on to users or hides it. People read probability words as less extreme than intended and in line with their prior beliefs (Budescu et al., 2014), and displayed confidence changes how far users follow an algorithm (Zhang et al., 2020). A binary label states a categorical judgment, whereas a probabilistic one qualifies its own claim, which should lower its perceived certainty and its diagnosticity, that is, how informative users find it. Platforms almost always use binary labels. Gong et al. (2026) find that a hedged AI label ("Suspected AI, please verify") increases information avoidance, but no study has tested how certainty framing affects belief in labeled and unlabeled content.

What labels imply. Because detection is imperfect, most fake content never carries a label. This makes the indirect effect on unlabeled content critical for policy. Pennycook et al. (2020) show that warnings on some false headlines raise the perceived accuracy of the others, as if users infer that unlabeled content has been checked. Verification tags on true items remove this effect. Users thus reason about the labeling regime as a whole, in addition to the labeled item. For AI labels, a first study finds an implied authenticity effect of about one fifth of the direct effect, but it pooled both label types (Pawelczyk et al., 2026). Detection uncertainty also works at the level of the regime.

Process model. We combine these findings into a process model with three pathways. Pathway (a) concerns the labeled item. A label is a signal about the content, and its effect on believability depends on how diagnostic users find it. If users treat labels as generic cues, diagnosticity is the same for all label targets. If they process the content, it varies with what the label says. Certainty framing acts at the same point. Pathway (b) concerns unlabeled items. A labeling regime invites the inference that unlabeled content has been checked, and this inference should be stronger the more a label claims. A veracity label makes a claim about truth, a provenance label only about origin. Veracity regimes should create a stronger implied truth effect, a prediction that Pawelczyk et al. (2026) state for harm-based labels but test only with label types pooled. The inference should also weaken when the regime admits that it can fail, through probabilistic framing or an explicit note on coverage. Pathway (c) covers platform-level outcomes of wrong moderation decisions and guides the agenda below.

A Preliminary Taxonomy of Content Labels

The taxonomy follows Nickerson et al. (2013). Its meta-characteristic is the design features of platform-issued, user-visible labeling interventions for content classified as fake or AI-generated that shape user perception. We classify only labels that users actually see. Watermarks count only if they appear in a visible label. General warnings shown before a feed, downranking and deletion are neighboring interventions outside the scope.

We developed the taxonomy in three iterations. The first, conceptual-to-empirical iteration derived five dimensions from the literature. The second, empirical-to-conceptual iteration classified six prominent platform labels and added creator self-disclosure, a source the literature had missed. After two rounds of methods review with language models, we made three changes. We added an inclusion check, which asks whether a label refers to a single piece of content and says something about its truth, origin, verification or context. We split intrusiveness into access friction and sharing friction and defined clear thresholds for explanation depth. This led to six dimensions. In the third iteration, we applied the taxonomy to real labels. We selected them on purpose to cover as many label designs as possible, so the sample does not show how common each design is. The labels come from four groups of platforms, namely social networks, public messaging channels, news and search services, and e-commerce and review sites. They span 2020 to 2026 and include labels that platforms have since discontinued. Of the 25 objects, 18 passed the inclusion check, 17 real labels and one constructed probabilistic label. During this iteration, one rule revision merged two characteristics of D6. Nickerson et al. (2013) end the development only when no characteristics change, so the taxonomy is not yet final.

Dimension	Characteristics	Use in Study 1
D1 | Label target	veracity (3) · misleadingness (1) · provenance (8) · verification (3) · context (3)	manipulated (veracity vs. provenance)
D2 | Displayed source	platform/system · fact-checker · community · creator · credential issuer · not attributed	constant (platform/system)
D3 | Certainty framing	categorical (14) · none (3; context labels only) · probabilistic (1, constructed)	manipulated (binary vs. probabilistic)
D4 | Access friction	none · notice · partial overlay · full overlay	constant (none)
D5 | Sharing friction	none · notice · confirmation · restricted	constant (none)
D6 | Explanation depth	none · short rationale · extended context	constant (none)

Table 1: Preliminary taxonomy of user-facing content labels

Before human coding, we tested the rulebook with four large language models from different providers. Each model coded all objects from the rules alone, with the same prompt. We read disagreement with our reference coding as a sign of an unclear rule and report agreement between the models (Fleiss' κ), which does not depend on our reference. Certainty framing fell to $\kappa = 0.38$ [−0.02, 0.64], well below the threshold set in advance, because the rulebook did not define a verdict. Once it did, and the D1–D3 consistency rule was completed, agreement rose to $\kappa = 0.92$ [0.69, 1.00] (other dimensions 0.81–1.00). Because we had revised the rules on the same objects, we then applied them unchanged to seven test objects with combinations of characteristics the rules were not built on. Expected codes were set before any run. The models agreed at $\kappa = 0.87$ [0.69, 0.99]

across all cells and matched 91% of the expected codes. Certainty framing reached only $\kappa = 0.58$ [0.19, 0.87], and two of the three disagreements came from earlier inclusion and label-target decisions. By a rule set in advance (at least three of four models deviating in the same direction), one expectation counted as refuted, since all models read a partially verified seal as categorical. Agreement between language models says nothing about human coders. Until human coding is complete, the taxonomy is empirically informed but not validated.

We isolate label target (D1) and certainty framing (D3) for four reasons. First, in our process model D1 drives the implied truth inference most directly, and D3 determines how diagnostic a label appears. Second, evidence on both is thin. Direct effects of D1 are contested, and no study has compared its spillover. For D3, the only related study measures information avoidance, not belief (Gong et al., 2026), and no study has crossed the two dimensions. Third, both dimensions can be manipulated while all others stay constant. Fourth, the pair matters in practice. None of the 17 documented interventions states its verdict as a probability where users first see it (Table 1), and Article 50 requires provenance disclosure but does not specify its framing. We fix D2 because labels attributed to a system may appear neutral (Horne, 2025), D4 and D5 because more intrusive labels are not always more effective (Liban & Horne, 2026), and D6 because explanations did not improve trust calibration in a controlled test (Zhang et al., 2020).

Proposed Experiment

Design and treatments. The experiment will be a preregistered, vignette-based feed study with a 2×2 between-subjects design plus a no-label control. It crosses label target (veracity vs. provenance) with certainty framing (binary vs. probabilistic). The four label templates look identical and name the same source, the platform's detection system. They read "False information", "Likely false", "AI-generated" and "Likely AI-generated". Probabilistic framing is verbal, because senders usually express uncertainty in words (Dhami & Mandel, 2022).

Stimuli and feed. In the stimulus pool, truth and origin vary independently. If all false items were AI-generated, a provenance label would also signal falsehood. To create an AI-generated item, we give a language model the facts of a human-written item and keep the model's natural style, because this style is part of what AI-generated text looks like in the field. Items are matched on length and topic, and emotionality, plausibility and style are pretested against truth status. Each set of facts exists in a human and an AI version, which rotate across participants.

Each participant browses 18 items in two blocks. The treatment block of 12 items carries the labels, on the six false items in veracity conditions and on the six AI-generated items in provenance conditions. The number of labels is the same in all conditions, and labeled false AI-generated items appear in every condition, which provides a common comparison set. The test block contains six unlabeled, human-written items, three true and three false, identical across conditions. A label share of 50 percent is far higher than on real platforms. We chose it on purpose as a proof of concept. Because such a high share may reveal the purpose of the study, we measure the consequences of labeling only after the labeled block.

Hypotheses. The confirmatory hypotheses concern labeled content and adjudicate between the generic-cue and semantic accounts, using the false AI-generated items labeled in every condition.

> *H1 (label target):* Veracity labels reduce the believability of labeled false AI-generated items more strongly than provenance labels on the same items.
>
> *H2a (certainty framing):* Binary labels reduce believability more strongly than probabilistic labels.
>
> *H2b (mechanism):* The framing effect is mediated by perceived label diagnosticity.
>
> *H3 (exploratory, spillover):* Unlabeled false items are perceived as more accurate under labeling than in control, more so under veracity than provenance regimes, and less so under probabilistic framing.

Measures. Participants rate each item on perceived accuracy with the item of Pennycook et al. (2020) and sharing intention. Accuracy ratings serve as the main measure of believability. A subset also answers a three-item believability scale (Moravec et al., 2020). Right after the test block, two items measure the inference about the regime, namely how complete users think moderation is and whether unlabeled posts have been checked. Neutral questions about the platform surround them to limit recall bias and demand effects. The post-questionnaire measures trusting beliefs in the platform (McKnight et al., 2002) and perceived label certainty and diagnosticity, the mediator for H2b. A manipulation check asks what the label claimed, which separates a

failed manipulation from real generic-cue processing. A framing check asks participants to translate the label wording into a number, to verify that the probabilistic template conveys lower certainty. We also pretest this.

Analysis and power. We analyze the item ratings with linear mixed models with crossed random intercepts for participants and items. H1 and H2a are tested with planned contrasts on the labeled items. H2b is tested with a preregistered mediation analysis. Because the mediator is measured after the outcome and not manipulated, this analysis can support the mechanism but cannot prove it. An analytic power analysis, checked by simulation, assumes intraclass correlations of 0.30 for participants and 0.10 for items. With about 120 participants per condition ($N \approx 600$, Prolific), it gives 80% power for differences of at least 0.18 SD for label target and 0.16 SD for framing. We report H3 and the mechanism items as estimates with confidence intervals and adjust their p-values for multiple testing (false discovery rate).

Research Agenda and Next Steps

The next step consolidates the taxonomy. Human coders will code all objects independently with the frozen rules. We compute κ before any adjudication and accept a dimension at $\kappa \geq 0.70$. Before finalizing the taxonomy, we will assess the objective and subjective ending conditions following Nickerson et al. (2013). Only then do we preregister and run Study 1, so that its treatments rest on confirmed dimensions.

Two further studies build on Study 1. A follow-up study manipulates uncertainty at the regime level. It crosses label density with an explicit notice that not all problematic content can be detected. Because platforms often show veracity and provenance labels together, it also combines both label targets on the same items. The false-positive side of detection uncertainty (pathway c) is also on our agenda. Wrongful warnings taint true news (Freeze et al., 2021), and a later study will test whether labeling is more error-tolerant than deletion.

Expected Contributions

Once completed, the research project can contribute to theory, method and practice. For theory, it can provide a validated taxonomy of user-facing content labels that shows which design features platforms combine and where practice leaves gaps. It can also test the process model, which separates detection uncertainty at the level of the single item from uncertainty at the level of the labeling regime. This includes a direct test of whether users process labels as generic caution cues or by their content, for labeled items (H1, H2a, H2b) and, exploratively, for spillover to unlabeled items (H3). For method, it can provide a treatment design derived from the taxonomy. For practice, it can show how platforms should design labels to meet the transparency duties of the EU AI Act and the DSA.

Acknowledgements

Generative AI was used in three roles. For the taxonomy, four commercial large language models from different providers coded the taxonomy objects and the test objects. Their agreement is part of the instrument evaluation reported above. For the methodological setup, a language model served as a sparring partner. In this role it worked like a critical colleague. The authors used it to turn early ideas into concrete design options and then to stress-test them. The ideas and all design decisions are the authors' own. As a writing aid, a language model helped to shorten and simplify author-written text. The authors checked all output and take full responsibility for the final version.

References

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, *3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508–512. https://doi.org/10.1038/nclimate2194

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, *26*(6), 514–526. https://doi.org/10.1016/j.tics.2022.03.002

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Gong, Z., Peng, D., Cui, J., & Lv, Z. (2026). The paradox of AI content labeling: How clarity influences information avoidance via cognitive dissonance on social platforms. *Frontiers in Psychology*, *17*, 1751670. https://doi.org/10.3389/fpsyg.2026.1751670

Horne, B. D. (2025). Does the Source of a Warning Matter? Examining the Effectiveness of Veracity Warning Labels Across Warners. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 823–836. https://doi.org/10.1609/icwsm.v19i1.35848

Kim, H. M., Lee, E.-J., Park, J. W., & Walter, N. (2026). Synthetic News, Natural Doubts? A Meta-Analysis of Credibility Perceptions of AI-Generated News. *Cyberpsychology, Behavior, and Social Networking*, *29*(5), 281–288. https://doi.org/10.1177/21522715261439452

Liban, H. S., & Horne, B. D. (2026). The Role of Intrusiveness and Color in Content Warning Label Effectiveness. *ACM Transactions on Applied Perception*, *23*(3), 1–22. https://doi.org/10.1145/3805797

Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, *8*(10), 1957–1967. https://doi.org/10.1038/s41562-024-01973-x

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research*, *31*(3), 987–1006. https://doi.org/10.1287/isre.2020.0927

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26

Pawelczyk, F., Dimmery, D., & Yan, P. (2026). Implied Authenticity Effect? The Impact of Explicit Labels on AI-Generated Content. *Proceedings of the International AAAI Conference on Web and Social Media*, *20*(1), 1738–1766. https://doi.org/10.1609/icwsm.v20i1.42721

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, *250*, 123941. https://doi.org/10.1016/j.eswa.2024.123941

Trattner, C., Forstner, S. L., Starke, A. D., & Knudsen, E. (2026). C2PA Provenance Labels Increase Trust in Digital News Platforms Across Western Countries. *Proceedings of the International AAAI Conference on Web and Social Media*, *20*(1), 2267–2279. https://doi.org/10.1609/icwsm.v20i1.42749

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, *4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852