RESEARCH-IN-PROGRESS

Towards Designing Content Labels under Detection Uncertainty

A taxonomy and experimental research agenda

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

Workshop on e-Business (WeB) · Lisbon 2026

0 of 17

real platform labels state their verdict as a probability.

Detection is probabilistic. Labels sound certain. Nobody has tested what that does to belief in labeled and unlabeled content.

1 Problem

Detection is unreliable. Adversarial attacks can push a state-of-the-art detector to almost zero accuracy (Rabhi et al., 2024).

Missed fakes look checked. Warnings on some false items raise belief in the unlabeled rest: the implied truth effect (Pennycook et al., 2020).

Wrong warnings taint true news. Overused warnings erode trust in credible sources. (Freeze et al., 2021).

Regulation mandates labels anyway. EU AI Act Art. 50, applicable since 2 August 2026, requires AI disclosure but leaves its framing open.

RESEARCH QUESTION

Do users read what a label says, or treat any label as a generic caution cue? Semantic: Altay & Gilardi (2024) | Generic cue: Pawelczyk et al. (2026)

2 Process model: three pathways

(a) Labeled items · Study 1, H1–H2b

Label design (Label target × Certainty framing) → Perceived label diagnosticity → Believability of labeled items (−)

(b) Unlabeled items · Study 1, H3 exploratory

Selective labeling regime (claim scope, admitted fallibility) → Inference “unlabeled = checked” → Perceived accuracy of unlabeled false items (+)

(c) Moderation errors · later study

Wrong label vs. wrong deletion → Platform-level outcomes

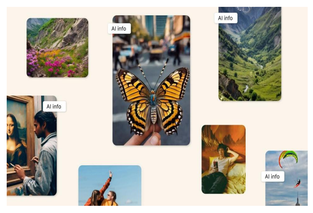
Detection uncertainty reaches users twice: in how certain a single label sounds, and in how much content the labeling regime can cover.

3 A preliminary taxonomy of user-facing content labels

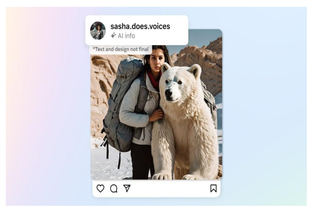
UNVALIDATED DRAFT

Method: Nickerson et al. (2013) · 3 iterations · 25 objects, 18 passed the inclusion check

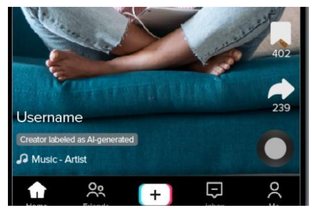
EXAMPLES: REAL AI LABELS ON PROMINENT PLATFORMS



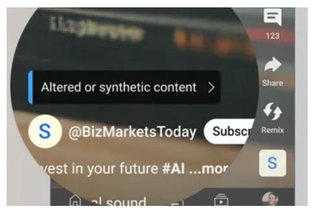
Meta · “AI info”
Facebook, Instagram, Threads. Set by platform from industry signals. Coded: provenance · platform/system · categorical
Source: Meta Newsroom, about.fb.com/news/2024/04 (5 Apr 2024)



Instagram · “AI info”
Creator self-disclosure variant. Coded: provenance · creator · categorical
Source: Meta Newsroom, about.fb.com/news/2024/02 (6 Feb 2024)



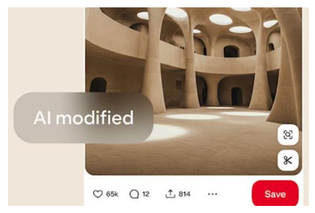
TikTok · “Creator labeled as AI-generated”
Creator toggle; automatic “AI-generated” label via C2PA since May 2024. Coded: provenance · creator · categorical
Source: newsroom.tiktok.com/generative-ai-content-labels (21 Sep 2023)



YouTube · “Altered or synthetic content”
Creator disclosure; shown in description or on the player. Coded: provenance · creator · categorical
Source: blog.youtube/news-and-events/disclosing-ai-generated-content (18 Mar 2024)



LinkedIn · “Cr” Content Credentials
“Cr” icon top left; automatic when C2PA metadata is present. Coded: verification · credential issuer · categorical
Source: linkedin.com/help/linkedin/answer/a6282984 (help page, n.d.)



Pinterest · “AI modified”
Set by platform from metadata and classifiers; bottom left of Pin close-up. Not part of the coded sample
Source: newsroom.pinterest.com/news/introducing-gen-ai-labels (30 Apr 2025)

DIMENSION	CHARACTERISTICS (COUNT IN SAMPLE)	STUDY 1
D1 · Label target What the label makes a claim about	**veracity (3)** · misleadingness (1) · **provenance (8)** · verification (3) · context (3)	**manipulated**
D2 · Displayed source Named origin of the label, if any	platform/system · fact-checker · community · creator · credential issuer · not attributed	constant: platform
D3 · Certainty framing Wording of the label’s confidence	**categorical (14)** · none (3, context labels only) · **probabilistic (1, constructed)**	**manipulated**
D4 · Access friction Extra steps before content is visible	none · notice · partial overlay · full overlay	constant: none
D5 · Sharing friction Extra steps before content can be shared	none · notice · confirmation · restricted	constant: none
D6 · Explanation depth How much the label says about why	none · short rationale · extended context	constant: none

Sample chosen for variety of designs, not frequency: social networks, public messaging channels, news and search services, e-commerce and review sites, 2020–2026.

ITERATION 1
conceptual-to-empirical
ScanLit in combination with Consensus and NotebookLM

ITERATION 2
empirical-to-conceptual
Labels collected on 10 prominent platforms

ITERATION 3
conceptual-to-empirical
Rules revised after LLM pretests

EVALUATION
rulebook check
LLM transfer test done; human coding pending

4 Stress-testing the rulebook with four LLMs

Before human coding, four LLMs from different providers coded all objects from the rules alone. Disagreement flags an unclear rule.
Reported: agreement between models (Fleiss’ κ, 95% bootstrap interval).

	OBJECTS	RULES	SCOPE	AGREEMENT (FLEISS’ κ)	
DEVELOPMENT					
1 First rulebook “verdict” undefined	development	v1	D3		**0.38**
↓ *rule revised: “verdict” defined, objects unchanged*					
2 Revised rules in-sample check	same	**revised**	D3		**0.92**
TRANSFER					
↓ *rules frozen, applied to 7 unseen objects*					
3 Transfer test rulebook carries over	**7 unseen**	**frozen**	**all cells**		**0.87**
↓ *same test, D3 shown separately*					
4 Transfer test D3 still the weakest	same 7	frozen	**D3 only**		**0.58**

Bar: κ from 0 to 1 · black line: 95% interval ([−0.02, 0.64], [0.69, 1.00], [0.69, 0.99], [0.19, 0.87]) · dashed: acceptance threshold κ ≥ 0.70 · green tag: changed from the step before · other dimensions after revision: 0.81–1.00

91% of expected codes matched in the transfer test. One expectation refuted: all four models read a “partially verified” seal as categorical.

Limit. Agreement between LLMs says nothing about human coders. Human coding with frozen rules is the next step.

5 Proposed experiment (Study 1)

Preregistered feed study, 2×2 between subjects plus no-label control. Same template and named source (the platform’s detection system); only the wording differs.

	Binary	Probabilistic
Veracity	**False information**	**Likely false**
Provenance	**AI-generated**	**Likely AI-generated**

Feed: 18 items per participant, truth and origin varied independently

Treatment block: 12 items, 6 labeled (dark)

Test block: 6 unlabeled, 3 true, 3 false

H1	Veracity labels reduce believability of labeled false AI-generated items more than provenance labels.
H2a	Binary labels reduce believability more than probabilistic labels.
H2b	Mediated by perceived label diagnosticity.
H3	(exploratory) Unlabeled false items seem more accurate under labeling; more under veracity, less under probabilistic framing.

N ≈ 600 (Prolific), 5 conditions x 120 · linear mixed models, crossed random intercepts for participants and items · 80% power for 0.18 SD (target) and 0.16 SD (framing)

6 Agenda and your input

1 Human coding, frozen rules (accept at κ ≥ 0.70).
2 Preregister and run Study 1.
3 Follow-up: label density × explicit coverage notice.

TELL ME WHAT YOU THINK

- Is a 50% label share acceptable as proof of concept?
- Verbal (“likely”) or numeric (“80%”) uncertainty?

References

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made. PNAS Nexus, 3(10), pgae403.

Freeze, M., et al. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. Political Behavior, 43(4), 1433–1465.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. EJIS, 22(3), 336–359.

Pawelczyk, F., Dimmery, D., & Yan, P. (2026). Implied authenticity effect? Proc. ICWSM, 20(1), 1738–1766.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect. Management Science, 66(11), 4944–4957.

Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications, 250, 123941.

Poster & paper

Theuerkauf, R. (forthcoming). Towards designing content labels under detection uncertainty: A taxonomy and experimental research agenda. *Workshop on e-Business (WeB 2026)*, Lisbon.