Emotional Optimization by Newsroom AI Needs Behavioral Evaluation

Bence Bago[1] and Jean-François Bonnefon[2]

[1]Department of Social Psychology, Tilburg University, Tilburg, NL
[2]Toulouse School of Economics, CNRS (TSM-R), Toulouse, France

Standfirst

When newsroom AI engages in emotional packaging, it changes how a story feels, shaping whether people approach, trust, and share it. As these systems are designed and deployed, AI and behavioral researchers should work together to test whether emotional packaging helps accurate information reach readers, and whether it deepens division.

Emotional packaging as an optimization target

Major news organizations including the BBC (*1*), Financial Times (*2*), *The Economist* (*3*), and *The New York Times* (*4*) have already integrated Artificial Intelligence (AI) into their operations, using AI systems for tasks ranging from summarization to drafting headlines. As these systems scale, they are poised to reshape not only what news reaches audiences, but how that news is made to feel. Newsroom AI may soon optimize headlines, summaries, tone, or imagery by predicting the emotional reactions they will elicit in different readers. The same core facts may thus be packaged to elicit more curiosity, reassurance, threat, anger, or outrage, depending on which reaction is expected to increase engagement. Here we show why emotional optimization by newsroom AI

is a joint challenge and opportunity for the behavioral and machine intelligence communities. We focus on two systemic outcomes of emotional optimization (See Box 1): epistemic polarization, when different audiences increasingly live in different informational environments and endorse different beliefs; and affective polarization, when groups develop increasingly negative sentiments toward one another.

Attention to polarizing outcomes is not new. Research on social media and recommender systems has established a great deal about how platforms select, rank, and distribute already-produced content, and how engagement based on this distribution can contribute to polarization (*5*). However, emotional optimization by newsroom AI concerns a different point in the information pipeline. It occurs upstream of distribution: AI changes the emotional packaging of a given factual story before that story is ranked or delivered. The relevant effects are then observed downstream of distribution, at the engagement stage, where the package can influence whether readers approach or avoid the story, what they expect to feel, whether they trust it, and how they respond to the groups it concerns. This upstream intervention with downstream behavioral effects creates a distinct object of study. The same factual content can produce different reader responses depending on how it is packaged before distribution. AI makes this possibility scalable and systematic by allowing newsrooms to learn emotional cues from audience data, generate alternative packages, and optimize them across stories and audience segments. Thus, findings from research focused primarily on distribution cannot simply be transferred to emotional optimization at the generation stage. Likewise, technical and policy interventions designed to steer distribution toward beneficial outcomes cannot simply be transplanted to the generation stage, because they act on different mechanisms and points in the causal chain.

Emotional packaging is not an entirely new phenomenon either. Editors have always chosen headlines, images, and ledes with some expectation about how readers will react. What changes with AI is the possibility of learning these expectations from audience data, optimizing them at scale, and personalizing them across segments. In other words, emotional packaging can become less like an editorial craft and more like an objective function. We summarize recent behavioral findings suggesting that conditional on how such emotional optimization is performed, society will move toward greater or lesser polarization; and we consider how these findings set the stage

Societal Outcomes: Epistemic and Affective Polarization
We consider how emotional optimization by newsroom AI can bridge or deepen social divides, by reducing or amplifying two forms of polarization.
Epistemic polarization
Increasing disjunction in the information different groups see or seek, leading them to live in different epistemic ecosystems and endorse different beliefs. For example, climate skepticism is self-reinforcing because skeptics select out of exposure to climate science news (*6*).
Affective polarization
Increasingly negative attitudes between groups, including colder feelings, lower trust, and elevated levels of anger, fear, or contempt. For example, selective exposure to like-minded media is associated with greater dislike across partisan lines (*7*).

Box 1

for a productive collaboration between behavioral scientists and machine intelligence scientists. Our goal is not to tell machine intelligence scientists what systems they should build. Rather, we argue that behavioral science can help specify the outcomes that should be measured when emotional reactions become optimization targets.

How engagement learning can polarize

We first consider how emotional optimization may increase epistemic polarization by strengthening the affective cues people use to decide which information to engage with in the first place. Epistemic polarization is not only driven by what information is available to people, but by what information people are willing to approach. Tone, imagery, or headlines can be cues for readers to anticipate how an article will make them feel, and to decide whether to click or disengage. By tuning these affective cues, emotion-aware AI systems can shape who is exposed to what information.

This matters because the affective cues that govern approach and avoidance are learned from the audiences that provide the feedback. When emotion-aware AI systems are trained on engagement data from homogeneous readerships, they learn which

tones, imagery, and headlines work for that audience, and then intensify them. The resulting emotional packaging increases resonance within the core audience while further repelling readers who were already less likely to engage. Consider climate change coverage. Research shows that climate skeptics actively avoid scientific information which would challenge their emotions, beliefs, or behaviors (*6*). If an AI system is trained on engagement patterns from a predominantly climate-concerned audience, it may intensify affective cues (including signals of doom and moral condemnation) that maximize engagement among the climate-concerned, while making skeptics less willing to approach the content at all. The result is a widening gap in who encounters high-quality information about climate change, further excluding the audience who would benefit the most from exposure to the scientific consensus.

The same mechanism can amplify affective polarization, with a different outcome: not unequal exposure to facts, but increasingly hostile sentiment between groups. Systems trained on core-audience engagement may discover that clicks come more easily from fear, outrage, or anger directed at another group, either by leaning on the audience's negative perception of that group(*8*) or on its negative meta-perception, that is, how negatively it believes it is judged by them (*9*).

The feedback loop is then straightforward: as divisive emotional packaging performs better, newsroom AI systems learn to produce more of it, wrapping facts in hostile or sensational language, raising the ambient level of anger and contempt across group lines and, over time, intensifying affective polarization. These dynamics can be further amplified once such content enters social media recommendation systems, which propagate it based on their own emotion signals (*10*), generating cascades that exceed any single newsroom's intentions or control. This is not a claim that every newsroom system will increase polarization. It is a claim that aggregate engagement can be a misleading success metric, because the same increase in engagement may be produced by expanding cross-cutting exposure, or by increasing antagonistic arousal inside an already homogeneous audience.

Behavioral evidence for depolarizing packaging

The same mechanism by which emotional optimization can deepen epistemic and affective polarization also creates a path to bridging divides. Rather than amplifying affective cues that repel and antagonize readers outside a core audience, emotion-aware AI could attenuate these cues in order to bridge over epistemic and partisan lines. Our own research provides direct evidence for epistemic depolarization by emotion-aware AI (*11*). In a controlled experiment with 2,000 US participants, we used an open-weights large language model to rewrite headlines of climate-related news articles so that they would feel less threatening to skeptical readers while remaining accurate to the article's core content. For example, *'Aircraft Turbulence is Worsening with Climate Change'* became *'Birds May Hold the Key to Predicting Turbulent Skies,'* a headline that remained true and relevant to the factual claims made in the article.

Such reframings altered the emotional expectations of skeptics, decreasing their anticipations of disagreement, regret, and other negative affects; and the more they did, the more likely skeptics were to engage with the news articles. Importantly, this impact was highest on the most skeptical participants, those otherwise most prone to information avoidance, with no corresponding drop in engagement of participants who already accepted climate change. In sum, emotion-aware AI was able to reduce the defensive reactions of skeptics and their information avoidance behavior, at no cost to the engagement of other readers, without relying on ethically questionable tactics such as false balance or both-sideism that might legitimize unfounded skepticism. This experiment illustrates the contribution behavioral science can make: the outputs of the system were embedded in a causal test of human reactions, providing the evidence needed to decide whether an optimization strategy carries negative social externalities beyond its performance on engagement.

While there is no comparably direct evidence that emotion-aware newsroom AI can achieve affective depolarization, there is indirect evidence that exposure to less hostile emotional cues can depolarize. In a large-scale, preregistered field experiment (*12*), participants used a browser extension which altered their X feed to downrank posts expressing partisan animosity for one week. This intervention led to warmer feelings toward their political outgroup, with no detectable partisan asymmetry. Al-

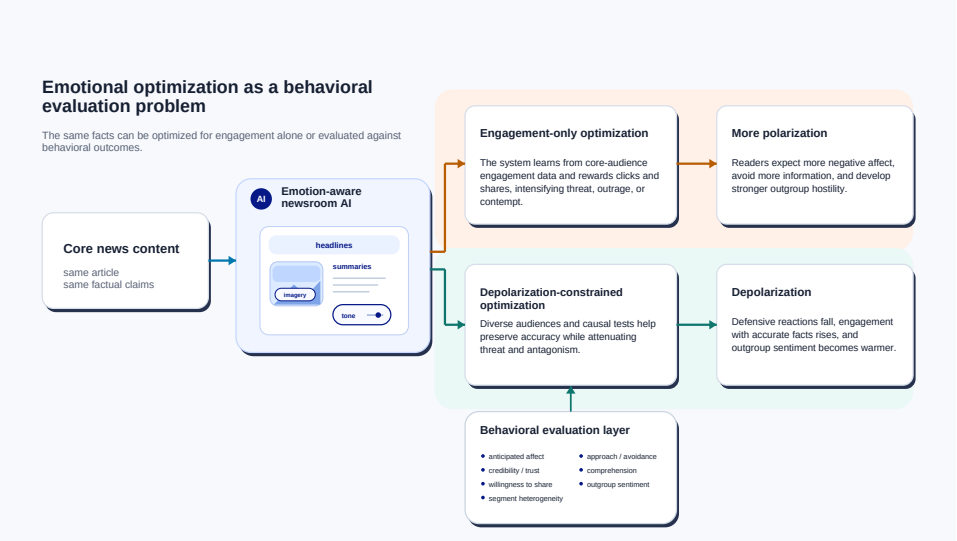

Figure 1: Conceptual overview of emotional optimization by newsroom AI. The same core news content can be packaged through headlines, summaries, tone, and imagery in ways that either intensify engagement-driven affective cues and increase polarization, or preserve accuracy while attenuating threat and antagonism. Behavioral evaluation can test these effects on anticipated affect, approach or avoidance, trust, comprehension, sharing, outgroup sentiment, and heterogeneous audience responses.

though this intervention acts on feed ranking rather than on the emotional packaging of individual items, this can serve as indirect evidence of what emotional packaging could achieve: shifting the affective tone of what people routinely encounter can move outgroup sentiment. Instead of relying only on downstream reranking to reduce hostile content, emotion-aware newsroom AI could directly attenuate antagonistic cues in headlines and summaries, reducing the probability that political news is experienced as contemptuous or enraging, and thereby creating conditions for affective depolarization.

What behavioral science can bring

The first contribution of behavioral science is to clarify the specific emotional reactions that matter for polarization, beyond the umbrella term of "emotion". Curiosity, anxiety, anticipated regret, moral condemnation, contempt, and feelings of humiliation are not interchangeable in their effects on defensive avoidance or increased social hostility. A behavioral evaluation of emotional optimization therefore needs to specify which affective states are being targeted, attenuated, or amplified, and why.

The second contribution is outcome selection. If the outcome is only aggregate engagement, an optimization can look successful while increasing polarization. A more informative evaluation would measure approach and avoidance, comprehension, perceived credibility and trust, anticipated disagreement, willingness to share, outgroup warmth or contempt. Different studies will not need all of these outcomes, but they should make explicit which behavioral pathway they are testing. In the case of epistemic polarization, the key pathway may be that attenuating threat increases approach to accurate information. In the case of affective polarization, the key pathway may be that attenuating antagonistic cues reduces anger and contempt across group lines. Emotional packaging may also have outcomes that are separate from polarization. Repeated exposure to threatening or highly negative news may affect readers' mood, anxiety, or well-being (*13*), even when it does not change what information they receive or how they feel about other groups. These outcomes need not be measured in every study, but they may be important when a system changes the emotional tone of news. In such cases, researchers could assess mood, anxiety, or depressive symptoms alongside the study's primary outcomes.

The third contribution is segment-level evaluation. Emotional optimization can produce different results for different audiences, and its effects should be analyzed accordingly. Our climate headline experiment illustrates this: the strongest benefit appeared among the most skeptical participants, and an aggregate click-through metric could have diluted it into a small average. The reverse is also possible, with an aggregate benefit hiding the fact that one segment is being systematically repelled or antagonized. Behavioral science is well equipped to design these heterogeneity tests and to specify in advance which segments are theoretically relevant.

The fourth contribution is causal evaluation. Publicly observable headlines and social media traces can reveal useful patterns, but they cannot by themselves establish what a newsroom system is optimizing for, or whether a different packaging would have produced different effects. Behavioral experiments can create the missing counterfactuals. They can compare original headlines, engagement-optimized headlines, and depolarization-constrained headlines for the same articles, then estimate their causal effects on different audiences. This is where collaboration with machine intelligence researchers becomes especially valuable: controllable generative systems can produce the candidate variants, while behavioral experiments can test what those variants do to people.

Toward collaborative evaluation infrastructure

Even without privileged access to newsroom AI systems, researchers can already evaluate whether emotional packaging could be optimized toward depolarization. Using publicly available articles together with their headlines, summaries, and imagery, researchers can quantify antagonistic affect cues (e.g., outrage, threat, contempt) and test whether alternative, accuracy-preserving packaging can attenuate those cues. They can then run controlled experiments to estimate how such packaging changes information avoidance, trust, and outgroup sentiment across audience segments. Browser extensions and custom news clients can extend such tests into the field, re-rendering headlines and summaries for consenting participants to preserve ecological validity without requiring cooperation from newsrooms or platform. However, this outside view has clear limits. While it can prove the possibility of depolarization, it cannot reliably establish what newsroom AI is currently optimizing for, based on which training data and reward functions. Nor can it distinguish emotion-aware optimization from ordinary editorial variation, unrelated A/B testing practices, or platform-driven differences in distribution. In other words, external research can benchmark feasible depolarizing designs and estimate their effects, but it cannot make string claims about real-world deployment.

This is where an infrastructure for collaboration would help. A useful model comes from emerging EU work on a voluntary code of practice for the transparency of AI-

generated content (*14*). The draft code sketches a transparency infrastructure by asking signatories to provide a verification interface, compliant with data protection rules, that legitimate parties (including researchers) can use to check whether content was generated or altered by their AI systems. While that interface would only allow users to verify that content has been AI-altered (and not how), a similar architecture could be adapted to emotional optimization.

In particular, a standardized, privacy-preserving interface could allow qualified third parties to request shadow-mode outputs for a pre-specified sample of articles, together with minimal documentation of the objective and constraint used. What we mean by shadow mode is that, for a given sample of articles, the newsroom or vendor would run the same system twice: once under its deployed objective (for example, maximize engagement) and once under the same objective conditional on attenuating antagonistic cues. The second output would be generated solely for evaluation, providing independent researchers with a clean counterfactual benchmark. This would not require newsrooms to disclose model weights, training data, or user-level engagement logs. It would only require the ability to observe paired outputs under controlled conditions, and to test their likely behavioral effects.

Such infrastructure would make the collaboration between behavioral science and machine intelligence more concrete. Machine intelligence researchers can make emotional packaging more controllable, more faithful to source material, and more robust. Behavioral researchers can test whether these controls have the intended effects on anticipated affect, information avoidance, trust, and outgroup sentiment. Emotional optimization sits exactly at the boundary between controllable generative systems and measurable human reactions, which means that it will allow for evaluation methods that are jointly technical and behavioral.

Acknowledgments

JFB acknowledges support from grant ANR-17-EURE-0010, grant ANR-23-IACL-0002, grant ANR-25-APEV-0001 and the research foundation TSE-Partnership

Statement of competing interests

The authors declare no competing interests.

References

1. BBC, *What we're doing with AI* (2026; https://www.bbc.co.uk/aboutthebbc/reports/policies/what-were-doing-with-ai/).
2. Financial Times, *Our approach to using AI* (2026; https://aboutus.ft.com/company/our-standards/ai).
3. The Economist, *How are we using AI?* (2026; https://myaccount.economist.com/s/article/How-we-handle-AI-generated-content).
4. The New York Times, *How The New York Times uses A.I. for journalism* (2026; https://www.nytimes.com/2024/10/07/reader-center/how-new-york-times-uses-ai-journalism.html).
5. S. Milli *et al.*, Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus* **4**, pgaf062 (2025).
6. T. P. Newman, E. C. Nisbet, M. C. Nisbet, Climate change, cultural cognition, and media effects: worldviews drive news selectivity, biased processing, and polarized attitudes. *Public Understanding of Science* **27**, 985–1002 (2018).
7. E. Kubin, C. Von Sikorski, The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* **45**, 188–206 (2021).
8. M. Falkenberg, F. Zollo, W. Quattrociocchi, J. Pfeffer, A. Baronchelli, Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nature Communications* **15**, 9560 (2024).
9. J. Lees, M. Cikara, Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B* **376**, 20200143 (2021).

10. W. J. Brady *et al.*, Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature human behaviour* **7**, 917–927 (2023).

11. B. Bago, P. Muller, J.-F. Bonnefon, Using generative AI to increase sceptics' engagement with climate science. *Nature Climate Change* **15**, 1176–1182 (2025).

12. T. Piccardi *et al.*, Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science* **390**, eadu5584 (2025).

13. M. Skovsgaard, K. Andersen, Conceptualizing news avoidance: Towards a shared understanding of different causes and potential solutions. *Journalism studies* **21**, 459–476 (2020).

14. "Second Draft Code of Practice on Transparency of AI-Generated Content" (European Commission, AI Office, Mar. 5, 2026), (`https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content`).