Harvard Kennedy School Misinformation Review[1]
July 2026, Volume 7, Issue 3, Special Issue on Fact-Checking

DOI: https://doi.org/10.37016/mr-2020-197
Website: https://misinforeview.hks.harvard.edu

Commentary

Fact-checking what matters: How a harms-based model for selecting claims works

Not all misinformation consequences are equal. Faced by hundreds of thousands of false claims online and offline every day, fact checkers need a robust way to identify the important ones to check. This scalable model—used by fact checkers in trials in Europe, Africa, and the Middle East since 2024—helps forecast the potential imminent and cumulative harms of different false claims and is an early warning system for society that focuses efforts on factually false claims that cause real-world harms. This is how it works.

Authors: Peter Cunliffe-Jones
Affiliations: Communication and Media Research Institute, University of Westminster, UK
How to cite: Cunliffe-Jones, P. (2026). Fact-checking what matters: How a harms-based model for selecting claims works. *Harvard Kennedy School (HKS) Misinformation Review*, *7*(3).
Received: January 16th, 2026. Accepted: April 4th, 2026. Published: July 7th, 2026.

Identifying broad and specific misinformation effects

Accurate forecasting of any complex phenomena requires a combination of a robust model and good data (Benjamin et al., 2018). As a visiting researcher at the University of Westminster between 2021 and 2024, I developed a model for fact checkers, researchers, policymakers, and platforms to assess the potential of specific examples of false or misleading information to cause or contribute to specific substantive consequences or harms.

With many false claims in circulation, fact checkers and researchers have to choose what to check. The harms-risk assessment model, first outlined in my 2025 book, *Fake News – What's the Harm?*, uses a combination of findings from prior studies on the effects of false information in particular fields (Ecker et al., 2024; Loftus, 2005; Thielman et al., 2014) and empirical evidence in fact checks on the nature of a particular claim, the audience's responses to it, and the context in which the consequences might occur, to identify the potential of different claims to contribute to specific harms.

The model—used since 2024 by fact checkers in a series of studies and trials in Europe, Africa, and the Middle East (AFCN, 2025; CAMRI, 2025)—recognizes at the outset the potential of all substantively false or misleading information to contribute at some incremental level to the broad effects of misinformation and disinformation on public trust and cohesion (Hameleers et al., 2020; Lynch, 2016). Going beyond this, it applies evidence of key differences between the topic, nature, source, size, and characteristics of the audience or subject, and the context of false claims that caused or contributed to real-world consequences, and those of outwardly similar claims which caused no such ill effects to identify the

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

potential for specific harms. In 2026, the model is being used in a study for the UK AI Security Institute on the effects of AI-generated misinformation.

Specific effects: A predictive model built around three questions

The harms-risk assessment model deployed in these studies and trials uses a dual key approach to assess the potential of particular false claims to cause or contribute to substantive consequences or harms. The model defines *substantive consequences* or *harms* as objectively verifiable substantive injuries to the interests of an individual, group, or wider society. These include negative effects on physical health; violence, unrest, and war; the distortion of public policy; or negative impact on businesses or the economy. Neither a false understanding nor a fleeting emotional response to false information meets the necessary level. Thus, first, evidence from existing research embedded in the model (Cunliffe-Jones, 2025, pp. 208–228) must show that such claims have caused or contributed to such harms in the past. Second, empirical evidence about the claim, audience, and context must show that the conditions for these harms to occur exist.

The Cunliffe-Jones (2025) study examined 250 factual claims that fact checkers had identified as in some way inaccurate or misleading and identified key differences between false claims that did cause or contribute to harmful effects and outwardly similar claims that did not. Based on the differences identified, the model distinguishes between claims that may or may not have the substantive potential to cause such a real-world harm by asking the following three questions in succession.

Whether the claim, if believed, causes a substantively false factual understanding

To determine a substantive potential for harm, the first task for fact checkers using the model is to determine—through analysis of content, context, and audience responses—the effect of a claim, if believed, on the audience's factual understanding of the event or situation. A claim is found to be either accurate or only narrowly inaccurate when (1) either all elements of the claim are proven to be correct and in context, or (2) those elements that are incorrect are immaterial to the understanding the claim causes or so marginally inaccurate that they do not affect the essential understanding formed. For example, a news report about an air crash might misidentify the exact type of plane involved, but could still create an accurate understanding that a plane has crashed, why, and how many people were killed because the type of plane was not material to the key understanding created. By contrast, if the key elements of the claim are substantively incorrect or are correct but are distorted or taken out of context in a way that causes a false understanding, the claim would be found to be substantively false. For example, a report that a healthy person has died or misidentifying the winner of an election creates a substantively false understanding. The key question is the effect on the accuracy of understanding if the claim is believed.

Whether the claim is seen as true by enough people to cause specific harm if they act on it

Where evidence in the fact check shows that an inaccurate claim is only narrowly inaccurate, fact checkers using the model would record the claim as having no substantive potential to cause specific substantive harm as misinformation (see Figure 1, finding *i). Where the evidence in the fact check shows that the claim is substantively false, the model finds that it adds at a tiny incremental level to reasons to distrust information. Or, if an accurate claim is misidentified as false, it is found to add at a similar level to the audience's undue distrust in information (see Figure 1, finding *ii). This does not, however, tell the user whether the claim has the potential to cause a more specific real-world harm.

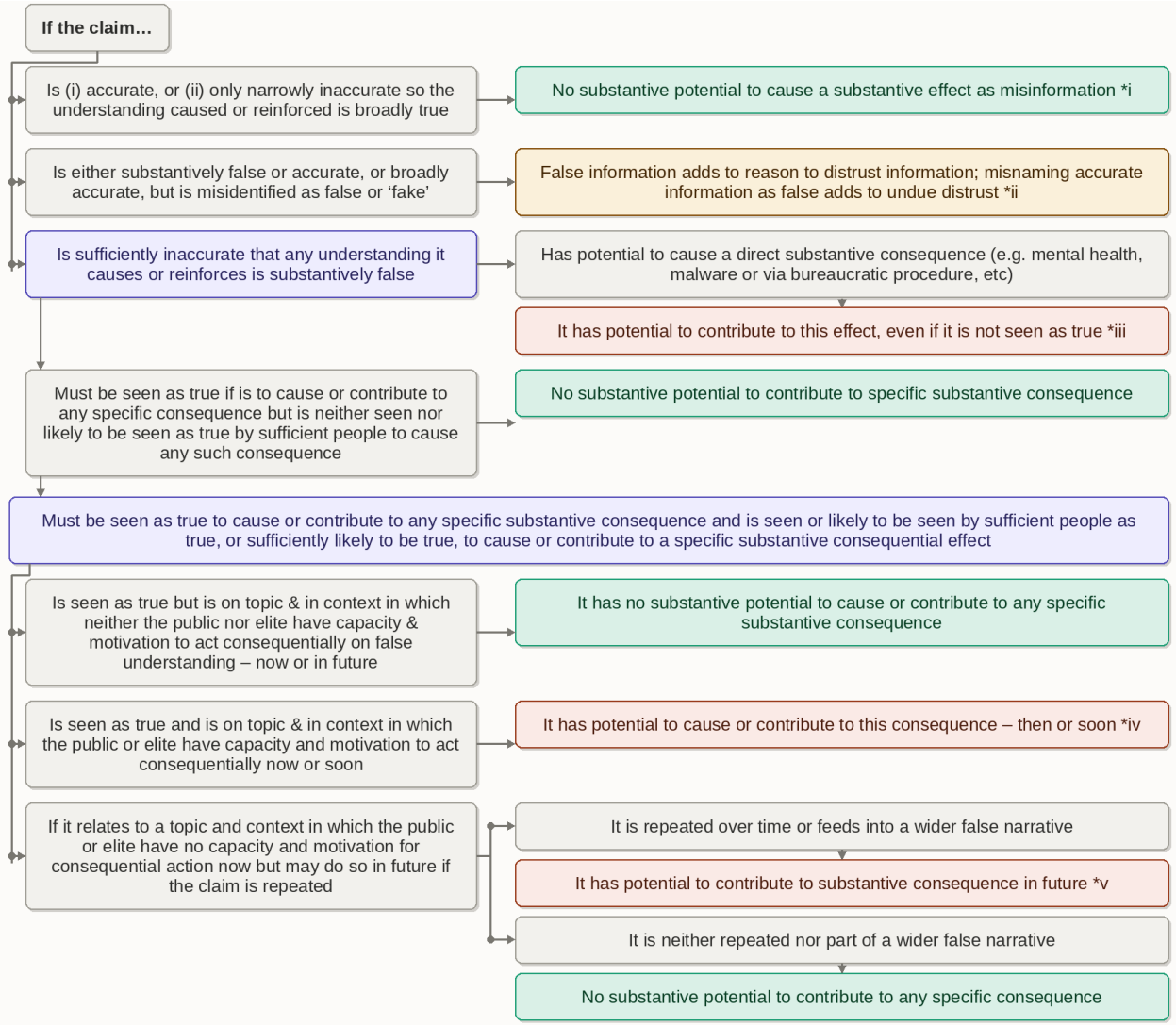

Figure 1. How three broad factors—(i) the degree of falsity, (ii) perceived accuracy, (iii) audience capacity & motivation to act—shape potential for substantive consequences.

To answer that question, the model next identifies whether the claim has the potential to have specific consequences even if it is not viewed as true by those who see it. In the case of some types of claims observed in the Cunliffe-Jones (2025) study, the model shows that, in particular circumstances, mere publication of a particular falsehood can harm the mental health of a vulnerable individual, be used as clickbait to spread a computer virus to a targeted audience, and/or can have a functional effect in a bureaucratic procedure, with no requirement that the claim be seen as true (see Figure 1, finding *iii). In these cases in the study, the mental health of people traumatized by earlier events and subsequently subject to false claims was affected by false claims, clickbait spread computer viruses, and bureaucratic processes adopted false numbers, all regardless of whether the audience believed the claims (Cunliffe-Jones, 2025, pp. 117–127). To identify where this might happen, the model takes account of factors such as the known history of the subject of the false claim that might make them more vulnerable to certain false claims and how certain data is used in certain functional processes.

In most cases, however, evidence shows that a false claim does have to be understood as true, or sufficiently likely to be true, for someone to act, in order to cause most types of consequences. In cases

seen in the Cunliffe-Jones (2025) study, when individuals sought to buy a fake medicine or rushed to protect children they wrongly believed were in danger, they did so not on a whim but because they believed the claim was true or sufficiently likely to be true for them to act. At the same time, the number of people needed to act to cause a given effect varies depending on the type of consequence concerned. In 2016, many thousands of Americans subscribed to the so-called "Pizzagate" conspiracy theory—a false belief that a pedophile network of high-powered individuals exploited vulnerable children detained in the cellar of a Washington DC pizza restaurant favored by the political elite. While many people believed the false claim, it only took one person to take action for an attack on the establishment to occur (Haag & Salam, 2017). By contrast, for a consumer boycott to significantly affect a major company's bottom line or for voters to affect an election outcome, thousands need to believe and act on it.

To determine whether sufficient numbers of people believe a claim to cause a specific consequence, fact checkers using the model should first assess the numbers—on a scale from an individual to dozens or hundreds to thousands—required to cause the consequence, then review the spread of and audience responses to the claim. They can do this using written or oral statements and online reactions, which have been shown to be reliable indicators of belief (Kosinski et al., 2013) and evidence of real-world behavior. Where such evidence is not available, users can review the claim for five factors, discussed in Cunliffe-Jones (2025; pp. 66–69) and identified in existing literature as increasing audience perceptions of a claim's credibility: (1) the plausibility (Madrid-Morales et al, 2021), (2) ease of processing (Powell et al, 2015), (3) repetition and coherence of the claim with an existing narrative (Fazio et al, 2022), (4) the claim's affective or likely affective impact (Brady et al, 2017), and (5) the perceived credibility of the source or perceived source of the claim (Traberg et al, 2024).

Whether those who believe the claim have the capacity and motivation to act on the false understanding it causes, either now or in the future, if this or similar claims are repeated

The fact that sufficient people believe a false claim to cause a particular consequence does not mean a consequence will follow. This requires that those who believe the claim to have both capacity and motivation to act in a way that would cause the effect, either then and there (see Figure 1, finding *iv) or in the future, if the claim is repeated over time (see Figure 1, finding *v). Fact checkers using the model could assess this based on both the context and the location and power of the audience to act consequentially on the false understanding caused and evidence of their motivation to do so.

Mapping harms that false claims may cause to individuals and/or society

One of the challenges for weather forecasters, meteorologist and mathematician Edward N. Lorenz reasoned in 1972, was the lack of both a robust model and good data on the variables for predicting future events. Describing what became known as "the butterfly effect"—the idea that, in a complex, dynamic system such as the weather, the "flap of a butterfly's wings in Brazil might cause a tornado in Texas"—Lorenz argued that: "We do not know how many butterflies there are, nor where they are all located, let alone which ones are flapping their wings at any instant" (Lorenz, 1972). Since then, however, what meteorologists call their "useful forecast skill" has been vastly improved by both better data and better modelling that takes account of multiple variables in ways unimaginable at the time (Benjamin et al., 2018).

The challenge of predicting human behavior depends, in some cases, on even more variables than the weather. However, the trials that fact checkers in Europe, Africa, and the Middle East have conducted of

the harms risk model since 2024 show good early potential to create better, more testable data than we have had before to measure the extent to which particular false claims (or butterflies) do, or have substantive potential to, cause or contribute to particular effects, who they might affect, and with what severity or duration. As a study of the 2024 trial, conducted by researchers at the University of Wisconsin-Madison and provided to the University of Westminster, found, "The process can help focus limited resources on the most urgent claims … (and) generates both case studies and aggregate data" (CAMRI, 2025). The model is "[t]he basis for developing a rigorous system that can be used by fact-checking organisations, researchers, platforms and others to identify false information that is potentially harmful to individuals or society" (CAMRI, 2025).

The researchers found that the process of identifying the potential harms of false claims is useful in focusing fact checkers' efforts on more consequential claims. Second, the work of mapping the specific proven and potential real-world effects of specific types of mis/disinformation could create an early warning system to help counter real harms. Third, this work can help, at the same time, to identify and prevent cases of over-zealous regulation of false content that can be robustly identified as having no substantive potential for harm.

Questions still to be answered and how to test the model

The judgements required in predicting the potential effects of different forms of false information are complex, and ensuring consistency across individuals and organizations is a challenge. One goal for the model in 2026 is to widen the number of experts drawn from different disciplines, perspectives, and geographies who can take the time to review both the theory of the model and its inner workings: the criteria that fact checkers are asked to apply and the consistency with which they do so. Another is for researchers to independently compare the predictive data produced and evidence of real-world outcomes. Identifying the potential for violent social unrest, harms to health, or even the economy that false claims cause or contribute to is a useful step toward mitigating or countering those harms.

Bibliography

Arab Fact Checkers Network (AFCN). (2025, December 7). *AFCN conducted "Fact-Checking Harmful Information" training at ARIJ25 Forum*. https://arabfcn.net/en/events-ar-en/2025/12/07/afcn-conducted-fact-checking-harmful-information-training-at-arij25-forum/

Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., & Schlatter, T. W. (2018). 100 years of progress in forecasting and NWP applications. *Meteorological Monographs 59*(1), 13.1–13.67. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0020.1

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

CAMRI. (2025, April 2). Trial finds predictive model helps fact checkers identify false claims with potential to cause harm. *Communications & Media Research Institute.* https://camri.ac.uk/blog/2025/04/02/trial-finds-predictive-model-helps-fact checkers-identify-false-claims-with-potential-to-cause-harm/

Cunliffe-Jones, P. (2005) *Fake news: What's the harm?* University of Westminster Press. https://doi.org/10.16997/mpub.14614695

Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*, 29–32. https://doi.org/10.1038/d41586-024-01587-3

Fazio, L., Pillai, R., & Patel, D. (2022). The effects of repetition on belief in naturalistic settings. *Journal of Experimental Psychology*, *151*(10), 2604–2613. https://doi.org/10.1037/xge0001211

Gichuhi, G. (2019, July 12). *Kenyan gospel music star Ringtone Apoke is alive and well*. Africa Check. https://africacheck.org/fact-checks/meta-programme-fact-checks/kenyan-gospel-music-star-ringtone-apoko-alive-and-well

Haag, M., & Salam, M. (2017, June 22). Gunman in 'Pizzagate' shooting is sentenced to 4 years in prison. *The New York Times*. https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html

Hameleers, M., van der Meer, T., & Brosius, A. (2020, May 31). Feeling "disinformed" reduces compliance with COVID-19 guidelines. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-023

Kosinski, M., Stilwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*(15), 5802–5805. https://www.pnas.org/doi/10.1073/pnas.1218772110

Kulundu, M. (2019, September 13). *Zambian farmer 'greatly distressed' by posts falsely identifying him as victim of South African violence*. AFP Fact Check. https://factcheck.afp.com/zambian-farmer-greatly-distressed-posts-falsely-identifying-him-victim-south-african-violence

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning and Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

Lynch, M. P. (2016, November 28). Fake news and the internet shell game. *The New York Times*. https://www.nytimes.com/2016/11/28/opinion/fake-news-and-the-internet-shell-game.html

Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanu, E., Tully, M., Umejei, E., & Uzuegbunam, C. (2021). Motivations for sharing misinformation. A comparative study in six sub-Saharan African countries. *International Journal of Communication*, 15, 1200–1219. https://ijoc.org/index.php/ijoc/article/view/14801/3378

Powell, T. E., Boomgaarden, H. G., de Swert, K., & de Vreese, C. H. (2015). A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication*, *65*(6), 997–1017. https://doi.org/10.1111/jcom.12184

Thielman, N., Ostermann, J., Whetten, K., Itemba, R., Itemba, D., Maro, V., Pence, B., & Reddy, E. (2014). Reduced adherence to antiretroviral therapy among HIV-infected Tanzanians seeking cure from the Loliondo healer. *Journal of Acquired Immune Deficiency Syndrome*, *65*(3), 104–109. https://doi.org/10.1097/01.qai.0000437619.23031.83

Traberg, C. S., Harjani, T., Roozenbeek, J., & van der Linden, S. (2024). The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, 14, Article 4205. https://doi.org/10.1038/s41598-024-54030-y

Acknowledgements
The author would like to thank Arwa Kooli and Fatima Bani Ahmad (Arab Fact-Checkers Network), Cayley Clifford (Africa Check), Estelle Peard (AFP), Joseph O'Leary (Full Fact), and all others who participated in the first trials of the model in 2024 and 2025.

Funding
Initial funding was received in 2020 from the Facebook Journalism Project, Google News Initiative, and Luminate to support the research. They neither sought nor had any influence over the scope or findings of the research. Further support was received from Google News Initiative in 2024 for one of the trials.

Competing interests
The author declares no competing interests.