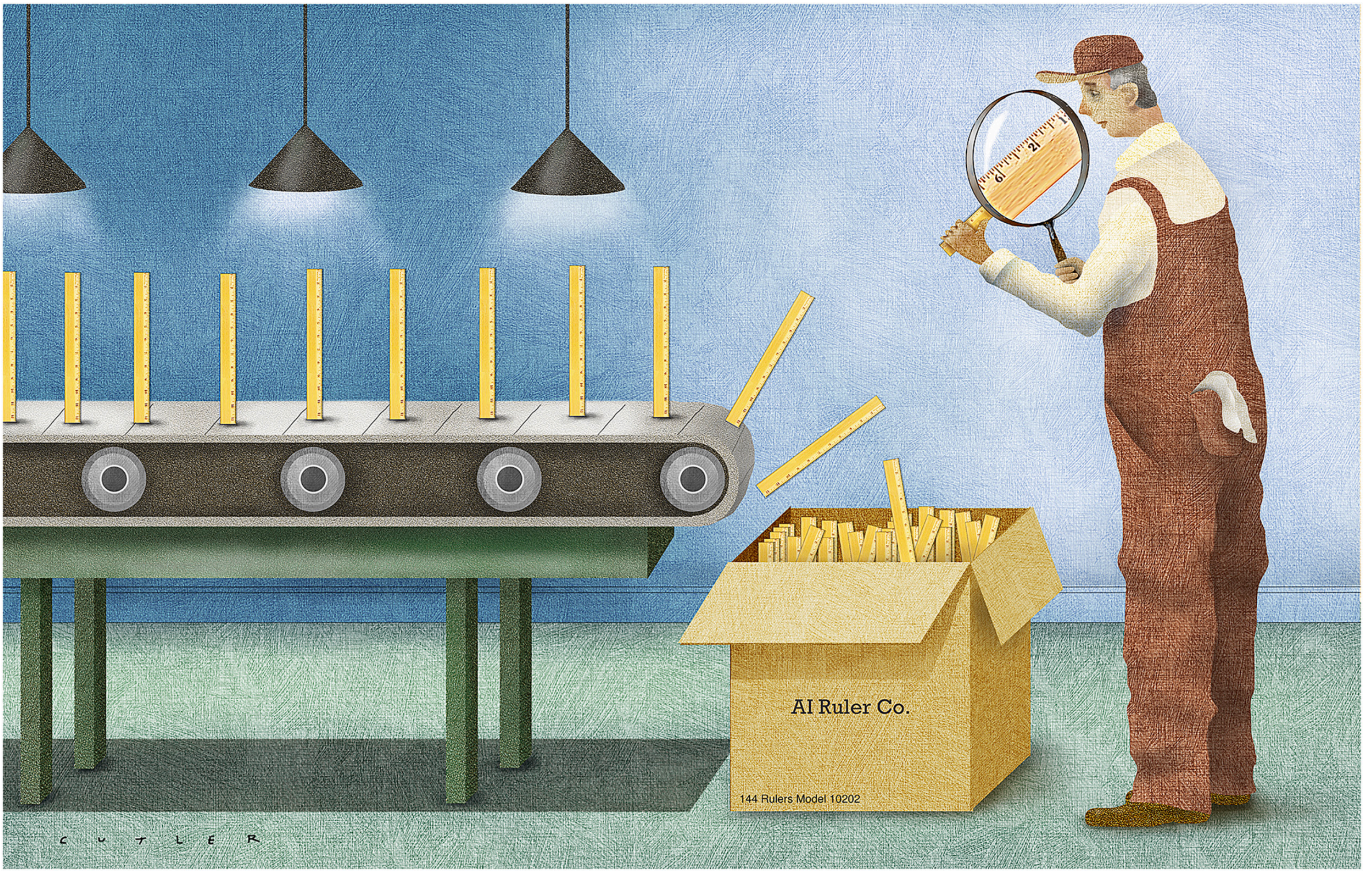
PNAS OPINION

The "synthetic validity" problem in AI-generated science

Jacob D. Teeny[a,1], Andrew Luttrell[b,1], and Dongchan Lee[a,1]



Large language models may at first blush provide convincing results. But does AI-generated content genuinely embody the concept it is meant to represent? Across multiple fields of science, researchers are beginning to draw conclusions from AI-generated materials whose fidelity to the concepts under study hasn't been adequately verified. Image credit: Dave Cutler (artist).

Already, artificial intelligence (AI) is assisting scientists—searching the literature, drafting code, and analyzing data. Increasingly, however, it is manufacturing the raw materials of science itself. Chemists select drug candidates proposed by generative models. Medical researchers train diagnostic software on synthetic patient records. Educators deploy exam questions written by chatbots. And across the behavioral sciences—psychology, political science, communication, and beyond—researchers now use AI to write the experimental stimuli, survey questions, and data-coding schemes on which their scientific studies are built.

The appeal is obvious: AI produces in minutes what once took research teams months, promising to advance science at a scale no human team could match. But does AI's output really capture what researchers intend?

There are well-publicized concerns about "AI hallucinations"—fabricated content presented as convincingly real. For example, an analysis found that 93% of ChatGPT-generated medical references were fabricated or inaccurate (1). However, we argue that AI presents a more insidious risk for science than hallucinations—namely, AI-generated content that *resembles* the concept under study rather than accurately *representing* it. For example, if a communication scholar asks AI to "write a persuasive message tailored to introverts," it will produce something in response. Whether that message reflects psychologists' nuanced understanding of introversion—or just a stereotype about staying home on a Friday night—is another matter entirely.

At the core of this concern is what science has long referred to as "validity"—that is, whether its materials and measures truly capture what they claim to study. We contend that current AI tools call for attention to a new form of validity that we term *synthetic validity*. This is evidence that AI-generated content genuinely embodies the concept it is meant to represent. The worry is akin to an astronomer with

Author affiliations: [a]Marketing Department, Kellogg School of Management, Northwestern University, Evanston, IL 60208; and [b]Psychology Department, Ball State University, Muncie, IN 47306

Author contributions: J.D.T., A.L., and D.L. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jacob.teeny@kellogg.northwestern.edu, alluttrell@bsu.edu, or dongchan.lee@kellogg.northwestern.edu.

Published September 9, 2026.

an uncalibrated telescope. The images may look intriguing or even compelling, but an uncalibrated lens distorts what it is showing, such that no astronomer trusts the view before checking the instrument. Across multiple fields of science, however, researchers are beginning to draw conclusions from AI-generated materials whose fidelity to the concepts under study has been minimally verified.

Although troubling for research across the sciences (e.g., AI-proposed drug molecules that prove impossible to synthesize), we suspect that this issue is especially relevant to the social sciences. These fields require scholars to translate abstract social concepts into concrete, measurable responses that cleanly isolate those concepts from other, often co-occurring experiences.

AI tools stand to industrialize these problems, taking humans out of the process and luring researchers into hastily verified conclusions at large scales.

And the challenge is by no means unique to the AI era. Even years ago, an audit found that nearly half the measurement scales in a leading psychology journal were used without cited evidence of their accuracy (2); another, that more than 40% of experimental procedures had never been validated beyond seeming reasonable to their designers (3). AI tools stand to industrialize these problems, taking humans out of the process and luring researchers into hastily verified conclusions at large scales.

Personalized Persuasion: A Case Study

Consider an area of research in which we, the authors, have particular expertise: personalized persuasion. This is the practice of making a message more influential by tailoring it to the psychology of its recipient, such as advertising a car's adventurousness to thrill-seeking consumers, while advertising its safety to security-minded ones (4, 5). Decades of experiments support the specific cognitive mechanisms by which this works. Personalized content is more attention-drawing and easier to process, moving the recipient's opinion closer to what is advocated in the message. However, these psychological processes only activate when a message speaks to the deeper reasons people hold their opinions (6)—their motivations, personal values, and characteristic ways of thinking—and fail when the message mismatches or only superficially references them (4).

Recently, however, high-profile studies have used AI to test personalized messaging at scale and have reported results that seem to overturn established consensus. In these experiments, AI systems were prompted to compose messages tailored to individual voters or debate participants one-on-one using their personal information. Researchers found that personalization conferred little or no additional persuasive influence (7–9). These findings made headlines and offered regulators a reassuring implication: perhaps fears of AI-powered microtargeting in elections and advertising are overblown.

But consider these experiments' synthetic validity. In three of them (7, 8), the personalized messages were tailored to demographics, such as a person's age, gender, or geographic location—surface characteristics that no existing theory predicts should make tailoring persuasive (4). Indeed, upon reanalysis of some of these data, the "personalized" messages failed to feel personally relevant to a large sample of respondents (10). Unless the messages activate the psychological processes inherent to the advantages of personalization (i.e., increased attention and processing ease), even if researchers had directed AI to target psychologically meaningful characteristics, the appeal would not have been more effective than a generic one. And today's models, if instructed to write for a particular personality trait, will tend to produce generic or stereotypical content (11, 12) rather than grounding their arguments in psychological theory or the substantive issues that make a tailored message resonate.

AI-generated content will also go beyond a prompt in unpredictable ways, preventing clear conclusions about which aspects of that content drive observed effects. Indeed, generative systems rarely produce the same output twice: the same request, run again, returns content that differs in numerous, sometimes imperceptible ways. Suppose a researcher prompts AI to produce one campaign appeal for younger voters and another for older ones. If the latter comes back longer and written in a warmer register (more personal, more emotive), then any difference in persuasion might be due to length or register, not the tailoring itself. Without careful attention to the content deployed in a study, these differences (or other, unnoticed ones) could account for the results (or lack thereof) that scientists attribute to a broader theory. In other words, scholars may claim to discover basic principles when in fact the results are specific to unidentified idiosyncrasies in the AI-generated study content. Because research often speaks to live regulatory debates—for example, the regulation of algorithmic influence in political advertising—the stakes of synthetic validity extend far beyond the academic record.

Beyond the Science of Communication

These concerns are not confined to the psychology of persuasion. More broadly within behavioral science, AI is now creating images intended to evoke specific emotions (13), writing questionnaire items intended to measure personality traits (14), and sorting thousands of open-ended text responses into conceptual categories (15). Researchers presumably apply standard checks, but quick reviews rarely catch synthetic-validity failures. Imagine an AI-built questionnaire that attempts to measure the anxiety people feel when using social media. Its items might sound reasonable, and scores on it could correlate with stress, worry, and negative mood. By every convenient indicator, it seems fine. But it may not measure social media-fueled anxiety at all—only people's general distress about present-day life. It would be like asking AI to design and build a thermometer. It looks every bit like a thermometer, and its readings rise and fall with the day. But its casing absorbs the afternoon sun, so what it faithfully records is not the temperature of the room at all.

Notably, behavioral science is not alone in these threats. When Google DeepMind announced that an AI system had discovered hundreds of thousands of new stable crystals—hailed as an order-of-magnitude leap in materials science—chemists who inspected the haul found "scant evidence" that

the compounds were genuinely novel, credible, or useful (16, 17). When an autonomous robotic lab reported synthesizing dozens of new compounds, the paper was later corrected after chemists showed that most were simply known materials misidentified as new (18, 19). Even AlphaFold, the celebrated protein-structure predictor, comes with a warning from structural biologists that its outputs are "valuable hypotheses" that "do not replace experimental structure determination" (20). And even more recent research has shown that many AI-designed protein structures might at first seem viable but ultimately describe chemically impossible designs that would collapse or fail in the real world (21).

Across science, the pattern recurs. And now, AI systems are actually grading the output of other AI systems (22). In each case, content that resembles the real thing may be quietly substituting for content verified to be the real thing.

Our argument is not that researchers should abandon or discredit research using AI-generated materials. It is that these materials deserve the same scrutiny as any scientific instrument, if not more. A human researcher, however imperfect, brings an understanding of the deeper concept to the task. AI brings naïve pattern-matching.

Three Safeguards

Drawing on behavioral science's century-long tradition of asking whether instruments and stimuli reflect what they claim (3, 23, 24), we propose three safeguards, applicable in any field.

First, specify before generating. The instructions given to the AI must contain the broader theory, not just the topic. The prompt should define the concept, describe how the effect is supposed to work, and state which features of the output are of value and which must be held constant. A prompt intended to generate tailored messages, for example, should name the motivation it must engage; and one intended to generate drug molecules should name the molecule properties the user intends to preserve. Just as important is documentation—the exact prompts, the model and version used, its settings, and the rules for selecting or editing outputs. Without that record, no reviewer or future researcher can tell which decisions reflect the scientist's judgment and which reflect the machine's idiosyncratic moves.

Second, verify before deploying. Outputs must be independently checked before they enter a study—by human judges for subjective qualities and by computational or experimental tests for objective ones. Do independent raters interpret the materials as intended? Do a tailored message and its generic counterpart differ only in the intended ways? Do an AI's proposed discoveries survive a check against existing databases? Because the same prompt yields different outputs across runs, generating materials should be a cycle—generate, check, refine, regenerate—with a record of what was kept and why. AI makes each cycle faster; it does not make the cycle optional.

Third, make sure to validate against reality. Once results are in, researchers must confirm that the AI's materials did what they were meant to—judged against an external, real-world standard, not the AI's own confidence. A generated molecule or a predicted protein structure is only a hypothesis until it is synthesized and measured in the laboratory; an AI's subjective evaluations must be checked against human judgment. As noted, the reanalysis showing that "personalized" messages felt no more relevant than generic ones (10) was precisely this check. Across fields, the test is the same: a valid instrument should show its strongest effect on the thing it is meant to capture and progressively weaker effects the farther one moves from it (3).

Calibrating the Instruments

These safeguards matter most in combination because the failures compound. A vague prompt, unverified outputs, and lack of after-the-fact validation yield a study in which validity was assumed at every link in the chain and confirmed at none. Scale makes this worse: a million responses to content that does not capture the intended concept will generate less understanding than a hundred responses to content that captures it adequately. Yet, the former types of data are already being published, cited, and used to inform both science and policy—and the history of behavioral research shows how easily failed replications get blamed on theories when the study's materials were at fault (25).

One might counter that researchers who refine AI materials until they pass verification are stacking the deck, engineering their way to a preferred result. But this confuses engineering the outcome with engineering the instrument. Verifying that a message genuinely speaks to introverts' actual motivations does not guarantee that it will persuade introverts; it guarantees that whatever the experiment finds is informative for the theory. A calibrated telescope does not promise that the astronomer will find the star they hypothesize exists. It promises that what they see is really there.

None of these problems are new or are unique to AI. But AI transforms their scale and attracts researchers from fields with little tradition of instrument validation. Journals, reviewers, and funders can act now while conventions are still nascent. They could, for example, require researchers to document how they produced AI-generated tools and materials and show evidence of their validity, making this as routine as ethics approval or data availability is today.

Across science, AI-based instruments can clearly be powerful. They are poised to become standard equipment. But before trusting what they show us, we must carefully calibrate what exactly we are looking through.

1. M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, L. E. Miller, High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* **15**, e39238 (2023).
2. J. K. Flake, J. Pek, E. Hehman, Construct validation in social and personality research: Current practice and recommendations. *Soc. Psychol. Personal. Sci.* **8**, 370–378 (2017).
3. D. S. Chester, E. N. Lasko, Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspect. Psychol. Sci.* **16**, 377–395 (2021).
4. R. E. Petty, A. Luttrell, J. D. Teeny, Eds., *The Handbook of Personalized Persuasion: Theory and Application* (Routledge, 2025).
5. J. D. Teeny, J. J. Siev, P. Briñol, R. E. Petty, A review and conceptual framework for understanding personalized matching effects in persuasion. *J. Consum. Psychol.* **31**, 382–414 (2021).
6. K. Joyal-Desmarais *et al.*, Appealing to motivation to change attitudes, intentions, and behavior: A systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. *Psychol. Bull.* **148**, 465–517 (2022).
7. K. Hackenburg, H. Margetts, Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403116121 (2024).
8. L. P. Argyle *et al.*, Testing theories of political persuasion using AI. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2412815122 (2025).
9. H. Lin *et al.*, Persuading voters using human–artificial intelligence dialogues. *Nature* **648**, 394–401 (2025).
10. J. D. Teeny, S. C. Matz, We need to understand 'when' not 'if' generative AI can enhance personalized persuasion. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2418005121 (2024).

11. L. Zhang *et al.*, LLMs are introvert. arXiv [Preprint] (2025). https://arxiv.org/abs/2507.05638 (Accessed 1 July 2026).
12. R. Saumure, J. De Freitas, S. Puntoni, Humor as a window into generative AI bias. *Sci. Rep.* **15**, 1326 (2025).
13. H. T. Chiu, H. I. Sou, Y. W. Lam, C. S. F. Ng, S. W. Wong, Beyond traditional stimuli: Validating AI-generated images for eliciting negative emotions in affect research. *PLoS One* **21**, e0342434 (2026).
14. F. M. Götz, R. Maertens, S. Loomba, S. van der Linden, Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychol. Methods* **29**, 494–518 (2024).
15. N. Than, L. Fan, T. Law, L. K. Nelson, L. McCall, Updating 'The future of coding': Qualitative coding with generative large language models. *Sociol. Methods Res.* **54**, 849–888 (2025).
16. A. Merchant *et al.*, Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
17. A. K. Cheetham, R. Seshadri, Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. *Chem. Mater.* **36**, 3490–3495 (2024).
18. N. J. Szymanski *et al.*, An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
19. J. Leeman *et al.*, Challenges in high-throughput inorganic material prediction and autonomous synthesis. *PRX Energy* **3**, 011002 (2024).
20. T. C. Terwilliger *et al.*, AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods* **21**, 110–116 (2024).
21. G. I. Makhatadze, The accuracy of electrostatic interactions captured by AI protein structure prediction models. *Proc. Natl. Acad. Sci. U.S.A.* **123**, e2609610123 (2026).
22. A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, D. Hupkes, "Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges" in *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM2)*, O. Arviv *et al.*, Eds. (Association for Computational Linguistics, 2025), pp. 404–430.
23. J. Loevinger, Objective tests as instruments of psychological theory. *Psychol. Rep.* **3**, 635–694 (1957).
24. S. Embretson, J. Gorin, Improving construct validity with cognitive psychology principles. *J. Educ. Meas.* **38**, 343–368 (2001).
25. A. Luttrell, R. E. Petty, M. Xu, Replicating and fixing failed replications: The case of need for cognition and argument quality. *J. Exp. Soc. Psychol.* **69**, 178–183 (2017).