Trust the human, learn from the AI?

How source attribution influences trust, comprehension, and financial engagement in an increasingly AI-enabled world

Research by Voya's Behavioral Finance Institute for Innovation

VOYA® Plan. Invest. Protect.

Artificial intelligence (AI) is rapidly becoming part of how financial education and retirement-planning content is created and delivered. Across the industry, AI-generated materials already appear on financial wellness platforms and retirement websites. Sometimes AI involvement is disclosed, and sometimes it is not.

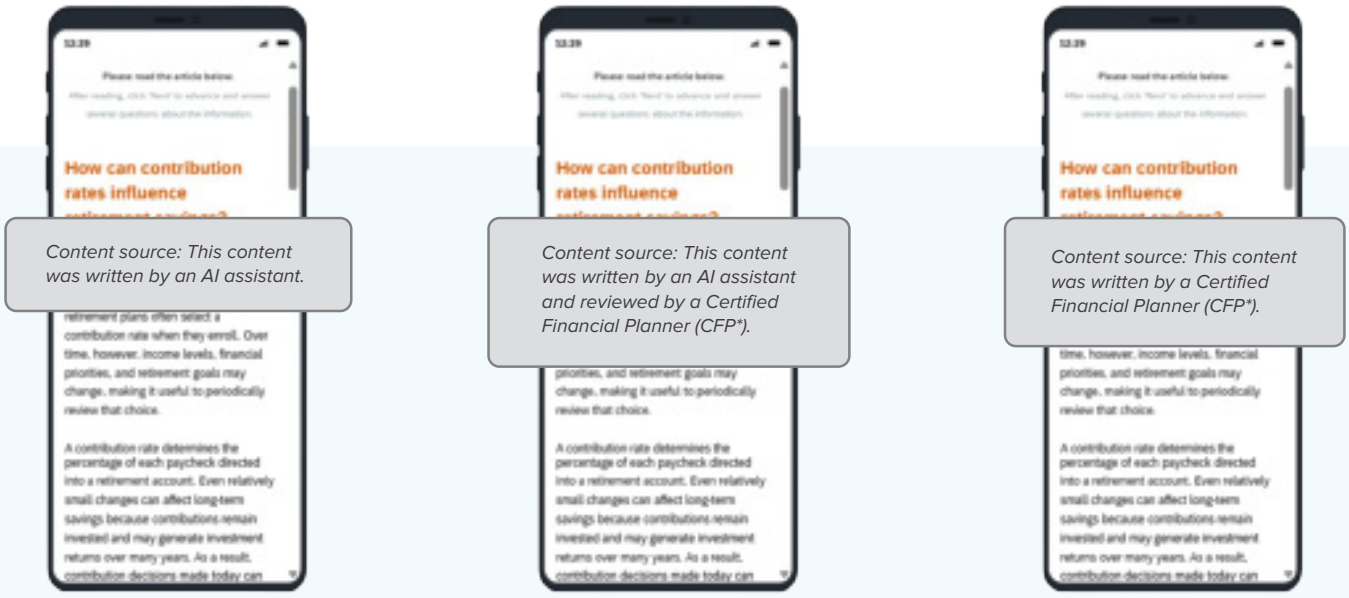
Less clear is how those disclosures shape participant responses. Does telling people that financial content was created by AI change how much they trust it? Does disclosure affect how accurate or useful they find the content? Does it alter how well they understand the information being communicated, or whether they are willing to act on the guidance they receive?

To test these questions, Voya's Behavioral Finance (BeFi) Institute for Innovation conducted a randomized experiment with approximately 600 participants who had employer-sponsored 401(k) plans or IRAs. Participants were randomly assigned to one of three source attribution conditions.

All participants reviewed the same financial education content, but it was presented as if it had been written by one of the following three sources:

- AI-generated by an AI assistant
- AI-generated by an AI assistant and reviewed by a financial professional
- Human-generated by a financial professional

Within each condition, participants viewed and evaluated three of six financial topics: Diversification, Roth versus Traditional IRA, increasing contributions, emergency savings, annuities, and accident insurance. The articles were matched in length, structure, reading level, and writing style. The research team examined how source attribution influenced content impressions, comprehension, and willingness to act on the recommendations.

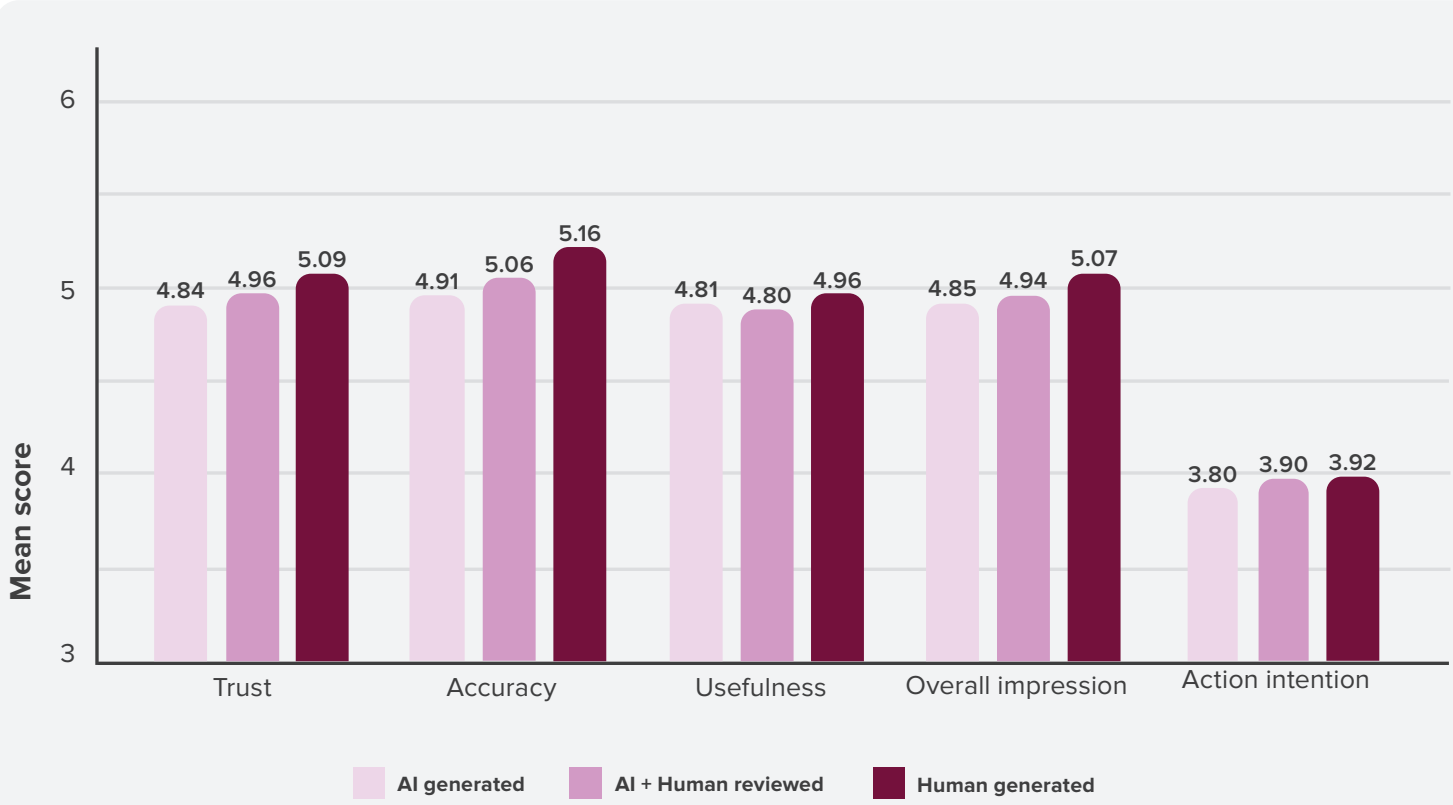


The findings revealed a key tension: trust and learning do not always move together. Human attribution increased trust, while AI attribution improved comprehension. In other words, source attribution shapes different parts of the participant experience in different ways.

Human attribution increases trust

Source attribution significantly shaped how participants evaluated the information. Human-attributed content generated the most favorable impressions, particularly on trust, compared with AI-attributed content. This is consistent with prior research that finds consumers often prefer human experts to AI systems, particularly for more consequential medical decisions. Content labeled as AI-generated and reviewed by a financial professional fell descriptively between the AI-only and human-only conditions, but its differences from either condition were not statistically significant. Practically, this suggests that adding human review may help, but it may not fully replicate the credibility benefits of content presented as human-generated.

Table: Content attributions impact evaluations of financial information, including perceptions of trust, p = .022, accuracy, p = .034, and overall impressions, p = .042[1]



Trust drives engagement

Trust did more than improve participants' impressions of the content. It served as a bridge between content experience and engagement, helping move participants toward action. Participants who trusted the information more reported being more willing to review their retirement investment allocations, increase contributions, evaluate annuities, and learn more about insurance products. Specifically, each one-point increase in trust was associated with a 0.41-point increase in average action intentions on the six-point intention scale. For plan sponsors and providers, this finding suggests that credibility cues may be especially important when the goal is to motivate next steps, not just communicate information.

Trust changes what happens next



Person 1 & 2 open retirement article to read

They view the content source



Person 1

“I’m not sure how much I trust this.”

Closes article



Person 2

“This seems credible, I should look into this further.”

They read the article

Log into their retirement account

Review contributions

Explore retirement income options

Save the information for later

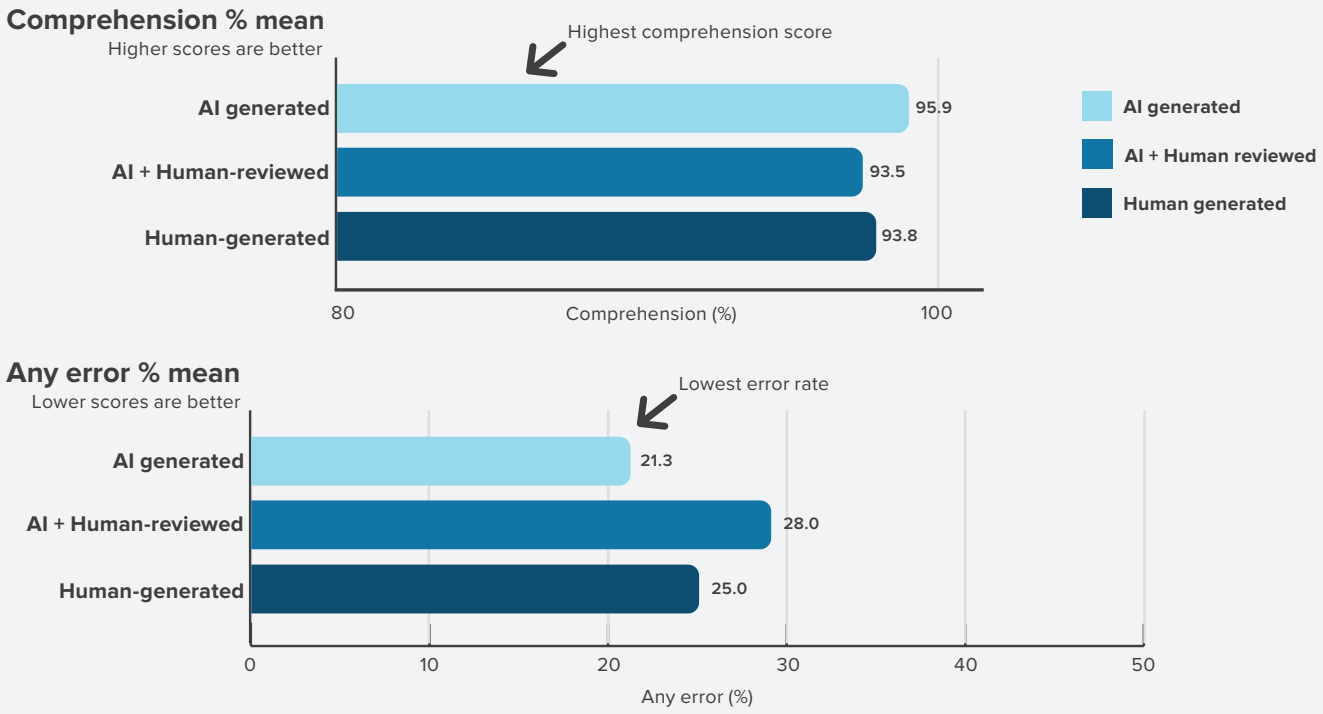
Key takeaway:
The findings suggest that source attribution influences engagement not because labels directly drive behavior, but because they shape trust, and trust increases people’s willingness to take the next step.

AI-attributed content shows a comprehension advantage

The surprising finding was that participants understood AI-attributed content the best despite trusting it the least. Participants exposed to AI-attributed content achieved the highest comprehension scores, made fewer errors, and expressed less uncertainty, challenging the conventional assumption that greater trust necessarily translates into better learning outcomes. While comprehension was high across all conditions, the relative differences were revealing.

One possible explanation is that AI attribution encouraged more deliberate or critical processing. Consistent with that possibility, reading time had a small positive relationship with comprehension suggesting that participants who spent longer with the content tended to understand it better. The evidence does not, however, conclusively show that reading time caused the comprehension advantage in the AI condition. Together, the findings suggest that while human attribution strengthens trust, AI attribution may prompt more careful engagement with information. In short, participants trusted the human more but learned more from the AI.

Table: AI-attributed content positively impacts comprehension, p = .010[1]



Key takeaway:

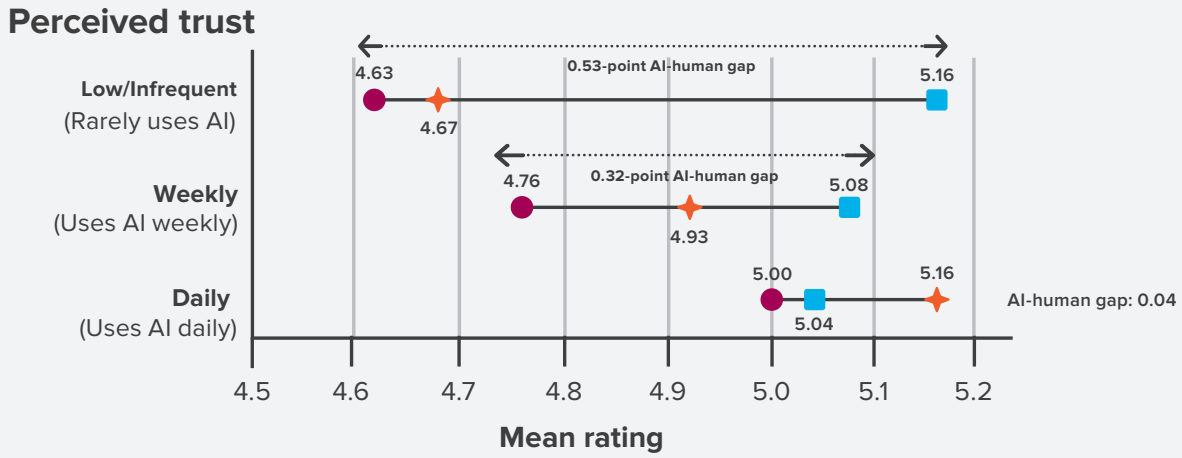
- The AI-generated condition had 2.4% higher comprehension mean than AI + Human reviewed, and 2.1% higher comprehension mean than Human-generated.
- The AI generated condition had 6.7% lower error % mean than the AI + Human-reviewed condition, and 3.7% lower error % mean than the Human-generated condition.

Familiarity with AI influences responses to AI-attributed content

AI familiarity influenced how participants responded to source attribution. Frequent AI users viewed AI-attributed content more favorably and were less sensitive to whether content was labeled as AI- or human-generated. Less frequent AI users reported lower trust and less favorable impressions of AI-attributed content, producing the largest gap between AI and human attribution. This makes AI familiarity potentially an important segmentation variable. As AI tools become more common in everyday work, financial, and consumer activities, participant reactions to AI-labeled guidance may shift over time.

The AI-human impression gap:

Table: AI usage shapes the effects of source attribution on perceived accuracy, p = .009 and perceived trust, p = .021



Perceived trust (Mean ratings); original AI usage scale has been truncated



Perceived accuracy (Mean ratings); original AI usage scale has been truncated

AI generated AI + Human reviewed Human generated

Key takeway: Experience matters.

Individuals who never use AI exhibit the largest difference in trust and perceived accuracy between human- and AI-labeled content, while frequent AI users show minimal differences in their evaluations.

The right source at the right moment: A framework for participant education and decision support

As AI becomes more integrated into participant experiences, the goal is not to maximize either AI or human involvement, but to leverage each where it adds the greatest value. The findings suggest that AI and human attribution support different purposes: human expertise strengthens trust, while AI attribution may enhance learning and comprehension.

This distinction has important implications for how financial institutions and plan sponsors can design benefits experiences. Rather than applying a single attribution strategy across all interactions, organizations may benefit from aligning AI and human credibility cues with participant needs, AI familiarity, and decision complexity.

The framework below illustrates one possible approach:

Table: A framework for participant education and decision support that leverages AI- and human-credibility cues.

Participant context	Primary goal	Content / Topic type	Recommended credibility cue	Example
Low-to-moderate complexity	Learning or Comprehension	• Emergency savings • Roth/Traditional IRAs • Contributions • Diversification	AI-led	• Content written by an AI Assistant • AI-powered retirement education
Moderate complexity	Planning	• Retirement readiness assessments • Savings projections • Goal setting and planning tools	AI + human-reviewed	• Content written by an AI Assistant and reviewed by a Certified Financial Planner (CFP®) • AI-powered recommendations validated by retirement specialists
Moderate-to-high complexity	Guidance or Decision Support	• Annuities • Supplemental health insurance (accident) • Retirement income products • Distribution decisions	Human-led	• Content written by a Certified Financial Planner (CFP®) • Content by retirement specialists

The opportunity is not to choose between AI and human expertise, but rather to design participant experiences that use each credibility cue where it is most likely to improve understanding, trust, and action. The future isn't AI versus humans, it is knowing when AI helps people learn and when human expertise helps them decide.

Three takeaways for plan sponsors and benefits leaders:

1. **Trust is a critical engagement lever.** Participants were more willing to take retirement-related actions when they trusted the information, and human attribution strengthened that trust. For benefits designers, this means credibility cues, such as financial professional endorsement, can be especially important when communications are intended to motivate action.

2. **AI may be especially useful for education and comprehension.** Even though AI-attributed content was trusted less, participants appeared to understand it better. Benefits designers may want to use AI-enabled tools for helping participants process information, while pairing AI with human expertise when decisions are consequential.

3. **Participant reactions to AI are not one-size-fits-all.** Familiarity with AI shaped how participants responded. Frequent AI users were more accepting, while less frequent users were more skeptical. Benefits designers should consider segmenting communication strategies by AI comfort level and decision complexity rather than applying the same AI disclosure or attribution approach across all participant interactions.

[1]Based on a Voya BeFi experiment conducted on August 26, 2026.

This report is for educational purposes only. Each plan must consider the appropriateness of the investments and plan services offered to its participants. Any financial services firm that adopts this strategy will need to ensure they are complying with regulatory requirements as well as firm policies around communications with the public and the use of AI.

Products and services offered through the Voya® family of companies.

6792046_0926 © 2026 Voya Services Company. All rights reserved. CN5946042_1028

VOYA® Plan. Invest. Protect.