Received: 27 October 2024 | Accepted: 31 October 2024

DOI: 10.1002/jcpy.1452

METHODS DIALOGUE

AI and the advent of the cyborg behavioral scientist

Geoff Tomaino | Alan D. J. Cooke | Jim Hoover

University of Florida Marketing Department, Gainesville, Florida, USA

Correspondence
Geoff Tomaino, University of Florida Marketing Department, 1454 Union Road, Gainesville, FL 32611, USA.
Email: geoffrey.tomaino@ufl.edu

Abstract

Large Language Models have been incorporated into an astounding breadth of professional domains. Given their capabilities, many intellectual laborers naturally question to what extent these AI models will be able to usurp their own jobs. As behavioral scientists, we performed an effort to examine the extent to which an AI can perform *our* roles. To achieve this, we utilized commercially available AIs (e.g., ChatGPT 4) to perform each step of the research process, culminating in an AI-written manuscript. We attempted to intervene as little as possible in the AI-led idea generation, empirical testing, analysis, and reporting. This allowed us to assess the limits of AIs in a behavioral research context and propose guidelines for behavioral researchers wanting to utilize AI. We found that the AIs were adept at some parts of the process and wholly inadequate at others. Our overall recommendation is that behavioral researchers use AIs judiciously and carefully monitor the outputs for quality and coherence. We additionally draw implications for editorial teams, doctoral student training, and the broader research ecosystem.

KEYWORDS
artificial intelligence, behavioral research, large language models

INTRODUCTION

The recent advent of Large Language Models (LLMs) since the launch of ChatGPT in November 2022 has entailed dramatic growth in the capability and public availability of AI technology, earning them mass adoption (Chui et al., 2023; Eastwood, 2024). They have proven useful in a wide breadth of domains, from the professional (e.g., interview prep and writing emails to colleagues) to the personal (e.g., financial advice and travel planning; Haan, 2023). Yet, a pervasive fear surrounding these AIs is their ability to usurp human labor (Caminiti, 2023; Haan, 2023). Given their extraordinary capability, it is natural for many intellectual laborers to question how and to what extent these AIs will be able to perform their own jobs (Amankwah-Amoah et al., 2024).

As behavioral scientists, we sought to examine the extent to which these AIs can perform *our* roles and, given any observed limitations, identify where the participation of trained human researchers is still vital. Specifically, we performed an exercise whereby we utilized commercially available AIs (OpenAI's ChatGPT 4, Bing's Copilot, and Google's Gemini) to perform each step of the research process.[1] That is, we took the role of the "cyborg behavioral scientist," offloading as much of the work of a research project as possible to AI, while still exercising final human judgment. This culminated in an AI-written manuscript, which is embedded in this paper.

More generally, throughout this paper, we provide a travelog of sorts, documenting our experience using these AIs, as they currently exist. We provide commentary regarding this experience and guidance for the behavioral sciences as we all collectively try to learn how to best incorporate these models into our workflows. We consider this an initial case study of the current capability of AIs to contribute broadly to behavioral research

[1] We utilized publicly available LLMs through their respective websites (e.g., chatgpt.com), rather than through APIs or other means of access.

Accepted by David Wooten, Editor; Associate Editor, Joel Huber

Geoff Tomaino, Alan D. J. Cooke and Jim Hoover contributed equally to this work.

Linked article: Blythe, P. A., Kulis, C., McGraw, A. P., Haenlein, M., Hewett, K., Yoo, K., Wood, S., Morwitz, V. G., & Huber, J. (2025). Comments on "AI and the advent of the cyborg behavioral scientist". *Journal of Consumer Psychology*, *35*, 316–328. https://doi.org/10.1002/jcpy.1453.

while documenting some of the specific tasks at which they succeed or fail, hopefully spurring more in-depth, systematic investigation of these tasks.

During this process, we attempted to intervene as little as possible in AI-led idea generation, empirical testing, analysis, and reporting. This allowed us to assess the limits of AIs in a behavioral research context and propose methods for where and how behavioral researchers can and should incorporate AI into their research process, given the current strengths and limitations of these models. While our work speaks primarily to behavioral researchers, we also draw implications for editorial teams, doctoral student training, and future directions for the research ecosystem.

AIs are, of course, not the first new technology to be adopted by researchers and incorporated into their work. For instance, researchers would laboriously conduct their statistical analyses using punch cards and other unwieldy computational machines before the onset of statistical processing software (Barnes, 1998). However, just as the introduction of this software shifted researchers' capabilities and, by extension, pursuits, we aim to interrogate the ways and the capability with which AIs will elicit similar transformation. Perhaps what makes AI most unique in comparison to other innovations is the breadth of applications it offers to a researcher's workload. For that reason, we believe it is especially urgent to understand how AIs will impact the varying tasks involved in behavioral research and to question the ideal form of this integration. More generally, we hope to provide researchers and the academic research community with guidance around how AI can and should be utilized in behavioral research, now and in the future.

Approach: “Gemini take the wheel!”

Rules of interaction

To provide the most complete test of the AIs' abilities, we adopted an approach in which we let the AI dictate the direction of the research to the limits of its capability at each stage. When possible, we tried to relegate our role to implementation, simply crafting prompts that gave the AI enough information and direction to make a valuable contribution. However, as will be noted throughout this manuscript, there were many instances where circumstances necessitated a deviation from this goal.

In general, we intervened for one of two reasons: correction and direction. By *correction*, we mean that we made edits or adjustments to the AI's output when that output did not defensibly meet the standards of an academic publication. For instance, in the research design stage, we found that elements of the AI-designed experiments had internal validity issues, and their procedures included frivolous components, some of which we felt would be necessary to amend. By *direction*, we mean that in some instances, we needed to evaluate multiple viable answers to a prompt. In these situations, we took an executive role in deciding which of the many responses to use in subsequent stages.

AIs utilized

We utilized three different AI models over the course of this research: OpenAI's ChatGPT 4; Bing Chat (now Copilot); and Google's Bard (now Gemini). We chose these AIs due to their prominence and accessibility at the time. That is, we intentionally eschewed alternatives from AIs still under significant development in the interest of consistency across steps, as well as the reliability of our findings in this work. That said, even the mainstream AIs we chose underwent changes over the course of this research, such as Bing Chat being rebranded as Copilot and Bard as Gemini at various points during this work. Where possible, we attempted to use the same underlying model (e.g., GPT-4) throughout our work, but the features and interface available varied some over time. Also, to note, despite Bing Chat being adapted from ChatGPT, we opted to still include it as a separate model due to its alternative programming to ChatGPT. These differences are evidenced in the different natures of the responses we received from Bing Chat versus ChatGPT. This decision was also consistent with most of the concurrent research comparing commercial AIs (e.g., Calonge et al., 2023; Zúñiga Salazar et al., 2023).

The AI interactions in this project occurred during the period of November 2023 through April 2024.

OUR JOURNEY

Because we wanted to involve AI at every possible stage of the research process, we initially divided the research process into six consecutive stages that we deemed necessary components of typical behavioral research and that spanned as much of the process as possible. These included: (1) Ideation, (2) Literature Review, (3) Research Design, (4) Documenting Results, (5) Extending the Research, and (6) Manuscript Preparation. These stages are broadly representative of mainstream experimental behavioral research processes (Johnston et al., 2020; Kite & Whitley, 2012).

Figure 1 depicts our idealized AI-enabled research process schematically, with gray and light green nodes representing the main and subsidiary research stages, respectively. Dark green nodes describe the inputs and goals of each (sub-)stage. Light blue nodes represent the main artifacts created at each stage. Finally, dark blue diamonds represent places where we envision researchers needing to interact with the AI.

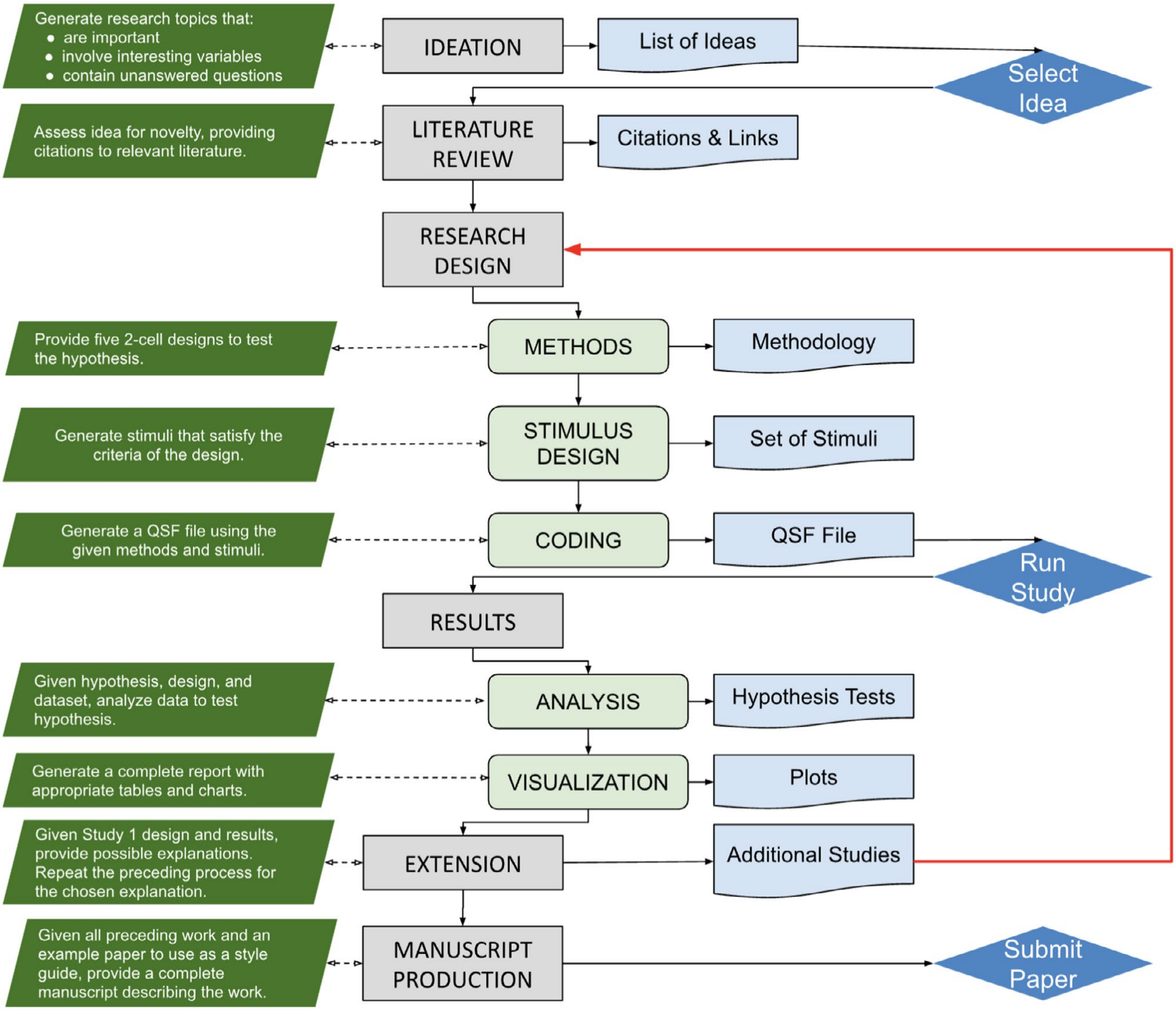
Generate research topics that:
• are important
• involve interesting variables
• contain unanswered questions
IDEATION
List of Ideas
Select Idea
Assess idea for novelty, providing citations to relevant literature.
LITERATURE REVIEW
Citations & Links
RESEARCH DESIGN
Provide five 2-cell designs to test the hypothesis.
METHODS
Methodology
Generate stimuli that satisfy the criteria of the design.
STIMULUS DESIGN
Set of Stimuli
Generate a QSF file using the given methods and stimuli.
CODING
QSF File
Run Study
RESULTS
Given hypothesis, design, and dataset, analyze data to test hypothesis.
ANALYSIS
Hypothesis Tests
Generate a complete report with appropriate tables and charts.
VISUALIZATION
Plots
Given Study 1 design and results, provide possible explanations. Repeat the preceding process for the chosen explanation.
EXTENSION
Additional Studies
Given all preceding work and an example paper to use as a style guide, provide a complete manuscript describing the work.
MANUSCRIPT PRODUCTION
Submit Paper

FIGURE 1 Flowchart of the AI-enabled research process.

Note that while these stages are typical of the research that is common in much of psychology and consumer research, they differ from those taken in many constructivist or postmodern research traditions (Fischer & Otnes, 2006). This does not imply that we believe that AI tools will not be useful in these instances. It merely reflects our lack of expertise within these research traditions. Furthermore, each of these six research stages is likely composed of multiple steps that may or may not be consecutive. Nevertheless, we enacted the six broadest stages as sequential: We attempted to perform, when possible, each stage using all three AIs. We would then judge the result that seemed most useful to us, and use that result, regardless of its AI source, as the input to the next stage for all AIs. More details of our journey, including sample prompts and AI responses, are documented in the Appendix S1. Furthermore, we have a more comprehensive documentation of our process stored here: https://github.com/g-tomaino/Cyborg-Behavioral-Researcher.git.

Of course, one of the most critical challenges of interacting with AIs as potential research partners is determining how to prompt the AI to provide a reasonable answer while minimizing the number of prompts necessary to achieve that answer. Assessing the correctness or reasonableness of an answer from an AI is key to the decision of when prompt engineering has been sufficient (Errica et al., 2024). This decision is driven by a domain assessment by the individuals conducting the tasking of the AIs. In our approach, if the AI result seemed to produce a reasonable response to the prompt for the research task, we accepted the result from the AI and did not attempt further prompting. If the result seemed incorrect or grossly incomplete, then additional prompting was attempted.

Ideation

In this first stage of the research process, the general idea for the research topic is developed. For human researchers, this stage often involves a long-standing interest in the topic and typically begins with a desire to extend existing research into new domains or phenomena, to explain currently unexplainable phenomena, or to systemize knowledge in a new, applied area (Cao et al., 2019). Despite this stage being perhaps the most personal to the researcher and therefore most intuitively performed by a human, in the interest of testing the complete sequence of a research project with AI, we feel it is important to consider ideation as part of our AI-driven process.

The methods for developing research theories and hypotheses are often broadly classified into one of three epistemological paradigms: a deductive approach uses theory to generate specific testable hypotheses, whereas an inductive approach attempts to infer general principles from a set of observations, and an abductive approach iterates the collection of data and refinement of theory to converge on empirical truth. LLMs operate by predicting patterns of words in response to input text string prompts using multilayered neural networks involving vast numbers of nodes, a process isomorphic to complex nonlinear regression. Since this is a chain of weighted sums, the process can be viewed as a form of deductive reasoning, albeit on a massive scale (Wolfram, 2023). Induction, on the other hand, occurs implicitly through the integration of a large number of training observations used to make the predictions. Therefore, AIs should likewise be capable of induction, but the quality of those inductions will depend greatly on the nature of the training data.

Our approach to research ideation using AIs, however, is perhaps best characterized as a specific form of abduction. Our experience involved describing the scope of ideas in which we were interested and the criteria we used to evaluate "good" ideas, and then having the AI list research questions that it deemed worthy of investigation. This perhaps most closely mirrors an inductive approach; however, we still characterize it as abduction since instead of pursuing a research direction based on data, we chose a direction based on a series of suggestions that were in turn based on a subset of published results existing in the training data. That is not to say that the AIs provided immediately usable ideas. Instead, we found ourselves using a drill-down approach, through which we prompted the AI models for more specificity on their responses that we judged to be most promising. While such an involved approach to prompting may not be necessary as these models develop, in their current forms we found this to be the most productive method of eliciting usable research ideas from the AI models.

It is a best practice in research ideation for teams to generate large numbers of ideas as the starting point for selecting the best ideas for research (Demszky et al., 2023; Stremersch, 2024). Similarly, we followed this practice by asking each AI to generate multiple responses from which to choose the response that is most interesting to research (Shaer et al., 2024). We began by asking the AIs for important topics in the field of consumer behavior where a meaningful contribution could be made. In response, the AIs gave us broad research spaces (e.g., ChatGPT 4's suggestion of looking into "Digital Consumption and Mental Health"). These ideas were not sufficiently refined for a researcher to pursue in a meaningful empirical project. They do, however, pose interesting territories for much broader research, such as review papers, or large exploratory studies.

The reason for the generality of these suggestions may be due to the data the AIs were trained on. AIs do not have access to the content of closed-access academic journals, so AIs cannot "read the literature." Instead, current AIs purportedly access only open-source websites and summaries, although there are nascent deals that may broaden their access (Clark, 2024). Thus, current AIs likely have a general sense of what the field is interested in, without understanding the field well enough to identify a meaningful gap and then independently suggest a research idea specific enough to pursue as a major empirical contribution.

We next decided among ourselves which from the large set of general research ideas we would be most interested in pursuing. We opted for ChatGPT 4's suggestion of "The Ethical Consumer: Factors Influencing Ethical Buying Decisions," both due to our personal tastes as researchers, as well as our judgment that this broad area likely had the most room for meaningful contributions. This is also a research area that none of us had experience in, allowing us to better adopt a naive outsider's perspective appropriate for this project.

We then asked the AIs for more specific research ideas within this topic. The resulting suggestions were much better defined, often containing an empirical prediction (e.g., "Ethical Consumption as a Luxury Good: Ethical consumption patterns will mimic those of luxury goods, with conspicuous consumption playing a role in the purchase of visibly ethical products.") and therefore more plausible research projects. We used our judgment regarding which idea from these lists of ideas was most likely to make a novel contribution, while also having sufficient face validity.

We chose to modify the following suggestion to serve as the target idea around which our program would be oriented:

> Ethical Fatigue: There will be a segment of consumers who experience "ethical fatigue" and become skeptical or indifferent to ethical branding due to overexposure to marketing messages about ethics.

Overall then, our experience of using AIs for research ideation showed that the initial suggestions from the AI were quite general, but with a researcher's guidance and refinement, useful ideas could be obtained (further notes on our experience at this stage can be found in the Appendix S1, pp. 1–2).

Literature review

Having chosen a research idea in collaboration with AI, we then undertook a literature review. In particular, at this stage we wanted to understand how to position our idea vis-à-vis existing work *and* whether it was indeed a novel idea. We found that the AI tools we used were lacking in this capacity.

When we presented the AIs with the ethical fatigue idea and asked them whether this idea was novel and worth pursuing, as well as for related literature, the AIs had very little to offer. They tended to make a great show of complimenting the idea and discussing practical contributions it might make, but they had a very shallow understanding of how the work might be positioned theoretically. Follow-up queries across the different AIs indicated that either the AI produced references that were not real papers (i.e., hallucinations) or were from open-access journals.

This limitation most likely comes from a more fundamental limitation whereby these models are unable to scan closed-access journals (Clark, 2024). Many of the top publications in the behavioral sciences are in hybrid open-access and closed-access journals. Not being able to refer to these publications significantly reduces the knowledge base of these AIs. In lieu of major academic publications among which we could position our work, we instead mostly received links to websites discussing tangentially related work; that is, when the links worked.

Overall, we were disappointed to find the AIs to be of minimal assistance at this stage in the research process (see Appendix S1, pp. 2–3). We instead conducted an independent review of prior work ourselves to ascertain the novelty and contribution of the research question.

Research design

We then asked the AIs to generate a sufficient test of the prediction. That is, we asked them to design an experiment that would test whether consumers experience "ethical fatigue," whereby higher exposure to ethical brand statements reduces the efficacy of subsequent ethical brand statements.

We found the AI outputs in this step to be generally useful. They provided designs that were fairly adequate in both internal and external validity and that clearly tested the key prediction. They were, however, not implementable without some alteration. For instance, Bing Chat's suggested design involved showing participants one, two, or four ethical statements from alternative brands before evaluating a focal brand's ethical statement. This introduces an obvious confound whereby not only does the number of previous ethical statements vary across conditions, but so does the number of previous statements of any kind. Thus, we made an adjustment to this design in which all participants were shown four statements about other brands and where we manipulated the number of these that were ethically related.

This resulted in the following experimental design: all participants were asked for their perception of Nike's ethicality using a four-item scale, and then saw four statements regarding four other brands with zero, two, or four of these statements being ethical, the rest being non-ethically-related brand positioning statements. Then all participants read the same ethical positioning statement: "Nike is committed to using sustainable materials and practices. We are also committed to promoting social responsibility and empowering athletes of all levels." We then concluded by reassessing participants' perceptions of Nike's ethics using the same four-item scale.

Another limitation of the AIs at this stage was that they required the researcher to already have a fair foundation in the experimental method. For example, not all of the AIs explicitly noted that participants should be randomly assigned across conditions, meaning the researcher would have to know to incorporate that design component themselves. They also varied in the detail with which they specified the nature of the manipulation or the dependent variables used to measure their effects. For instance, ChatGPT 4 suggested a pre–post measure of attitudes toward the target brand, while other AIs did not.

More generally, we found it necessary for us, as researchers, to intervene and ensure that the suggested experiments maintained internal validity. That is, the AI models seemed to have a strong grasp of what an experiment should resemble but were unable to sufficiently critically evaluate them for issues like confounds.

As a step within the research design stage, we prompted AIs with the chosen design, and asked them to generate experimental stimuli in the form of ethical and non-ethical brand positioning statements. We found that the AIs performed quite well at this task, providing us with realistic statements that clearly manipulated the construct of interest, consistent with prior work using AI for stimulus generation (Sarstedt et al., 2024). For instance, Bard provided the following ethical statement: "Seventh Generation is committed to using non-toxic ingredients and environmentally friendly practices. We believe in making a healthier home for everyone."

Given that we now had a complete research design, including procedure, stimuli, and measures, we wondered (in a flash of unbridled optimism) whether it would be possible to have the AIs actually create the experiments for us using Qualtrics' QSF (Qualtrics Survey Format) questionnaire schema, with Qualtrics being the survey platform supported by our institution. The answer was an unequivocal "no." Only ChatGPT 4 was able to provide us with a QSF file in response to our prompt, and when we examined its contents, we found that it was missing many tags that were crucial to the QSF schema. As a result, Qualtrics was unable to open the QSF file. Repeated attempts to improve the output, including providing information about the QSF schema, proved ineffective.[2] Instead, we added the resulting stimuli to a Qualtrics survey manually.

We then conducted the AI-designed study using an online sample recruited through Prolific Academic. We chose not to use AIs to generate human-subject-like data for reasons that we discuss in the General Discussion. As described in the AI-written manuscript, this study ultimately provided evidence in support of the AI's predictions.

As an additional note, we intentionally limited the final stimuli to text, rather than incorporating images or other forms of media that may have been appropriate for these experimental designs (e.g., showing a logo for each brand mentioned). While some AI models can generate visual stimuli, our initial attempts at prompting complementary images when designing these experiments resulted in incoherent and unusable content. While these engines may improve in their ability to generate useful visual stimuli, as well as stimuli of other types (e.g., audio), we did not find them capable of reliably doing so in their current forms.

In sum, we found the AIs to be useful at the design stage (see Appendix S1, pp. 3–6 for more detail). However, their value varied, with them showing excellent promise for stimulus creation, moderate promise as a partner in experimental design, and ineptitude at producing a Qualtrics QSF file.

Results and analysis capabilities

We were disappointed to find the AIs to be highly limited in their ability to analyze and report data (for more details, see Appendix S1, pp. 6–7). First, Google's Bard did not provide an option for us to upload a dataset for analysis, precluding it as a tool for this stage. We therefore only used Bing Chat and ChatGPT 4 in data analysis attempts.

We downloaded the comma-delimited file from Qualtrics, removed test participants, and then uploaded this cleaned file with no further alteration to the AIs. We also prompted the AIs with a description of the corresponding experiment. We did find that both AIs were able to decide on the correct analytical tools. That is, they both understood that an ANOVA should be used. Furthermore, they produced results that looked appropriate at first glance. However, troublingly, these analyses contained incorrect statistics. More generally, we were not able to find a mode of requesting analyses that a researcher could rely upon without verifying. Since the act of verifying involves conducting these analyses oneself, we advise that researchers do not use these models for data analysis in their current development.

A related step in results analysis is data visualization. Here we found that the AIs all provided appropriate and well-formatted charts when given clear prompts. However, because these charts may be based on faulty underlying analysis, we caution researchers to carefully verify the displayed results.

Extensions

In this stage, we asked the AIs to develop extensions of the original study (see Appendix S1, pp. 8–10 for more details). That is, we asked them what we should do next, after our initial findings, to enhance the contribution. The first step in this procedure was asking the AI models why they thought we got the results we did (i.e., we asked them for a mechanism). All of the models suggested multiple possible mechanisms. Most of these suggested mechanisms were tenuous in their ability to account for our effect. For instance, ChatGPT 4 suggested that "[i]nformation overload from multiple ethical statements could lead to cognitive dissonance, where participants struggle to reconcile the various claims they've encountered. This dissonance might make it harder for Nike's statement to stand out or be internalized effectively, leading to a smaller boost in ethical perception." Other suggestions were essentially redescriptions of the initial result, rather than a process mechanism.

All three AIs suggested some form of a saturation effect. For instance, ChatGPT 4's suggestion was as follows: "Exposure to multiple ethical statements from other brands might have saturated participants' perception of ethicality as a distinguishing feature. By the time they saw Nike's statement, the novelty or impact of ethical claims might have diminished, making Nike's statement less impactful or persuasive." As such, we asked all of the models to develop a moderation study to test for this mechanism. The models all required an explanation of how a moderation study should look; many seemed to have a difficult time understanding what moderation meant and what an appropriate design to test for moderation would be. After providing this explanation, the models all provided adequate experimental designs. We ultimately went with ChatGPT 4's suggestion to introduce a second factor of gain or loss framing of the positioning messages. ChatGPT 4 did not explain how this additional factor related to the mechanism we were pursuing, nor could we intuit an obvious link. Despite this, we wanted to honor the exercise of following the AI's guidance and ran the study anyway. While we did

[2] We were a bit surprised at this result since AIs have shown themselves to be adept at a range of programming tasks (Nam et al., 2024). We expect that our outcome may be due in part to the proprietary nature of the QSF schema. We were unable to find complete documentation of it online, which stands in stark contrast to the vast amount of information online regarding popular open-source programming languages like Python and R. It is possible that if Qualtrics provides more information about their schema, future AIs may be able to master this task, representing a major time-saving tool for researchers, especially for complex designs, or that Qualtrics may provide this service on their platform.

make edits to the AI-generated design for the first experiment, that was in the interest of internal validity. By contrast, for this additional study, we would have only been making edits in support of face validity, which we felt would not be in keeping with the spirit of this exercise. Likewise, we wanted to be open to the possibility that the AI may select a meaningful moderator even if it is unable to articulate its theoretical role.

We then repeated the process of experimental design (with our findings and experiences similar to those of Study 1) and ran the study ChatGPT 4 had suggested with a sample of undergraduate student participants, which yielded no significant results.

Manuscript production

After completing the empirical portion of the project, we turned to manuscript production (see Appendix S1, p. 10 for more details). By this we mean writing up the empirics, as well as providing a motivating front end and thoughtful discussion. We began by creating a series of prompts describing the idea and our results, which we gave to the AIs for background. This was followed by prompting for actual manuscript writing. While it was tempting to upload a representative paper as a model around which the AI could design the manuscript, this poses significant questions of copyright that we chose to eschew. Instead, we relied on the model's existing knowledge, telling it the manuscript is targeting a "top-tier consumer psychology journal."

The first friction in this process was the AIs' abilities to write all of the portions of a complete manuscript at once. That is, the AIs would frequently turn to bullet points or offer suggestions on how to write certain sections, rather than actually writing those sections. By breaking our requests for these sections up into separate prompts, however, we were ultimately able to prompt the AI for a complete manuscript (see EMBEDDED MANUSCRIPT). An exception to this, however, was the methods and results sections. We had to write much of these ourselves, given the high error rate for the AI models, as discussed in the preceding results section, and as likewise exhibited by a hallucinated citation in the reference section.

Overall, we found the writeup to be usable, but in need of significant editing (which we did not do, in the interest of transparently demonstrating the capability of AI at this stage). The various sections were written in a way that was on topic and informative, yet lacking much of the information and style a behavioral research reader would expect. Though a professional researcher would be able to use this output only as a starting point, we do believe that these models could prove useful for non-doctoral students (e.g., honors thesis students) whose outputs may require less rigor.

As an additional interesting note, we thought that it might be useful to have the AI models review and edit the final manuscript as a form of refinement. We found, however, that these same limitations we saw in manuscript production also applied to editing, making this effort largely futile.

EMBEDDED MANUSCRIPT

In this section, we present the manuscript output of the aforementioned AI interactions:

ETHICAL FATIGUE

Introduction

In recent years, the marketing landscape has seen a paradigm shift towards ethical branding, with companies increasingly promoting their commitment to social and environmental responsibilities. This trend aligns with a growing consumer demand for brands that not only offer quality products but also reflect personal values related to sustainability and ethical conduct. However, as ethical branding becomes more prevalent, a critical question arises: Does the constant bombardment of ethical marketing messages lead to "Ethical Fatigue" among consumers? This manuscript proposes the exploration of Ethical Fatigue, defined as the skepticism or indifference towards ethical branding caused by overexposure to marketing messages about ethics. This research aims to bridge a gap in the literature by focusing on the potential adverse effects of excessive ethical marketing on consumer perception and behavior.

Literature review

The concept of ethical consumption is well-documented, with studies emphasizing the importance of ethical considerations in consumer purchase decisions (Bray et al., 2011). Consumers increasingly seek out brands that demonstrate social and environmental responsibility, indicating a shift towards more conscientious consumption patterns (Carrington et al., 2010). However, research on the implications of overexposure to ethical branding is sparse, signaling an opportunity to explore the phenomenon of Ethical Fatigue.

Consumer skepticism towards brand motives in ethical marketing has been noted, with studies indicating that consumers often question the authenticity of brands' ethical stances, suggesting a potential precursor to Ethical Fatigue (Parguel et al., 2011). Additionally, the concept of advertising wear-out suggests that repeated exposure to a specific message can lead to diminished consumer attention and engagement, which could extend to ethical marketing messages (Ehrenberg et al., 2002).

Furthermore, the psychological impact of information overload in the digital age (Eppler & Mengis, 2004) may contribute to Ethical Fatigue, as consumers navigate an ever-increasing barrage of marketing communications, including ethical branding messages. This parallels findings in cognitive psychology that suggest excessive information can lead to decision fatigue, potentially influencing consumer reactions to ethical branding (Vohs et al., 2008).

Research gap

Despite the substantial body of literature on ethical consumerism and skepticism towards marketing communications, the specific concept of Ethical Fatigue—stemming from overexposure to ethical branding—remains underexplored. This manuscript proposes to fill this gap by investigating how repeated exposure to ethical marketing messages influences consumer perceptions, potentially leading to skepticism or indifference. This research will contribute to the broader discourse on ethical consumer behavior, marketing ethics, and consumer psychology by introducing and elucidating the concept of Ethical Fatigue.

Study 1: ethical advertising frequency's impact on brand perception

In our investigation of ethical advertising frequency's impact on consumer perception of a brand's ethical stance, we manipulated exposure to ethical versus non-ethical brand statements before evaluating the target brand. Participants were assigned to one of three conditions: a control group with no ethical statements, a low-frequency group with two ethical and two non-ethical statements, and a high-frequency group with four ethical statements. Our hypothesis posited that while exposure to the target brand's ethical positioning would generally enhance perceptions of its ethicality, this effect would attenuate as the frequency of ethical statements encountered increased.

We chose Nike as a target brand since it would be generally familiar to participants. We assessed perceptions of Nike's ethical position using an index of four measures:

- Nike is ethical.
- Nike cares about the environment.
- Nike is socially responsible.
- I would trust Nike to do what is right.

These measures were each assessed using a five-point Likert scale anchored at Strongly Disagree and Strongly Agree. This assessment occurred twice, once before and once after the brand positioning statements were presented. The other brand positioning statements were taken from one of two lists. The ethical statements were:

- TOMS is committed to giving away a pair of shoes for every pair purchased. We believe in making a positive impact on the lives of others.
- Warby Parker provides affordable eyeglasses to people in need. We believe in making quality eyewear accessible to everyone.
- Method is committed to using plant-based ingredients and sustainable packaging. We believe in making a positive impact on the environment.
- Seventh Generation is committed to using non-toxic ingredients and environmentally friendly practices. We believe in making a healthier home for everyone.

The non-ethical statements were:

- Tesla is the future of transportation. With its sleek design, powerful acceleration, and long range, Tesla cars are the perfect way to experience the thrill of driving.

- The Apple Watch is the perfect companion for your iPhone. With its stylish design, comprehensive health tracking features, and convenient notifications, the Apple Watch makes it easy to stay connected and informed, all while keeping you on top of your fitness goals.
- The Sony PlayStation 5 is the ultimate gaming experience. With its incredible graphics, powerful processor, and immersive gameplay, the PlayStation 5 takes gaming to the next level.
- LEGO is more than just a toy. It is a creative tool that allows children and adults alike to express their imaginations and build whatever they can dream of. With its endless possibilities and enduring appeal, LEGO is a timeless classic that continues to inspire generations.

In the low-frequency condition, two statements were randomly selected from each list. All other brand statements were presented one at a time in random order.[3] After they had been presented, we presented the following target brand positioning statement: Nike is committed to using sustainable materials and practices. We are also committed to promoting social responsibility and empowering athletes of all levels. We then reassessed participants' perception of Nike's ethical stance and additional process and demographic questions.

Analyzing responses from 399 Prolific participants, we conducted a repeated-measures ANOVA to compare scores on perceptions of the brand's ethical stance before and after exposure to the manipulation. We observed high agreement among our index measures (Cronbach's alpha = .933 and .95 in the pre- and post-manipulation measures, respectively.) We observed a marginally significant interaction between the time (before vs. after manipulation) and condition, $F(2, 396)=2.903$, $p=0.056$, $\eta^2=0.01$. Post- hoc analyses revealed a significant increase in ethicality perceptions from before to after manipulation in all conditions (M_diffs=0.429, 0.405, 0.280 in the control, low- and high-frequency conditions, respectively, all ps <0.001, see Figure E1). However, the magnitude of the increase was significantly greater in the control condition than in the high-frequency condition ($F(263)=2.17$, $p=0.031$) and marginally greater in the low-frequency than in the high-frequency condition ($F(263)=1.83$, $p=0.069$), supporting our hypothesis.

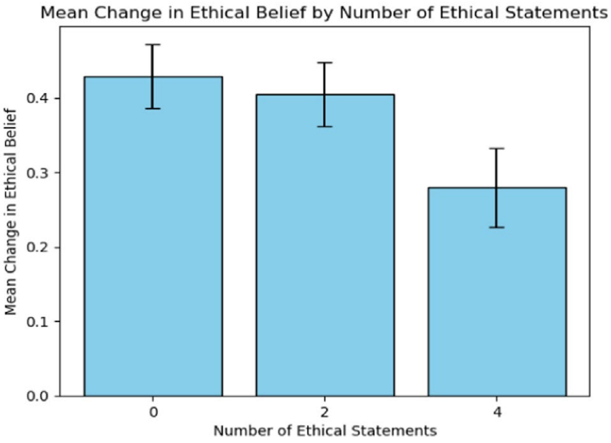

FIGURE E1 Mean change in ethical perception as a function of frequency of other brand ethical statements. Error bars represent 1 standard error of the mean.

Study 2: Frequency and framing's moderating effects on ethical positioning response

Study 2 extended our examination to the moderating effects of message framing (promotion vs. prevention) on responses to ethical positioning, considering both the framing and frequency of ethical statements. Participants were randomly assigned to one of four groups in a 2 (ethical statement frequency: control vs. high) ×2 (framing: promotion vs. prevention) factorial design. Our analysis focused on the interaction effects on brand perception following exposure. The other-brand positioning statements were framed as being either promotion-focused or prevention-focused and concerning ethical position or not. The content of the messages was held constant across framing manipulations. Examples include:

- Non-ethical promotion: Elevate your driving experience with Tesla, the pinnacle of

[3] Note from the human researchers: The preceding two sentences are unclear. In the control condition, we used only the non-ethical statements. In the low-frequency condition we used two statements from each list, selected randomly for each participant. In the high-frequency condition, we used only the ethical statements. Regardless, the four statements were presented one at a time in random order. We have not modified the original output to further highlight the limitations of the AIs at manuscript production.

automotive innovation. With its cutting-edge technology, Tesla offers unrivaled acceleration, superior range, and a design that turns heads. Embrace the thrill of the future today and join the revolution in transportation, making every journey an exhilarating adventure.

- Non-ethical prevention: Guard against outdated driving experiences with Tesla. Don't let traditional vehicles hold you back with their limited capabilities. Tesla's advanced features, including powerful acceleration and extended range, ensure you're always ahead, preventing the discomfort of frequent stops and the frustration of slow, unresponsive drives.
- Ethical promotion: Step into a brighter future with TOMS! For every pair of shoes you purchase, another pair is gifted to someone in need, spreading joy and making positive strides towards global well-being. Embrace the power of your purchase to transform lives and walk in solidarity with communities worldwide.
- Ethical prevention: Avoid contributing to global disparity with TOMS. Every pair of shoes you purchase prevents a child from going barefoot, tackling the issue of poverty and disease spread through unprotected feet. Join us in the fight against inequality and ensure no individual is left vulnerable due to lack of footwear.

The full set of other-brand positioning statements can be found in the Appendix S1. The target brand (Nike) ethical positioning statement, procedure and dependent variable index items were unchanged from Study 1.

434 undergraduate students at a large US university participated in return for extra credit in introductory classes. Again, the agreement between ethical perception measures was high (.933 and .949 for the pre- and post-manipulation measures). A 2 ×2 ANOVA revealed no interaction between the frequency of ethical statements and the framing on perceived ethicality of the brand, $F(1, 430) = 1.063$, $p = 0.303$, $\eta^2 = 0.0025$.

Specifically, the presentation of Nike's ethical positioning statement increased perceptions of Nike's ethics overall (M_diff=0.298, $t(434) = 10.32$, $p < 0.001$) but did not differ as a function of frequency ($t(434) = 1.525$, $p = 0.128$), framing ($t(434) = 0.970$, $p = 0.332$), or their interaction. The change in ethical perception was greatest in the control/prevention condition ($M = 0.392$) but this did not differ significantly from the other three conditions as seen in Figure E2.

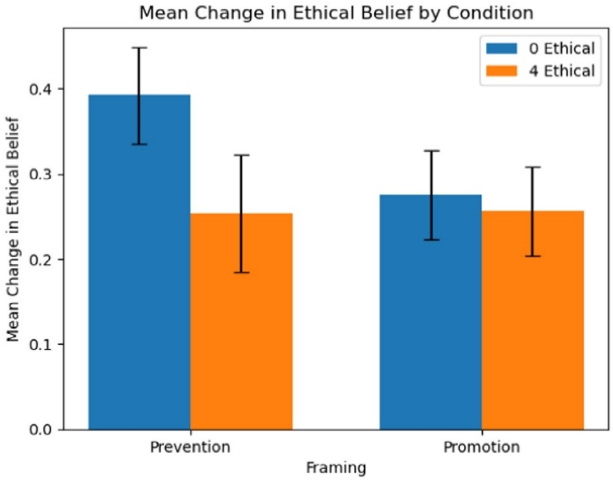

FIGURE E2 Mean change in ethical perception as a function of frequency of other brand ethical statements and message framing. Error bars represent 1 standard error of the mean.

Discussion

Our findings from Study 1 and Study 2 contribute to the understanding of ethical fatigue in consumer psychology. Study 1 demonstrates that while ethical advertising can enhance brand perception, overexposure to such messaging might lead to ethical fatigue, diluting the impact of these messages. Study 2 further elucidates the nuanced role of message framing, suggesting that the effectiveness of ethical positioning depends on both the frequency of ethical messages and their framing, with promotion framing being more effective at lower frequencies of ethical messaging.[4]

These results underscore the complexity of consumer responses to ethical branding efforts and highlight the need for brands to strategically manage the frequency and framing of their ethical

[4] Note from the human researchers: This interpretation of the results is notably incorrect. That is, there was no significant contrast such that differences in ethical perceptions varied in the promotion conditions as a function of the number of ethical messages.

communications to avoid ethical fatigue. Future research should explore additional factors that may influence the onset of ethical fatigue, such as the authenticity of the ethical stance and the consumer's pre-existing attitudes towards the brand.

General discussion

Our investigation into the phenomenon of Ethical Fatigue among consumers, through two empirical studies, has illuminated several notable findings, implications for marketing practice, inherent limitations, and promising directions for future research.

Summary of findings

The concept of Ethical Fatigue, characterized by consumer skepticism or indifference towards brands' ethical branding efforts due to overexposure, finds empirical support in our research. Study 1 revealed that while ethical advertising initially enhances brand perception, its effectiveness diminishes with increased frequency of ethical messages. This attenuation effect suggests a threshold beyond which additional ethical messaging may fail to further improve or might even harm brand perception. Study 2 expanded our understanding by exploring the interplay of message frequency and framing, finding that while ethical messaging generally boosts brand perception, this effect is not significantly influenced by the frequency of messages or their framing as promotional or preventive.

Implications

These findings hold substantial implications for marketers and brand managers. In an era where consumers are increasingly attuned to the ethical standing of brands, our research suggests that there is a fine balance between effective ethical communication and the risk of engendering Ethical Fatigue. Marketers are thus advised to strategically manage the frequency of their ethical messaging, ensuring that their communications are perceived as genuine and not merely as attempts to win consumer favor. Moreover, the nuanced role of message framing in ethical branding efforts suggests that brands should consider adopting a more diversified communication strategy that varies in content and style to maintain consumer engagement without triggering skepticism or indifference.

Limitations

Our study is not without limitations. First, the generalizability of our findings may be influenced by the selection of a single brand (Nike) as the focus of our studies, raising questions about how these insights apply across different industries and brand archetypes. Additionally, our participant pool, predominantly university students, may not fully represent the broader consumer population, particularly in terms of age, socio-economic status, and cultural background. Finally, our exploration of message framing focused on a binary categorization (promotion vs. prevention), potentially overlooking the effects of other framing strategies or message nuances.

Future directions

Future research should endeavor to address these limitations and expand the scope of investigation into Ethical Fatigue. Studies could explore a wider range of brands and sectors to assess the generalizability of our findings. Additionally, researchers should consider a more diverse demographic profile of participants to enhance the external validity of the results. Investigating the impact of different types of ethical messaging, beyond the binary framing of promotion and prevention, could offer deeper insights into how consumers interpret and respond to ethical branding efforts. Finally, the role of brand authenticity and consumer skepticism deserves further exploration, particularly in how these factors interact with the frequency and framing of ethical messages to influence brand perception and consumer behavior.

TABLE 1 Performance of each AI on each research stage (as of April, 2024).

Research stage	Substage	Bard/Gemini	BingChat/CoPilot	Chat GPT4
Ideation		L3	L2	L3
Literature Review		L2	L2	L2
Research Design	Methods	L3	L3	L3
	Stimulus design	L3	L2	L3
Results	Coding	L0	L0	L1
	Analysis	L0	L0	L1
	Visualization	L0	L0	L1
Extension		L1	L2	L2
Manuscript Prod.		L0	L0	L1

Note: L3, AI produced output at an acceptable level for academic research with little modification; L2, AI-produced output deemed valuable with difficulty or substantial modification; L1, AI-produced output requiring such oversight as to be without value; L0, AI unable to produce requested output.

In conclusion, our research sheds light on the intricate dynamics of ethical branding and consumer perception, offering valuable insights for both scholars and practitioners interested in navigating the complexities of ethical marketing. By carefully balancing the frequency and framing of ethical messages, brands can foster positive consumer perceptions while avoiding the pitfalls of Ethical Fatigue, ultimately contributing to a more sustainable and ethical marketplace.[5]

GENERAL DISCUSSION

We utilized various AIs to perform as much work as possible in the development of a behavioral research project. We did this across various steps, noting the quality of the assistance offered by the AI, ranging from useful to unusable (see Table 1 for a comprehensive summary of our evaluations). In general, we found that these AIs can offer some assistance, but their value stops there, at assistance. They proved to be unreliable tools for multiple aspects of a behavioral research project. They did not exhibit sufficient quality in their taste at the ideation stage to be utilized without some degree of oversight from a trained researcher. They also made experimental design mistakes that would prove fatal to a paper if left uncorrected. They furthermore exhibited significant limitations in their sourcing and writing of literature reviews. We also found that the analyses conducted by these AIs, for those AIs that even made an attempt possible, were not to be trusted. Thus, in their current manifestation, we think it is most appropriate to classify these AI tools as a potential service to a researcher, not as an autonomous partner.

Our observations are consistent with work in other domains wherein a human and AI working in tandem have proven to be an effective combination. For instance, researchers have found that an expert working with AI can improve the detection of breast cancer while reducing the number of false positives (Dembrower et al., 2023). Similarly, research has found that call center performance goes up when AI is incorporated, but also that AI cannot take the place of humans in terms of empathy and handling complex situations that the AI has not been trained on (Ferraro et al., 2024). More generally, at certain stages of research, we found that AI can save the researcher time and augment their creativity. However, we also found that major researcher input is still necessary to make for a successful collaboration.

We summarize these reactions in Table 1. We hope this table is useful as a tool to researchers that currently want to incorporate AI into their workflow, as a "point in time" reference for the field to understand the efficacy of AIs in their current form in a research capacity, and as a spotlight for all concerned parties on where these models most need improvement.

Advice on incorporating AIs into behavioral research

Based on our experience, we developed a revised version of Figure 1, in which we laid out the various steps of a research project and how an AI can be integrated into these steps. In particular, in Figure 2, we make adjustments to reflect a more practical incorporation of an AI into one's research. (We removed the descriptions of each stage, as they were redundant.) In this figure, orange boxes represent steps wherein we believe a researcher would need to provide significant input in order to have a successful collaboration with an AI.

Some of this involvement entails major oversight over the AI's output. For instance, we found that the initial ideas

[5] Note from the human researchers: The Ehrenberg et al. (2002) reference is an outright hallucination, while other references contain small errors, such as incorrect page numbers.

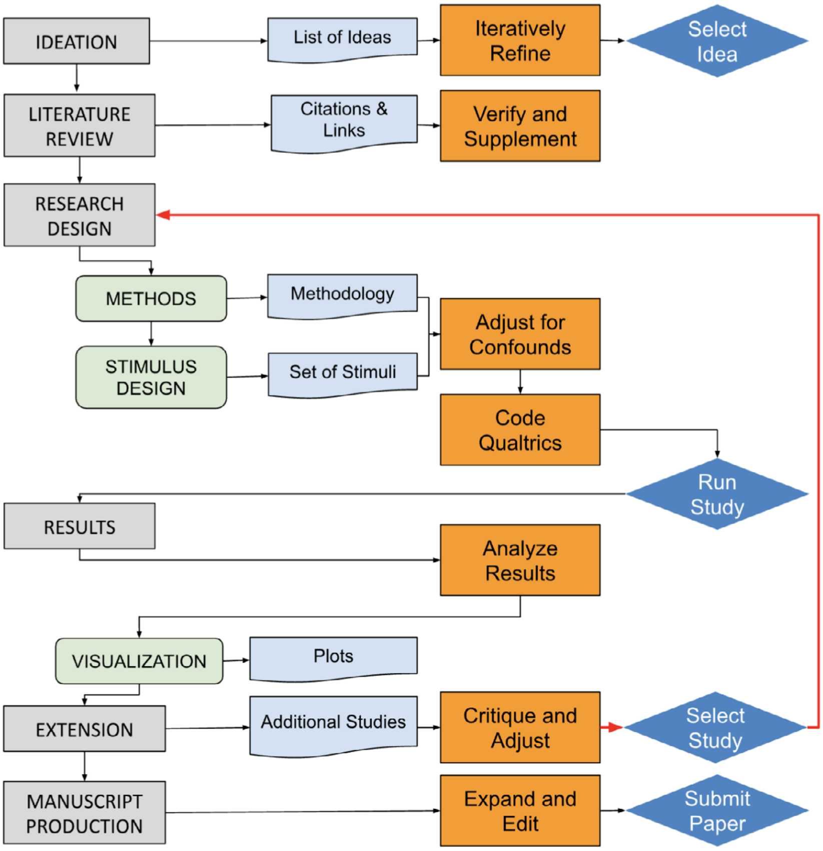

FIGURE 2 Flowchart of our recommended AI-enabled research process.

suggested by an AI model will not likely be immediately usable and will thus require meaningful iteration and refinement from the researcher. For other portions of the research process, we have altered this process to eliminate AI involvement entirely. For instance, based on our disappointing experiences attempting to involve the AIs in analyses, we recommend that the researcher(s) perform this step entirely on their own.

Thus, Figure 2 represents a prescriptive procedure for behavioral researchers using AI models based on the capabilities we observed in our experience.

Persistence of noted limitations

Throughout this exercise, we noted some key limitations in these AI models as tools for a behavioral researcher. We expect that, given the pace of AI development, some of these limitations could be resolved relatively soon, under certain circumstances, while others would require seismic shifts in the nature of these models.

Currently addressable limitations

One of the most important limitations we observed in these AI models was their inability to perform an adequate review of the literature. This is a crucial element of a research project, both at the ideation stage and at the writing stage. We expect that this limitation in large part derives from these models being unable to train on papers in many top journals in the field because they are closed-access. This problem, however, is not due to an inherent technical limitation of these AI models. Rather, it is a question of licensing. In principle, were major academic publishers, such as the Oxford University Press or John Wiley and Sons, to strike a licensing deal with the developers of any AI model, this model would then be much more adequately trained to support academic inquiry. This may, in fact, provide an opportunity for large academic institutions that possess the resources and motivation to develop and train such AIs in-house. However, the financial viability of such a model is not yet clear and could present meaningful operational challenges to the academic ecosystem. That is, there may be significant opportunity costs to publishers having their content integrated into and accessed through these AI models.

Another limitation that could be amended relatively quickly is the AIs' inability to perform reliable data analyses. In particular, solutions are being developed that can take descriptions of analysis methods, write code, and then execute the code on datasets. Tools like LangChain and GitHub Copilot allow multiple AI analysis steps to be generated in code, executed, and then returned as a consolidated solution. These kinds of solutions will likely be developed into custom GPT tools suitable for researchers in the near future (Pokhrel et al., 2024).

More challenging limitations

Other limitations of the AI models in an academic context are that they are less readily surmountable due to the technical nature of these models. These are largely to do with the AI models' inadequacies as project partners. For instance, we noted how a researcher would need to apply discernment before pursuing a research direction suggested by an AI. This may be due to a fundamental limitation of their architecture—they are capable of inductive and deductive reasoning, but may fail at abduction, because abduction requires that the AI generate specific testable hypotheses from incomplete data (Bell et al., 2024; Larson, 2022; Littlefield, 2019; Medianovskyi & Pietarinen, 2022; Nepomuceno-Fernández et al., 2022; Takyar, 2023). This represents a meaningful limitation, for as Charles Sanders Peirce, considered the philosophical progenitor of abduction, noted abduction "is the only logical operation which introduces any new idea." Likewise, we found that the AI models were unable to develop justified, theoretically relevant moderators for the effect we pursued. This is because an AI cannot possess such "taste." Rather, it is simply probabilistically selecting words in response to a prompt, instead of appealing to some larger agential function. This mode of processing is inherent to AI models in their current forms. Thus, without a sizable change in the structure of these models, it is unlikely that they can be relied on for outputs that resemble major executive functioning.

This limitation is particularly relevant to the literature review stage of research. That is, even if AI models are granted access to currently closed-access journals, they may not exercise sufficient discernment in these reviews. For instance, AI models will often eschew referring to higher quality primary sources, despite their being open access, and instead refer readers to more diluted secondary sources for information (Wong, 2024). As such, while access to closed-access outlets would improve many aspects of working with AI models, it will not necessarily solve issues related to their discernment.

By their nature, AI models have a particular way of producing their output. We focused on one particular project to understand how AI can be used at various stages in behavioral research, meaning we did not have a large set of outputs among which to observe systematic patterns. However, prior work has documented that when generating output for behavioral research designs, biases can affect these designs (Bail, 2024; Demszky et al., 2023). While the researcher can verify the internal validity of such designs, their scaled usage may result in systematic biases across the field in the types of stimuli used, for instance.

An additional limitation pertains to something we did not attempt in this exercise: synthesizing data through AI. Instead, we opted to use real participants recruited through Prolific and a university research pool. While it would have been tempting to have "closed the loop" and allowed the AI models to synthesize data, current work suggests that synthesized participants can be useful in piloting ideas, but not for more legitimate testing (Abdurahman et al., 2024; Hutson, 2023). Thus, to credibly ascertain the viability of the AI-generated idea in this exercise, we felt it necessary to use actual human participants.

Finally, another area where we did not test the capabilities of these models was in text analysis. This was simply because our experiments did not involve any open-text measures. However, while recent research is encouraging, current work on the subject of whether AI models can consistently analyze text data finds results both supporting (Kozinets & Seraj-Aksit, 2024) and cautioning against the indiscriminate use of AI models in this capacity (Chew et al., 2023; Tai et al., 2024; von Rütte et al., 2024; Yadkori et al., 2024).

Relevance of findings outside of experimental behavioral research

We pursued this exercise largely from the perspective of behavioral researchers such as ourselves. Yet, we believe that much of what we found likewise has relevance outside of this domain.

First, we think that much of our experience will likewise be relevant to researchers who take a qualitative approach. In particular, a frequent tool used by such researchers is natural language processing. It would clearly be tempting for these researchers to pass their datasets to an AI model and ask for sentiment analyses. Yet, as our investigation shows, they may not receive a usable output. The odds of there being a meaningful error in the output are too high for it to be independently reliable. In the case of traditional experimental research, the researcher can at least audit an AI's analytical output. However, the way in which an AI engages in natural language processing will almost necessarily be a black box, making the reliability and consistency of an AI's output in this capacity unverifiable (Chew et al., 2023; Tai et al., 2024; von Rütte et al., 2024; Yadkori et al., 2024).

AI is also increasingly being used as a research tool in natural science domains, for instance, in genetic research (Vilhekar & Rawekar, 2024). It is crucial to note that these researchers utilize AI models designed for precision, rather than the AIs we tested. This means that researchers in natural sciences need to select the correct, likely more specialized model regardless of the intelligence exhibited by popular AIs like ChatGPT.

We also believe our findings are highly relevant to researchers in Humanities-related fields, such as History.

The prevalence of AI "hallucinations," whereby they confidently present patently false information, is already well documented (Alkaissi & McFarlane, 2023), and the ways in which that could seriously undermine an assertion from a historian, for instance, are clear. However, our work highlights yet another potential drawback for these types of researchers. In particular, we find that the AI models had major limitations in their ability to position work in existing literature, given their limited access to that literature. This is especially crucial to a researcher wanting to understand an idea's novelty. While behavioral researchers can simply scan top journals in their field for a research question and relatively easily ascertain its novelty, this would be significantly more challenging for a researcher in a field that instead relies on entire books. A historian would have a significantly more challenging time validating an idea's novelty, meaning they would instead need to turn to their own knowledge base or perform their own survey of the literature, rendering the AI's assistance in this matter largely moot.

Implications for the academic ecosystem

Our findings regarding the capabilities of these AI models have implications for multiple facets of the academic ecosystem.

Development of AI as a research tool

One way the contributions of typical AIs to behavioral research are currently stymied is by their inability to access closed-access journals. The situation suggests an undesirable equilibrium of mutual dampening: The AIs can't progress because they can't access the *opera omnia* of a given discipline, and the science that relies on these models cannot progress because it lacks effective research tools. One result may be an increased impetus for researchers to publish in open-access journals to improve the odds of their research being cited by AI models assisting other researchers.

A second possibility would entail the development of large, well-resourced organizations that could negotiate access for AI models to otherwise closed-source publications. This might involve large universities or consortia—the same ones that buy subscriptions to the closed-access journals currently—or specialized AIs that recognize the value of these publications to their AI's success in this domain. However, it is worth noting that both of these may undermine the democratization of behavioral science that some envision will come with the integration of AI into an academic context.

A final possibility involves a company that provides paid access to its AI as a research tool and uses that revenue to fund its paid subscriptions. Such a model may have the perverse effect of creating an increasingly adept AI research tool while simultaneously weakening researchers' ability to competitively conduct their research without using the tool.

Authors

We depict at length the prospects for AI in behavioral research. There are some tasks that it can currently do well, other tasks that it cannot, and many where its performance is variable. Beyond these considerations of performance, though, it is also important to consider how these tools ought to be handled from an open science perspective. Good science requires precision and attribution. Thus, we urge authors to communicate their research process, including AI involvement, to readers and editorial teams. We suggest that authors document their AI use, perhaps using a framework like that in Figure 2, including the AI used and possibly even prompts and outputs in a Appendix S1.

Reviewers

Our findings show that it would be of great disservice to any field to allow an AI model to lead a review. At best, we found that these models had a superficial understanding of behavioral research. Just as we needed to guide the AI model for a usable direction to pursue, so too would a reviewer need to have sizable oversight over an AI to ensure a coherent review. However, in order for a reviewer to successfully implement this oversight, they would need to have thoroughly read and understood the paper they are reviewing, effectively eliminating the value in using the AI in a critical capacity.

Moreover, these models are inherently probabilistic. This is acceptable for a researcher engaging in ideation, since if they do not like an output, they can always request another. However, a reviewer offloading work to the AI would be doing so to generate an evaluation. As stated, if they already had an evaluation of their own against which to judge the AI, they would not be using the AI this way. Thus, a reviewer utilizing an AI model would effectively be leaving the fate of the paper up to some degree of randomness, doing a disservice to the authors and the field.

Furthermore, as authors continue to adopt AI into their writing, the styles in these outputs will naturally be reflective of the way in which the AI model writes. Reasonably then, an AI review could respond favorably to this style, resulting in an incentive for authors to offload their writing entirely to AIs, at the expense of their unique contributions and inputs to this writing.

This is all made especially troubling by the already rising prevalence of academic reviewers utilizing AIs (Singh Chawla, 2024). For these reasons, we advocate for reviewer policies, such as that of the journal *Management Science*, which places clear, absolute responsibility for the review on the reviewer. This type of policy does not prohibit a reviewer from using AI to consolidate their notes and construct a more cogent review, for instance, but it does, however, help protect authors from both AI biases, as well as intellectual property concerns that come with their work being uploaded to an AI.

Editorial teams

No doubt, AI is already being used in behavioral research. How extensively and effectively is an open question. Our research suggests that AIs are currently dull tools for most tasks. But they will undoubtedly get better. It will be important for editorial teams to perform periodic research among their authors and reviewers regarding if and how they are using AI. Good data are crucial for sound policy. We encourage editorial teams to be open to research that has been done with AI assistance but to consider policies that would require authors to provide full disclosure regarding their AI use (see the *Journal of the Association for Consumer Research*'s policy on AI use for a positive example). It may also be prudent to prohibit AI use for specific tasks where AI results are poor and impactful and where authors are often given the benefit of the doubt, such as reporting/citing literature and data analysis. Nevertheless, review teams and editorial staff will need to remain vigilant against hallucinatory citations and bogus analyses lest they contribute to the poisoning of the well.

Our preceding section argues forcefully that AIs should not be used to determine the intellectual content of the peer-review process, though assistance with organizing and writing is reasonable. Ensuring this constitutes another crucial role for editorial teams. They should be vigilant in protecting authors from the inappropriate use of AI by reviewers. This could be done both through journal policy and through increased evaluation of the validity of cited work, the meaningfulness of proposed alternative mechanisms, the appropriateness of analytic methods, the coherence of results, and the realism of proposed implications. It is likely that while AI may improve efficiency for some researchers, it will substantially increase challenges for editors of that work.

PhD training

We found that there are indeed some ways for these AI models to assist with research. For instance, we found them capable of usable experimental designs and helpful in the ideation stage. Thus, current and future PhD students can be reasonably expected to use these tools going forward. Yet, this raises an important question of what this means for PhD training. As we found, it is essential that the researcher using an AI in this manner exercise their taste and judgment over the AI outputs. As such, while it may be prudent to train PhD students in how to use AI to further their research and improve their productivity, it is crucial that they continue to receive holistic training that will enable them to judge when and how they should augment or override the AI's directions. It is hard to imagine this being accomplished without broad exposure to the basic literature and research methods. Some of this exposure can actually come from intentional, targeted interactions with AI models (e.g., a PhD student asking for help navigating research on persuasion). However, these limited use cases certainly do not undermine the value in the field's current model of PhD training.

Research assistance

Behavioral research is often conducted with considerable assistance, both from research assistants and PhD students. This can be at simple tasks, such as compiling a list of eligible stimuli, or the more nuanced work of programming an experiment on Qualtrics. We found that AI models can, in some instances, perform this type of labor. For instance, the stimuli we requested from the AI models were largely usable. Yet, we do not believe that AI models will make this type of research assistance obsolete. For instance, there are concerns around the capabilities of these models to reliably code data on many constructs of interest (e.g., creativity), a common task for research assistants (Chew et al., 2023; Tai et al., 2024; von Rütte et al., 2024). Moreover, they cannot perform in-person work, such as managing a mock store in an in-person lab study. Thus, we believe that research assistance will still be necessary in future research.

Future directions for investigating AI capabilities in research

Throughout this process, we aimed to provide a holistic analysis of how AIs can be incorporated into the behavioral research process. Yet, there is additional important work to be done by engaging in more focused evaluations of some of the steps we have highlighted here.

One step that we believe would benefit from increased attention is the ideation stage. We attempted to follow commonly accepted best-prompting practices in our approach (Shaer et al., 2024) to provide generalizable documentation of the AI models' performances. However, it could be that a more tailored prompting approach is useful in research contexts to extract the best possible ideas from AI models. For instance, providing the AI with papers that represent the author's tastes (without violating copyright law) may help improve the relevance of AI-suggested ideas.

An additional area that we would encourage other researchers to investigate is AI's capability to generate non-text stimuli. As noted, we found the visual stimuli generated by these models to be unusable in our experimental context. However, as these models improve, their abilities to generate images, audio, and other forms of non-text stimuli will likely also improve. Thus, while it is difficult to forecast when this will become a germane area of inquiry, when it does, understanding how to best prompt AI models for effective, unbiased stimuli of all sorts will be highly valuable.

Relatedly, we observed the capabilities of these AI models in generating stimuli for a specific research question. While we found these models to present generally usable stimuli, after a few researcher adjustments, they may not be capable of universally producing valuable experimental stimuli. Thus, we also suggest that future research examine the contexts under which AI models generate more or less effective stimuli. For instance, AI models may prove less effective at generating stimuli that have an element of provocativeness, given guardrails in their programming.

Another area that would benefit from further research would be examining a wider breadth of available models and systematically evaluating them on the research stages we have laid out. We have performed this exercise for only three mainstream AI models in the interest of providing an initial, comprehensive step in understanding how these models function in a research context. However, it may be that models that exist on the periphery could exhibit surprising capabilities on some of the steps we have examined here.

Finally, we recommend targeted research on the manuscript production stage. We found the models to be highly resistant to writing lengths appropriate for a journal submission. Moreover, what they did write was fairly superficial. However, having an AI model perform this step well could represent an extremely meaningful time-saving mechanism for researchers. Thus, research on how to effectively utilize AI models for manuscript drafting would be particularly valuable.

More generally, the focus of our work was on utilizing AI as completely as possible to conduct a representative behavioral research project. By contrast, we believe work that focuses more deeply on specific stages, especially as these models evolve, will be highly valuable to behavioral research moving forward. We hope that our work can provide one initial benchmark against which future advances in AI capabilities may be judged.

Final recommendations

In light of our experience, we have certain specific recommendations to researchers wanting to incorporate AI models into their processes.

First, we hope to encourage a high skepticism for the outputs of these models. That is, these models are often very fast and very impressive with their outputs, but that should not be mistaken for absolute accuracy. Researchers using these models should thus adopt a policy of treating the AI outputs as starting points and suggestions, engaging themselves in a practice of verification and refinement. These steps may, at least initially, require more rather than less time and effort.

Second, we urge editorial teams to consider and study the presence of AI in research and set policy decisions accordingly. From an evaluation perspective, it may be necessary to treat AI research papers with additional skepticism, knowing that work conducted with the assistance of AI is liable to have serious mistakes. Ideally, however, researchers could agree to a system of noting which aspects of their work were done with AI assistance, so that added attention would only need to be focused on these “cyborg” sections. Moreover, as noted, we recommend to editorial boards of journals to largely prohibit the use of AI on the side of reviewers as critical tools. This is to preserve fairness for authors in the review process.

Third, we encourage researchers to reflect on the evolving nature of their role in a research project in light of these AI advancements. These tools can do a great deal of legwork and act as useful ideation tools. However, as we find, the researcher still has a vital place in the process, acting as a director and critic of the AI, not an equal partner.

Finally, we note that this work primarily focused on a question of whether AI *can* perform the role of the behavioral scientist. A perhaps equally germane question is one of whether it *should*. That is, we, and likely our readers as well, take a great deal of pride in the work we do as researchers. The specific steps that bring each of us joy (and angst) as researchers are likely as varied as the research in which they are used. Nevertheless, a focus on only the capabilities and efficiencies of AI may miss the point: As these tools evolve, it will be up to each individual researcher to decide for which steps of the research process they want to become a cyborg behavioral researcher, and for which they would like to remain simply human.

ACKNOWLEDGMENTS

We thank Paul Andrew Blythe, Michael Haenlein, Kelly Hewett, Chris Janiszewski, Rich Lutz, Peter McGraw, Asaf Mazar, Vicki Morwitz, Stacy Wood, and Kiwoong Yoo for their helpful and insightful feedback on earlier drafts of the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Geoff Tomaino https://orcid.org/0000-0002-0034-1679
Alan D. J. Cooke https://orcid.org/0000-0002-5106-6179
Jim Hoover https://orcid.org/0000-0001-5432-6530

REFERENCES

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*, 245. https://doi.org/10.1093/pnasnexus/pgae245

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Amankwah-Amoah, J., Abdalla, S., Mogaji, E., Elbanna, A., & Dwivedi, Y. K. (2024). The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda. *International Journal of Information Management*, *79*, 102759. https://doi.org/10.1016/j.ijinfomgt.2024.102759

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Barnes, T. J. (1998). A history of regression: Actors, networks, machines, and numbers. *Environment and Planning A: Economy and Space*, *30*(2), 203–223. https://doi.org/10.1068/a300203

Bell, J. J., Pescher, C., Tellis, G. J., & Füller, J. (2024). Can AI help in ideation? A theory-based model for idea screening in crowdsourcing contests. *Marketing Science*, *43*(1), 54–72. https://doi.org/10.1287/mksc.2023.1434

Bray, J., Johns, N., & Kilburn, D. (2011). An exploratory study into the factors impeding ethical consumption. *Journal of Business Ethics*, *98*(4), 597–608. https://doi.org/10.1007/s10551-010-0640-9

Calonge, D. S., Smail, L., & Kamalov, F. (2023). Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. *Journal of Applied Learning & Teaching*, *6*(2), 346–357.

Caminiti, S. (2023). *The more workers use AI, the more they worry about their job security, survey finds*. CNBC.

Cao, C., Cao, X., Cashman, M., Kumar, M., Timoshenko, A., Yang, J., Yu, S., Zhang, J., Zhu, Y., & Wernerfelt, B. (2019). How do successful scholars get their best research ideas? An exploration. *Marketing Letters*, *30*, 221–232.

Carrington, M. J., Neville, B. A., & Whitwell, G. J. (2010). Why ethical consumers don't walk their talk: Towards a framework for understanding the gap between the ethical purchase intentions and actual buying behaviour of ethically minded consumers. *Journal of Business Ethics*, *97*(1), 139–158. https://doi.org/10.1007/s10551-010-0501-6

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv* preprint arXiv:2306.14924.

Chui, M., Yee, L., Hall, B., Singla, A., & Sukharevsky, A. (2023). *The state of AI in 2023: Generative AI's breakout year*. McKinsey & Company.

Clark, D. (2024). *Academic publishers and AI do not need to be enemies*. Times Higher Education (THE).

Dembrower, K., Crippa, A., Colón, E., Eklund, M., & Strand, F. (2023). Artificial intelligence for breast cancer detection in screening mammography in Sweden: A prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, *5*(10), e703–e711.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5

Eastwood, B. (2024). *The who, what, and where of ai adoption in America*. MIT Sloan.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from Organization Science, accounting, marketing, MIS, and related disciplines. *The Information Society*, *20*(5), 325–344. https://doi.org/10.1080/01972240490507974

Errica, F., Siracusano, G., Sanvito, D., & Bifulco, R. (2024). What did I do wrong? Quantifying LLMs' sensitivity and consistency to prompt engineering. *arXiv* preprint arXiv:2406.12334.

Ferraro, C., Demsar, V., Sands, S., Restrepo, M., & Campbell, C. (2024). The paradoxes of generative AI-enabled customer service: A guide for managers. *Business Horizons*, *67*(5), 549–559. https://doi.org/10.1016/j.bushor.2024.04.013

Fischer, E., & Otnes, C. C. (2006). Breaking new ground: Developing grounded theories in marketing and consumer behavior. In Russell W. Belk (Ed.), *Handbook of qualitative research methods in marketing* (pp. 19–30). Edward Elgar Publishing.

Haan, K. (2023). *24 top AI statistics and trends in 2024*. Forbes.

Hutson, M. (2023). Guinea pigbots. *Science*, *381*(6654), 121–123. https://doi.org/10.1126/science.adj6791

Johnston, J. M., Pennypacker, H. S., & Green, G. (2020). *Strategies and tactics of behavioral research and practice*. Routledge.

Kite, M., & Whitley, B. E. (2012). *Principles of research in behavioral science*. Routledge Academic.

Kozinets, R. V., & Seraj-Aksit, M. (2024). Everyday activism: An AI-assisted netnography of a digital consumer movement. *Journal of Marketing Management*, *40*(3–4), 347–370. https://doi.org/10.1080/0267257x.2024.2307387

Larson, E. J. (2022). *The myth of artificial intelligence*. Harvard University Press.

Littlefield, W. J. (2019). *Abductive humanism: comparative advantages of artificial intelligence and human cognition according to logical inference*. Master's thesis. Case Western Reserve University.

Medianovskyi, K., & Pietarinen, A. V. (2022). On explainable AI and abductive inference. *Philosophies*, *7*(2), 35.

Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., & Myers, B. (2024). Using an LLM to help with code understanding. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, *2*, 1–13. https://doi.org/10.1145/3597503.3639187

Nepomuceno-Fernández, A., Soler-Toscano, F., & Velázquez-Quesada, F. R. (2022). Abduction from a Dynamic Epistemic Perspective: Non-omniscient Agents and Multiagent Settings. In S. Velázquez-Quesada (Ed.), *Handbook of abductive cognition* (pp. 1–29). Springer International Publishing.

Parguel, B., Benoît-Moreau, F., & Larceneux, F. (2011). How sustainability ratings might deter 'greenwashing': A closer look at ethical corporate communication. *Journal of Business Ethics*, *102*(1), 15–28. https://doi.org/10.1007/s10551-011-0901-2

Pokhrel, S., Ganesan, S., Akther, T., & Karunarathne, L. (2024). Building customized chatbots for document summarization and question answering using large language models using a framework with OpenAI, lang chain, and Streamlit. *Journal of Information Technology and Digital World*, *6*(1), 70–86.

von Rütte, D., Anagnostidis, S., Bachmann, G., & Hofmann, T. (2024). A language Model's guide through latent space. *arXiv* preprint arXiv:2402.14433.

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*(6), 1–17.

Shaer, O., Cooper, A., Mokryn, O., Kun, A. L., & Ben Shoshan, H. (2024). AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference* *on Human Factors in Computing Systems* (pp. 1–17). Association for Computing Machinery.

Singh Chawla, D. (2024). Is CHATGPT corrupting peer review? Telltale words hint at AI use. *Nature*, *628*(8008), 483–484. https://doi.org/10.1038/d41586-024-01051-2

Stremersch, S. (2024). How can academics generate great research ideas? Inspiration from ideation practice. *International Journal of Research in Marketing*, *41*(1), 1–17.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 1168. https://doi.org/10.1177/16094069241231168

Takyar, A. (2023). *AI in product development: Use cases, benefits, solution and implementation*. LeewayHertz.

Vilhekar, R. S., & Rawekar, A. (2024). Artificial intelligence in genetics. *Cureus*, *16*(1), e52035. https://doi.org/10.7759/cureus.52035

Wolfram, S. (2023). What is ChatGPT doing … and why does it work? Stephen Wolfram - Writings. https://www.stephenwolfram.com/

Wong, M. (2024). Generative AI can't cite its sources. *The Atlantic*. https://www.theatlantic.com/technology/archive/2024/06/chatgpt-citations-rag/678796/

Yadkori, Y. A., Kuzborskij, I., György, A., & Szepesvári, C. (2024). To believe or not to believe your LLM. *arXiv* preprint arXiv:2406.02543.

Zúñiga Salazar, G., Zúñiga, D., Vindel, C. L., Yoong, A. M., Hincapie, S., Zúñiga, A. B., Zúñiga, P., Salazar, E., & Zúñiga, B. (2023). Efficacy of AI chats to determine an emergency: A comparison between OpenAI's ChatGPT, Google bard, and Microsoft Bing AI chat. *Cureus*, *15*(9), e45473. https://doi.org/10.7759/cureus.45473

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology*, *94*(5), 883–898. https://doi.org/10.1037/0022-3514.94.5.883

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tomaino, G., Cooke, A. D. J., & Hoover, J. (2025). AI and the advent of the cyborg behavioral scientist. *Journal of Consumer Psychology*, *35*, 297–315. https://doi.org/10.1002/jcpy.1452