

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/176247>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems

Anjana Susarla, Ram Gopal, Jason Thatcher, and Suprateek Sarker

## 1. Introduction

Generative Artificial Intelligence (AI) broadly refers to a class of AI models that apply neural network logic to a variety of structured and unstructured data to generate seemingly new content. While generative AI techniques have been around for some time, the release of ChatGPT unleashed an explosion of conversations in the popular press, in online forums, and in academic circles. ChatGPT aroused concern because it made accessible the broad and open set of functionalities of generative AI models through simple queries to a browser-based interface. Within two months of its beta release, ChatGPT attracted over 100 million users.

Conversations about what generative AI tools mean for human learning, knowledge creation, and the future of work have consumed traditional news media and social media. Discussions have ranged from dismay at the new Promethean moment that generative AI represents for humanity (Freidman, 2023) to extolling the possibilities these tools present for renewed creativity and positive social change (Larsen and Narayan 2023). As conversations about generative AI have been echoed and amplified in offline and online forums, there has been a parallel rapid proliferation of generative AI-powered applications and corporate actions, resulting in what some are calling an AI arms race.

In our collective experience, we have never seen such an explosion of applied and academic interest in an IT tool. We find this surge of interest exhilarating, because we appear to be on the cusp of change in academic systems, such as those involved in developing or reviewing a manuscript; disconcerting, because generative AI and their capabilities are so often mischaracterized; and rich with opportunity, because understanding the changes these tools might bring requires an understanding not only of the technology and the data that inform it, but also of the ways that academics will choose to construct meaning, rules, and routines around the application of generative AI in their lives.

Academic conversations about generative AI have moved beyond bemoaning the decline of learning to thinking more critically about how scholars will and should use generative AI. They may help automate aspects of writing or serve as an aid to the creative process, but can they replace our accumulated experience or insights on the problems we have grappled with? Serious questions have also surfaced about the ability of humans to detect work supported by these tools. Nature reported on a recent study that tested our ability to spot artificial research-paper abstracts generated by ChatGPT (Else 2023). These abstracts were evaluated by a plagiarism detector, an AI-output detector, and by a group of blinded human reviewers (Gao et al. 2022). Soberingly, the blinded human reviewers could correctly identify only about 68% of the abstracts that were generated by ChatGPT, but incorrectly identified 14% of the original

(human-generated) abstracts as being generated by ChatGPT. As AI-generated content becomes indistinguishable from human-generated content, it is imperative that we establish norms, policies, and procedures regarding the use of these tools in the research process.

Recognizing that generative AI can be a double-edged sword for scholarly work, this editorial<sup>1</sup> seeks to start a conversation about how we believe generative AI tools, including but not limited to ChatGPT, should be used by the Information Systems (IS) community to conduct, and publish research. In doing so, we do not adopt a naïvely optimistic perspective, one that would normalize such technologies as nothing more than “calculators”, at least in the short run. Neither will we adopt a position of deep pessimism, as espoused by notable scholars such as Noam Chomsky, who believes that ChatGPT-style systems “may have some value for something,” but “it’s not obvious what,” and characterizes ChatGPT as “basically high-tech plagiarism” that will lead to avoidance of learning (Chomsky et al. 2023). Rather, we seek to offer a balanced rumination on the changing nature of scholarly work and offer thoughts on how the IS community can cautiously embrace the changes. We acknowledge that we take some liberties in speculating about the possibilities and applications of generative AI in research and given the pace of change in the technology and tools, some of our assertions and suggested guidelines may eventually seem misplaced. However, we anticipate that by offering these preliminary thoughts based on our experience as authors, reviewers, and editors, we lay a foundation for more refined applications of generative AI to support scholarship that will appear over time in journals such as *Information Systems Research (ISR)*.

In the following pages, we first provide a brief overview of the development of large language models that can be adapted to a variety of downstream tasks – these are at the heart of tools such as ChatGPT<sup>1</sup> with the caveat that this is descriptive about what such models can do, and not intended to advocate for the broader adoption of such models in research. Next, based on the published literature and our “analysis” of selected papers, we highlight some of the possible benefits of using these tools for developing, reviewing, and reading/understanding papers;

---

<sup>1</sup> Though this piece constitutes an editorial, it diverges from the customary structure, format, and content typically found in ISR. The staggering pace of recent AI advancements has already begun to instigate seismic shifts in research conduct and communication. It is evident that all journals, including ISR, will experience the reverberations of this paradigm shift. As a community, it is incumbent upon us to unite and formulate guidance for the future, ensuring that ISR remains the preeminent platform for publishing and promoting innovative, transformative research within information systems. The impetus for composing this editorial in its current format and structure is the pressing need to furnish IS scholars, particularly junior researchers, with direction regarding the appropriate employment of generative AI in their ISR-targeted research. In doing so, we endeavor to outline potential trajectories for the journal to continue fulfilling its core mission. We encourage readers to perceive this piece not as the definitive word on the subject, or as advocacy for the use of generative AI in research, but as the genesis of conversations that will shape our collective future.

<sup>1</sup> We do not sharply distinguish between LLMs and generative AI models since both offer the promise of broad societal deployment. For instance, it is not inconceivable that researchers might employ tools such as DALL-E or Midjourney to study image analytics or employ video generation tools that can augment existing datasets and might even serve as an alternative to current models of synthetic control. It is also possible to use generative AI tools to assist in coding tasks or in data collection. The same set of principles that we articulate henceforth in the context of LLMs will apply to such modes of generative AI as well.

then, we highlight concerns about using these generative AI tools, as they stand today, in scholarly research activities, that could be dysfunctional or misleading. We end our editorial with some reflections and some thoughts for the future.

## 2. Language Models and their Evolution: A Brief Overview

In this section, we provide a brief introduction to the language models, that are the focus of the editorial, as they relate to scholarly work. While some readers may find this section useful, readers with advanced knowledge in this area may wish to skip this section.

The origins of language modeling can be traced back to the early days of statistical natural language processing (NLP), which emerged in the 1950s and 1960s. Early language models relied on probabilistic models that used statistical techniques to estimate the probability of a given word or string of words occurring in a language (Russell and Norvig 2003). These early models typically used *n-grams*, or sequences of *n* words, and utilized straightforward frequency-based statistics to estimate the probability of each *n-gram* occurring in the language.

The evolution from *n-gram* based models to *word embeddings* began in the mid-2000s with the development of the Word2Vec algorithm (Mikolov et al. 2013), which introduced a new method for capturing semantic knowledge. A further breakthrough in language modeling came with the introduction of the *transformer* architecture (Vaswani et al. 2017). The transformer architecture paved the way for the introduction of *Large Language Models* (LLMs) (Wei et al. 2022), which are trained on massive amounts of text data from diverse sources such as books, news articles, web pages, and social media posts. With advances in model building and model training, we have entered a new era with the emergence of what some have dubbed *foundation models* – models that are trained on massive amounts of data and tuned with billions of hyperparameters that can be adapted to a wide range of tasks, including some tasks that the model was not originally trained on (Wei et al. 2022). Generative models are now capable of creating realistic and creative content from various inputs, such as text, images, audio, or video. In the future, such models may be able to produce novel and complex content, such as interactive stories, immersive simulations, or personalized experiences. For example, a large language model for protein engineering was trained on 280 million protein sequences from over 19,000 protein families and used to identify protein sequences with a predictable function across large protein families (Madani et al. 2023), which the researchers likened to a language model that generates grammatically and semantically correct natural language sentences on diverse topics.

Given the increasingly general-purpose nature of these models, adapting models such as ChatGPT or DALL-E to new tasks can be done with little to no coding and can be as easy as describing a task in simple language. Using a chatbot such as ChatGPT or image generation tool such as DALL-E is as simple as entering requirements through a browser interface. Furthermore, humans can extend the functionality of these models by fine-tuning prompts that are phrased as instructions. Cognitive neuro-linguists have questioned whether the capabilities of generative AI truly demonstrate a human-like understanding of language (Mitchell and Krakeur 2023). Experts on fairness and transparency in AI have also warned that general-purpose AI models

trained on vast repositories of online data are little more than “stochastic parrots” (Bender et al. 2021), with large volumes of training data gleaned from the Internet potentially encoding hegemonic and ethnocentric worldviews, and that people using these models may be misled by the efficiencies of these models to overlook the societal biases and other problematic issues arising from the training data.

### **3. The Role of Generative AI in Enabling High-Quality Scholarly Work**

Generative AI does seem to offer opportunities to enhance specific areas of research, namely (i) problem formulation and research design, (ii) data collection and analysis, (iii) interpretation and theorization, and (iv) as an aid to composition and writing. We will discuss the opportunities as well as challenges in the next two sections. We briefly review each of these possibilities in this section and the next and illustrate them with examples.

#### **3.1 Generative AI in Problem Formulation and Research Design**

Generative AI tools can be used as a sounding board for emerging ideas. When ideating on research questions or hypotheses, one can enter key questions, concepts, and arguments as prompts, which can then contribute to an iterative process of idea formulation, assessment of the quality, novelty, or state of related research, and idea refinement. For example, if one were to extend Tarafdar et al.’s (2022) introspective study of *Information Systems Research* to examine the entire discipline, one could ask ChatGPT to help by listing the best journals that represent the discipline. We asked ChatGPT – “*what is the best MIS Journal?*” It returned a list of ten journals, missing two of the “AIS Senior Scholars elite journals” and adding an additional, lesser-known journal. We tuned our query to ask “*What is the best MIS journal to start a literature review?*” It returned a subset of five journals, with some explanation, opining that a journal other than ISR was a better starting point. When asked what each journal was good at, ChatGPT opined that ISR was best for technology adoption and implementation research. While the answers were not necessarily correct or complete, *ChatGPT can be used to support finding a starting point for research.*

ChatGPT was also able to offer suggestions for future research related to a recently published paper in ISR (e.g., Fugener et al. 2022), incorporating multiple perspectives that could be accommodated within the study’s setting and analysis. When prompted to redesign Fugener et al. (2022) study to consider algorithm aversion, a theoretical perspective that was not considered in the original paper, ChatGPT responded by proposing an experimental manipulation detailed in Table 1 below. The experimental manipulation suggests a set of nudges such as (i) giving participants the option to decline assistance from the AI vs. not, and (ii) manipulating the degree of delegation to the AI agent, and (iii) measuring participants’ attitudes towards AI. Similar patterns were observed with different prompts asking ChatGPT to redesign the study based on alternative theoretical traditions. The responses in Table 1 illustrate some of the ways in which these tools can augment the process of literature review and research design.

Table 1. Example set of prompts and response for a published ISR Paper (Fugener et al. 2022)

Prompt: “how would you redesign the study to consider algorithm aversion”

Edited and condensed response..

To redesign the study to consider algorithm aversion, the researchers could introduce a manipulation where participants are informed about the potential benefits of using the AI algorithm in the collaborative task, but they are also given the option to decline the AI's assistance and complete the task on their own. This would allow the researchers to measure the extent to which participants are reluctant to use the AI algorithm and whether this reluctance affects the collaborative performance between humans and AI. ...

To test whether algorithm aversion affects delegation decisions and performance, the researchers could manipulate the delegation role of the AI and participants in the task. For example, in one condition, the AI could delegate tasks to the participants, and in another condition, the participants could delegate tasks to the AI. The researchers could then measure the delegation performance and efficiency for each condition and compare them to the control condition where there is no delegation.....

Once a manuscript has been written, authors can potentially use ChatGPT to evaluate key elements of the paper's quality and completeness, and to identify additional studies that may be relevant. To illustrate this point, we describe our experience when we entered the text of Susarla et al. (2020), and asked ChatGPT to summarize the key points. We reviewed the summary provided by ChatGPT to see if the key points extracted by ChatGPT corresponded to what we felt was consistent with our understanding of the paper's thrust. Had the paper been in development, we could have used the summary to review why key points were missed, and others (over-)emphasized.

Further, we would suggest using supervised prompt fine-tuning to ask ChatGPT for insights into how IS researchers with different theoretical and paradigmatic perspectives might respond, thereby providing opportunities to address possible misunderstandings or discrepancies when the work seen from certain perspectives, prior to entering the peer review process. For completeness, one can enter the abstract or literature review and ask ChatGPT to suggest possible missing explanations or relevant papers. To identify possible missing explanations, we entered the abstract of Wong et al (2021) and asked ChatGPT for perspectives that might help to explain the findings. The tool responded with the usual suspects such as Social Learning Theory and the Technology Acceptance Model, but also with Social Identity Theory and Diffusion of Responsibility. When we compared this to the research model, we found elements of each theory in the whistleblowing framework that appeared in the paper. This is quite interesting, given that the Wong et al. paper was not part of the ChatGPT training data.

### ***3.2. Generative AI in data collection and analysis***

Generative AI can provide valuable support in assembling datasets or in identifying deep patterns in the data. Generative AI is well-suited to help researchers match data across archival sources. Manual, time-consuming tasks such as matching names, companies, or other variables can be somewhat automated. Somewhat, in the sense that once the records are matched, the researcher should verify the quality of the results. For example, when compiling a unique

dataset for their work on corporate social performance and data breaches, D’Arcy et al. (2020) drew together information from several sources to match firm data breaches, corporate social performance ratings, and related factors. In doing so, they had to match records that used sometimes inconsistent firm names, which is a detailed and time-consuming process. Generative AI tools such as ChatGPT are well-suited for such pattern-matching exercises. They can be used to identify possible names to search across heterogeneous archival sources, assess the likelihood of a match between records, and assemble a dataset that takes into account **each permutation of each company’s name**. Again, human supervision is required to ensure the search and matching exercise is properly executed, because the generative AI may lack the context or information in the training dataset to correctly match all the records. Researchers (e.g., Mollick 2023) have reported examples where they input data along with a data dictionary into ChatGPT along with a code interpreter to have ChatGPT analyze the results and provide preliminary conclusions. With a code interpreter plugin, ChatGPT also offers extensive visualization functionalities that may prove useful to researchers investigating a variety of topics.

Beyond simple pattern matching, generative AI can be applied to explore text data. Researchers can explore how different assumptions and definitions affect the type, content, and quality of insights derived from patterns manifested in the corpus text. In doing so, they can identify ways to complement analysis. For example, Zheng et al (2022) apply generative AI to surface patterns and critically evaluate those patterns, creating opportunities to confirm or disconfirm findings suggested by the data. When queried about a potential study using word embedding-based methods to study an IS specific phenomenon (see Table 2), ChatGPT suggested comparing across alternative methods such as Latent Semantic Analysis (LSA), or, better still, a hybrid approach that combines different techniques like LSA, topic modeling, sentiment analysis, and network analysis to get a more comprehensive understanding of the nuances of the problem domain. While this is no substitute for human reasoning and cognition, generative AI serves as an aid for a researcher seeking to identify and evaluate alternative approaches to solving problems.

Table 2. Example prompt and response for a recently accepted paper in ISR (Zheng et al. 2022)

<p>Edited and condensed responses for a prompt on how to use machine learning to study the same phenomenon</p> <ol style="list-style-type: none"> <li>1. Classification: A supervised learning algorithm like logistic regression or decision trees could be used to classify sellers into high-quality and low-quality categories based on their refund policies and platform insurance. This classification model could then be used to predict the quality of new sellers based on their policies.</li> <li>2. Clustering: An unsupervised learning algorithm like k-means clustering or hierarchical clustering could be used to group sellers based on their refund policies, reputation, popularity, or other characteristics. This could help identify patterns in seller behavior and reveal potential correlations between refund policies and seller quality.....</li> </ol>
--

With primary data collected directly through methods such as experiments or observational data curated through scraping websites or calls to an API, generative AI can be used to suggest approaches to analyzing and interpreting the data. While generative AI is not a substitute for

having a deep understanding of statistical or visualization tools, it can speed up the process of creating the syntax, estimating the results, and evaluating the quality of the results. Because one has far more control over the data manipulation and analysis with text-based commands than with a point-and-click interface, augmenting human ability to craft and evaluate such code for statistical analysis can provide researchers with opportunities to conduct more granular analysis. For example, a recently published paper in ISR (Oh et al 2023) uses clickstream and social media data to examine the complementary relationship between social media and news sites. While Oh et al.'s analysis appropriately focused on text and numbers relevant to their research question, using generative AI could have afforded opportunities to integrate additional data found in social media posts such as image content, geospatial data, and more.

### 3.3. Generative AI as an aid to composition and writing

Finally, generative AI can be used to support the writing process. One can enter draft text and ask it to assess the structure of arguments or the quality of writing. We asked ChatGPT to assess the quality of our first paragraph in this section. It returned four substantive suggestions for improving clarity, organization, evidence, and acknowledgment of limitations. Improving and shortening writing offers several benefits. It allows the author to independently assess the paper's structure, logic, and readability independently. By improving these fundamental elements of papers, ChatGPT could reduce the demands placed on peer reviewers. Additionally, it facilitates teamwork by improving the quality of the text exchanged between collaborators, before it is shared. ChatGPT can assist with formatting references to specific styles. The examples below, though not exhaustive, provide pointers on using ChatGPT to aid in composition. Generative AI tools can also provide holistic reviews of a paper's grammar, punctuation, complexity, and voice. Unlike widely used line-by-line tools such as Grammarly, generative AI can provide global support for a paper's technical aspects, making it more accessible to a broader audience. For example, as we were polishing this paper, we used ChatGPT to suggest consistent word choices among the four authors, ensuring a consistent voice in describing the opportunities and challenges of ChatGPT for IS research.

Table 3. Example of using ChatGPT in developing a manuscript

Writing Tasks	ChatGPT outputs
Overall writing style	Reviewing for inconsistent word usage
Proof-reading for consistency and clarity	Reviewing for clarity and organization
References	Formatting to approved styles

We would like to emphasize that our support for using generative AI to assist with writing should not be interpreted as support for having generative AI conduct the research or write the paper. Some suggest using tools like ChatGPT to generate draft sections, especially for more structured sections like methods and results. We believe that ceding control of the writing and knowledge-creation process to ChatGPT is a mistake, at least at this point, and we will elaborate more on this in the next section.



## 4. Concerns about Generative AI Tools Detracting from High-Quality Scholarly Work

In this section, we discuss the nature of the influence that generative AI may have on scholarly activities, such as (i) problem formulation, research design, and methodological critique, (ii) manuscript summaries and literature reviews. We critically examine possible undesirable outcomes that may result from applying generative AI tools. The primary concerns that emerge include: limitations and biases in the training data, hallucinations that lead to veracity and reliability concerns, violation of intellectual property rights, and the lack of appropriate depth and coherence in the outputs.

### ***4.1 The use of Generative AI associated with problem formulation, research design, and ,methodological critique***

Birhane et al. (2022) reviewed a large corpus of published work in machine learning research and found that values such as performance, generalization, quantitative evidence, and efficiency dominate as justifications for research. However, these publications rarely mention the negative potential of these technologies or question whether there is a societal need for them. Indeed, there is growing concern that the large-scale models underlying generative AI may prioritize predictive accuracy and performance over transparency, equity, and open access. Furthermore, training on digital traces and internet sources may replicate existing societal biases without adequate safeguards or protections. Word embedding methods, for example, tend to encode significant gender biases (e.g., Yang and Feng 2020). These biases can propagate to downstream tasks that rely on biased word vectors, such as analyzing user-generated content. More broadly, biases in machine learning models can arise from biases in training datasets, and users' perception of model outputs. Beyond these well-known concerns about machine learning, generative AI tools may have additional undesirable implications for scholars involved in developing, reviewing, and summarizing research papers. Because generative AI is trained using published papers, when asked to suggest research questions, these tools may suggest known questions found in the established corpus of a research community's work, rather than forward-looking state-of-the-art questions at the leading edges of a topic. Similarly, when asked to design a research study, ChatGPT may overlook relevant, emerging approaches. Reviews conducted with these tools may selectively prioritize variables, methods, and findings from dominant sub-disciplines and traditions in the training data. For example, reviews provided by ChatGPT may apply standards associated with a dominant method, such as econometrics, to a paper that comes from a different research tradition, such as interpretivism. There may also be significant factual errors in the output of these tools, which we address in our discussion on "hallucinations."

Generative AI models also present *interpretability* concerns due to the type of architecture and the amount of data used to train them. Unlike predictive models commonly associated with AI, generative models learn a probability distribution over an extremely high dimensional space, making it difficult to interpret their outputs. This lack of interpretability can make it difficult to

understand research design suggestions and critiques, as there is virtually no explanation provided for why a particular option is considered suitable while others are not.

As another example, ChatGPT, in response to the prompt, ‘*How do we use mixed methods research design to study value destruction in information technology ecosystems?*’ proposed the sequential use of quantitative and qualitative data. While the proposed approach is valid, it is generic and lacks innovation. Mixed methods are often portrayed as using qualitative and quantitative data (Venkatesh et al. 2013), but they can involve multiple *approaches* that using similar or different types of data. Indeed, a recent study by Sahaym et al. (2022) used the hermeneutic circle to integrate analytical modeling and interpretive case study (involving qualitative data) to shed light on the issue of value destruction – there was no quantitative data involved.

With respect to refereeing conducted with such tools, they are likely to selectively prioritize variables, methods, and findings from some subdisciplines and traditions that are dominant in the training data. The reviews are likely to “parrot” criticisms that are no longer considered valid or legitimate. For example, we asked ChatGPT ‘*Can you provide a critique of the methods used in the study ‘Toward a Theory of Information Systems Security Behaviors of Organizational Employees: A Dialectical Process Perspective’ published in Information Systems Research?*’ by Karjalainen et al. (2019). Excerpts from the response received are presented Table 4 (problematic issues in bold, hallucinations which we discuss later are underlined):

**Table 4: Critique by ChatGPT of a Qualitative Study**

<ol style="list-style-type: none"><li>1. ....</li><li>2. Sampling: The study uses a <b>convenience sample</b> of employees from <u>two organizations</u>, which may <b>limit the generalizability</b> of the findings to other organizations and contexts.</li><li>3. Data Collection: The study uses a <u>mixed-methods approach with a survey and interviews</u> to collect data. While this approach provides rich data and allows for triangulation of findings, the data may be subject to <b>social desirability bias</b> and the interviews may not fully capture the nuances of security behaviors.</li><li>4. Operationalization of Variables: The study operationalizes security behaviors <b>using a multidimensional scale that includes cognitive, affective, and behavioral aspects of security</b>. While this scale provides a comprehensive view of security behaviors, <b>the measurement of security behaviors may still be subject to biases</b> and limitations.</li><li>5. Data Analysis: <u>The study uses structural equation modeling (SEM) to analyze the survey data and thematic analysis to analyze the interview data</u>. While these are appropriate methods for the research questions, the findings may be <u>subject to the assumptions of SEM</u> and the limitations of thematic analysis.</li></ol>
---

The critique offered by ChatGPT of this qualitative, interpretive paper has characteristics similar to those expected of a novice, unqualified reviewer. The critique misses the idea of “theoretical sampling,” frames the study as one involving “measurement” and “multidimensional scale”, and repeats concerns of “bias” and lack of “generalizability” that are not directly applicable to interpretive work. A reviewer, who is not trained in the interpretive tradition and who uses ChatGPT to aid her review is likely to submit a report that is inconsistent with the standards

associated with the research tradition to which the paper belongs. In addition, there are significant factual errors (the underlined texts) that we will delve into next.

#### **4.2. The use of generative AI in conducting manuscript or literature reviews**

The most pressing challenges are from hallucinations, and concerns regarding interpretability, consistency, intellectual property, and depth of the resulting analysis. In Natural language Processing (NLP), a hallucination is often defined as "generated content that is nonsensical or unfaithful to the provided source content"; it is a confident response by an artificial intelligence system that does not seem to be justified by its training data (Ziwei et al. 2022). Hallucinations can arise because the AI has been trained on vast amounts of data that contains a wide range of information, including both factual and fictional content. The phenomenon by which generative AI produces hallucinations is still not well-understood. Worryingly, generative AI models, even those that are designed to be less prone to hallucinations, have been shown to amplify hallucinations (Dziri et al. 2022).

The problem of hallucination was evident in responses to many of our prompts. For instance, in the critique of "*Toward a Theory of Information Systems Security Behaviors of Organizational Employees...*" above (see Table 4), the underlined text represents hallucinations. Only one (not two) organization was studied, this was not a mixed methods study involving a survey, and SEM was not applied to survey data! As another instance, consistent with past studies (e.g., Eysenbach 2023), ChatGPT provided several fabricated references along with legitimate references in response to the prompt "*Can you list the most important papers looking at the dark side of information technology use on employee well-being and productivity?*" What's more, when asked to summarize a recently published paper in ISR (Wang et al. 2021), ChatGPT not only made up an entirely fictitious paper title and set of authors (Silver et al. 2020), but also an entire set of fabricated claims about how the paper was providing an overview of reinforcement learning (RL) as applied to robotics, and the challenges of applying RL to robotics, such as the need for high-dimensional state and action spaces, the long-horizon nature of many robotic tasks, and the need for robust and reliable algorithms.

Similarly, the use of generative AI in literature reviews can be problematic in that the output may lack scholarly depth consistent with the expectations of a given research community. For example, when prompted to provide a literature review on information systems security behaviors (ISSB) of organizational employees, the resulting output from ChatGPT provided brief summaries of key themes and findings but lacked the theoretical framing and critical analysis expected in an academic literature review. We were curious to know how two leading IS security researchers (both serve on the board for Information Systems Research) would rate the review. One of them assessed the review as follows:

*... It is throwing out concepts or factors studied, without telling [us] what they mean, or what is the state of the art. The taxonomy or categories are overlapping, so the categorization does not make sense. Moreover, IS audience often discusses things based on theories. This is missing...*

*Just listing concepts randomly without their theory, does not tell what these concepts or factors really mean. Citations are missing.*

The other colleague was even more severe:

*I don't see this as a review at all... I see no depth for review purposes. In fact, in looking at the key constructs in this area, it only covers a few of them and glosses over the more interesting ones. No terms are defined, no citations are provided, no meaningful statistics are given, and I have no sense of who is writing this. So I would personally find this useless from a review perspective.... A key problem here is that when we review the literature, we're not just trying to get the high-level obvious stuff, we're trying to find the interesting tensions, metaphors, paradoxes, controversies, unresolved problems, and so forth.*

While the literature review by ChatGPT did provide a high-level summary of the topic, the summary lacked depth, citations, and theoretical grounding. Citations are crucial for readers and authors to judge the credibility and trustworthiness of the ideas presented. The lack of references and explanation of why certain themes were included or excluded made the review less credible and potentially violated intellectual property rights.

It is crucial for scholars to prioritize their understanding of the relevant body of work relevant literature over the suggestions provided by generative AI tools. This approach is imperative because the inherent biases of these AI models, such as ChatGPT, often reflect the corpus of documents on which they are trained. Specifically, these biases can manifest as misclassification of ideas or hallucinations when processing technical or literature-specific language. For example, in the context of the literature on fake news, specific terminology is used to describe false information shared on the Internet (e.g., Moravec et al. 2022). If the generative AI algorithm omits or replaces this language, it could lead a reader to misunderstand the paper's connection to broader research on fake news and online misinformation. Furthermore, in cases where the generative AI algorithm fails to find relevant content, it may generate incorrect citations or link the author to inappropriate sources. In both situations, the author must carefully review the results of all GPT instantiations to identify and correct any problematic results. Thus, while generative AI models hold significant potential in augmenting scholarly work, the authors must exercise caution and prioritize the comprehension of their work and relevant literature to ensure the accuracy and integrity of their research.

## **5. The Path Forward**

The rise of generative AI has significantly lowered the cost of knowledge acquisition, making it easier for individuals and organizations to engage with information and democratize knowledge. With generative AI, it is now possible to discover new insights and learn about complex topics with ease. It is also possible to develop texts that span genres, from memos to poetry, and to disseminate knowledge on a wide range of topics. The optimistic vision holds that such increased access to information can lead to more efficient knowledge creation, assessment, and consumption, and eventually to an informed and empowered society. Yet, as we have seen in

our earlier section, the role of generative AI in the creation, assessment, and summarizing of knowledge is not unproblematic.

The *raison d'être* of top journals such as ISR is to publish original research that contributes to the advancement of knowledge. Such knowledge creation involves a rigorous process that might include theorizing, hypothesis testing, data collection, analysis, and interpretation, and thoughtful evaluation that cannot be replicated by generative AI alone. As a result, the human endeavor associated with papers published by journals such as ISR remains critical to the advancement of knowledge. While generative AI tools do offer noteworthy benefits, such as those outlined in the third section of this editorial, they cannot at this time replace the role of top journals and human engagement in the process of knowledge creation, assessment, and dissemination, which are deeply embedded in the social context and thus involve many tacit rules and conventions. There is also the danger of offloading human cognition and intelligence to generative AI. In fact, we believe that normalizing the tools such as ChatGPT by thinking of them as being similar to “calculators” (that support or automate some straight-forward computing tasks) may be misplaced, given the issues related to bias, hallucinations /fabrications, lack of interpretability and consistency, lack of appropriate depth expected in a community, and potential violations of intellectual property, that we discussed in the fourth section of this editorial. These are not issues posed by the calculator, a **non-stochastic** technology. The technology and the social process of scholarly work will need to evolve significantly before the use of generative AI becomes as routine and unproblematic as the calculators of today.

We also contend that human supervision and oversight may be necessary for almost all stages of research because generative AI lacks the intuition, perspective, and rich understanding of the sociotechnical context related to scholarly work in our discipline. Further, given that generative AI models are rapidly changing, we believe that scholars who use such tools to support their research must, at the minimum, declare the tool(s) used, the version, and how it was used to support their work. For certain research activities, it would be advisable to provide more detail than a simple acknowledgment of the tool used, such as choices of the different prompts made to elicit suggestions. Furthermore, any use of these models should be done with a demonstrated awareness of possible biases, with de-biasing methods carefully adopted and explained. There are also serious questions as illustrated in our section on limitations. The use of generative AI tools could discourage deeper scholarly engagement by lacking theoretical depth, and encourage scholars to seek breadth rather than a depth of understanding. Finally, in light of the problems of bias, fabrication, and poor contextualization that can affect scholarly activities and consequently the quality of papers published in journals, we echo recent comments in the journal *Nature* (van Dis et al. 2023) that urged authors to take responsibility for fact-checking their text, results, data, code, and references, and expressed the need for discipline-specific norms as well as investment in truly open models.

While we are hopeful that this editorial will encourage adaptation of our research practices and refinement of norms, there are clearly many important questions that our community needs to think about, including *what scholarship will look like in the age of generative AI?* Will and can it

remain the same? What tasks might be automated, augmented, or untouched by generative AI? Table 5, adapted from van Dis et al. (2023), provides a useful set of questions, that might be abstracted to the future of work in general, and spawn valuable research studies.

Table 5. Questions for further exploration (adapted from van Dis et al. 2023)

- What are the different ways in which generative AI tools augment or assist human input?
- What steps in an AI-assisted research process require human verification? Relatedly, how do we adopt norms for review processes?
- How should we adopt policies to ensure research integrity?
- How can we build discipline-specific and independent open-source LLMs and other models to ensure IS artefacts and diverse conceptual paradigms are well represented?
- What quality standards should be expected of AI-augmented research (for example, transparency, accuracy, de-biasing, and source crediting), and which stakeholders are responsible for enforcing the standards?
- How can researchers ensure that AI-augmented tools promote equity in research, and avoid risks of widening inequities?
- What legal implications does the practice of AI-assisted research have for intellectual property and discipline-specific knowledge (for example, laws and regulations related to patents, copyright, and ownership)?

Looking to the future use of generative AI models in journals in IS, we note that while these models are pre-trained on a diverse range of internet text sources, they offer limited knowledge specific to *ISR* and other IS journals. Although some papers in the field may have been included in the training data through open-access repositories, preprint servers, or personal websites, the extent to which generative AI models incorporate research articles, particularly those not publicly available and easily accessible, remains uncertain. Consequently, their ability to provide accurate and up-to-date information on specific topics in information systems to support research activities is significantly restricted.

To effectively use the power of generative AI, the IS community may need to develop discipline-specific applications of such tools. ArXiv, which is an open-access repository for academic papers in various disciplines such as physics, mathematics, economics, and computer science, has recently developed a bespoke language model termed ArxivGPT which ‘summarizes arXiv papers and provides key insights’ (<https://twitter.com/ArXivGPT>). *ISR* has a unique opportunity to leverage its wealth of research, review processes, and manuscripts to create a valuable knowledge base. A phased approach could possibly begin with the collection of abstracts and fine-tuning, offering broad-stroke summaries and articulations of contributions. A more ambitious endeavor would involve building a knowledge base derived from paper drafts and reviews that have journeyed through the journal’s review process. Although challenging, this effort could provide immense support for early-career researchers, authors, reviewers, and editors and lead to a far more efficient manuscript development and editorial process.

In closing, the potential for generative AI in revolutionizing *IS* research lies in our ability to harness its strengths, address its limitations, and forge a path forward that illuminates the

profound impact of our work on the world. The future beckons and it is our responsibility to seize it.

## References

- Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots. in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610–623 (ACM, 2021). doi:10.1145/3442188.3445922
- Chomsky, N., I. Roberts, J. Watumull. “Noam Chomsky: The False Promise of ChatGPT,” *New York Times*. Retrieved online at <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- D’Arcy, J., I. Adjerid, C. M. Angst, and A. Glavas. (2020). Too Good to Be True: Firm Social Performance and the Risk of Data Breach. *Information Systems Research* 31(4), 1200–1223. <https://doi.org/10.1287/isre.2020.0939>
- Dziri, Nouha; Milton, Sivan; Yu, Mo; Zaiane, Osmar; Reddy, Siva (July 2022). "On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?". *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.38
- Else, H. (2023, January 19). “Abstracts written by ChatGPT fool scientists,” *Nature*, Vol 613: 423
- Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Med Educ* 2023; 9:e46885
- Freidman, T. (2023, March 23). “Our New Promethean Moment,” *New York Times* Retrieved online at <https://www.nytimes.com/2023/03/21/opinion/artificial-intelligence-chatgpt.html>
- Fugener A., J. Grahl, A Gupta, W Ketter. (2021) Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33(2):678-696.
- Gao, C.A, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, A. T. Pearson. (2022), “Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers,” *bioRxiv* 2022.12.23.521610; doi: <https://doi.org/10.1101/2022.12.23.521610>
- Karjalainen, M., S. Sarker, M. Siponen (2019) Toward a Theory of Information Systems Security Behaviors of Organizational Employees: A Dialectical Process Perspective. *Information Systems Research* 30(2):687-704
- Kwon H.E., S. Dewan. W. Oh, T. Kim (2023) Self-Regulation and External Influence: The Relative Efficacy of Mobile Apps and Offline Channels for Personal Weight Management. *Information Systems Research* 34(1):50-66.
- Larsen, B., and J. Narayan. (2023, January 9). “Generative AI: a game-changer that society and industry need to be ready for.” *World Economic Forum*. Retrieved online at <https://www.weforum.org/agenda/2023/01/davos23-generative-ai-a-game-changer-industries-and-society-code-developers/>

- Madani, A., Krause, B., Greene, E.R. *et al.* (2023). "Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*.  
<https://doi.org/10.1038/s41587-022-01618-2>
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. (2013). *Distributed representations of words and phrases and their compositionality*. Advances in Neural Information Processing Systems. arXiv:1310.4546
- Mitchell M, Krakauer DC. The debate over understanding in AI's large language models. Proc Natl Acad Sci U S A. 2023 Mar 28;120(13):e2215907120. doi: 10.1073/pnas.2215907120
- Mollick. E. (2023). @emollick Retrieved from Twitter:  
<https://twitter.com/emollick/status/1653945049275670528>
- [Patricia L. Moravec](#), [Antino Kim](#), [Alan R. Dennis](#), [Randall K. Minas](#) (2022) Do You Really Know if It's True? How Asking Users to Rate Stories Affects Belief in Fake News on Social Media. *Information Systems Research* 33(3):887-907
- Russell, S. J. and P. Norvig. (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2.
- Sahaym A., J. Vithayathil, S. Sarker. S. Sarker. N Bjorn-Andersen. (2022) Value Destruction in Information Technology Ecosystems: A Mixed-Method Investigation with Interpretive Case Study and Analytical Modeling. *Information Systems Research*. Forthcoming
- Susarla, A., M. Holzhaecker., & R. Krishnan. (2020). Calculative Trust and Interfirm Contracts. *Management Science* 66(11):5465-5484.
- Tarafdar M., G. Shan, J B Thatcher, A Gupta (2022) Intellectual Diversity in IS Research: Discipline-Based Conceptualization and an Illustration from *Information Systems Research*. *Information Systems Research* 33(4):1490-1510.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N Gomez, L. Kaiser, I. Polosukhin (2017). "Attention Is All You Need". arXiv:1706.03762
- Van Dis, E., J. Bolen, R van Rooij, W. Duidema, C.L. Bockting. (2023) ChatGPT: five priorities for research. *Nature*, 619, 224-226
- Wang, Y-Y., C. Guo, A. Susarla, V Sambamurthy. (2021) Online to Offline: The Impact of Social Media on Offline Sales in the Automobile Industry. *Information Systems Research* 32(2):582-604.
- Wei, J., T. Yi; R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, C. Donald, E Chi, T. Hashimoto, O. Vinyals, P Liang, J. Dean, W Fedus, (31 August 2022). "Emergent Abilities of Large Language Models". *Transactions on Machine Learning Research*. ISSN 2835-8856
- Wong, R. Y. M., Cheung, C. M., Xiao, B., & Thatcher, J. B. (2021). Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research*, 32(2), 561-581.
- Yang, Z., & Feng, J. (2020). A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 9434-9441. <https://doi.org/10.1609/aaai.v34i05.6486>
- Zheng, J., Y. Wang, Y. Tang. (2022) Platform Refund Insurance or Being Cast Out: Quantifying the Signaling Effect of Refund Options in the Online Service Marketplace. *Information Systems Research*, Forthcoming



Ji, Ziwei; Lee, Nayeon; Frieske, Rita; Yu, Tiezheng; Su, Dan; Xu, Yan; Ishii, Etsuko; Bang, Yejin; Dai, Wenliang; Madotto, Andrea; Fung, Pascale (November 2022). "Survey of Hallucination in Natural Language Generation" (pdf). *ACM Computing Surveys*. Association for Computing Machinery. **55** (12): 1–38.