Journal of the Association for Information Systems

Volume 25
Issue 1 *Special Issue: The Future Impact of AI on Academic Journals and the Editorial Process (pp. 37-181)*

Article 11

2024

Responsible Artificial Intelligence and Journal Publishing

Shirley Gregor
Australian National University, shirley.gregor@anu.edu.au

Recommended Citation
Gregor, Shirley (2024) "Responsible Artificial Intelligence and Journal Publishing," *Journal of the Association for Information Systems*, 25(1), 48-60.
DOI: 10.17705/1jais.00863
Available at: https://aisel.aisnet.org/jais/vol25/iss1/11

Journal of the Association for Information Systems (2024) **25**(1), 48-60
doi: 10.17705/1jais.00863

EDITORIAL

ISSN 1536-9323

Responsible Artificial Intelligence and Journal Publishing

Shirley Gregor[1]

[1]College of Business and Economics, Australian National University, Australia, shirley.gregor@anu.edu.au

Abstract

The aim of this opinion piece is to examine the responsible use of artificial intelligence (AI) in relation to academic journal publishing. The work discusses approaches to AI with particular attention to recent developments with generative AI. Consensus is noted around eight normative themes for principles for responsible AI and their associated risks. A framework from Shneiderman (2022) for human-centered AI is employed to consider journal publishing practices that can address the principles of responsible AI at different levels. The resultant AI principled governance matrix (AI-PGM) for journal publishing shows how countermeasures for risks can be employed at the levels of the author-researcher team, the organization, the industry, and by government regulation. The AI-PGM allows a structured approach to responsible AI and may be modified as developments with AI unfold. It shows how the whole publishing ecosystem should be considered when looking at the responsible use of AI—not just journal policy itself.

Keywords: Artificial Intelligence, Journal Publishing, Responsible AI, Human-Centered AI, Generative AI, AI Governance, Design Science Research

David Schwartz was the accepting senior editor. This paper was submitted on June 20, 2023 and underwent two revisions. It is part of the Special Issue on The Future Impact of AI on Academic Journals and the Editorial Process.

1 Introduction

Artificial intelligence (AI) is now receiving an unprecedented level of public attention. ChatGPT, a form of generative AI, was released in November 2022 for use by the general public in many contexts. Its level of adoption was astonishing, with more than 100 million monthly active users by the end of January 2023, just two months after its launch (Wu et al., 2023). From May to October 2023, when this opinion piece was prepared, multiple articles were appearing daily in the press on the use of generative AI. Topics of concern related to generative AI included the provision of false information (hallucinations), use in large-scale disinformation campaigns, bias, misuse by students, loss of jobs, risks regarding AI's ability to write code, and lack of respect for the rights of holders of the intellectual property used in training systems. On the other hand, some businesses were enthusiastic about its use. Nvidia, the dominant company producing chips used in generative AI, saw its share prices boom. Steps were being taken toward developing regulations and "guardrails" for the use of generative AI.

The aim of this opinion piece is to consider the *responsible use* of AI in relation to academic journal publishing. The understanding of what is meant by responsible AI varies, but recurring themes include ethical considerations, transparency, and a focus on human well-being (e.g., see Dignum, 2018). A similar perspective is being taken in *human-centered AI*, which is seen as an emerging discipline that aims to create AI systems that amplify and augment rather than displace human abilities (Geyer et al., 2022).

The issue of responsible AI is an important one for academic journals. For example, the *Journal of the Association for Information Systems* (JAIS) expects to publish work of "the highest quality scholarship in the field of information systems" and "rigorously developed contributions"[1] (i.e., not hallucinations). JAIS is also inclusive in terms of research approaches, meaning that it accepts work using the design science approach where the authors themselves may have constructed some form of AI (e.g., see Ptaszynski et al., 2019). Other concerns for academic journals include plagiarism, authorial responsibility, and appropriate source attribution. Some work on these concerns has already appeared (e.g., Dwivedi et al., 2023; Eke 2023).

Artificial intelligence, especially natural language processing and commonsense knowledge representation were among the research interests of Professor Phillip Ein-Dor, who is honored by this special issue. As one example of his contributions in this area, he was responsible as editor for the proceedings of the Fourth International Workshop on Artificial Intelligence in Economics and Management in Tel-Aviv, Israel in 1996 (Ein-Dor, 1996).

This opinion piece is being written at a time of very rapid change and it is difficult to give comprehensive coverage of the phenomena of interest. Some of what is written is likely to be outdated in the near future. The aim is, however, to show some of the principles of responsible, human-centered AI that currently exist and how they can be applied in the context of publishing in academic journals. These principles may have some longevity and can serve as a base for further discussion.

This paper first discusses different views of both artificial intelligence and responsible AI. I present well-recognized principles of responsible AI and demonstrate their applicability to publishing academic journals in terms of a matrix for governance structures for human-centered AI, based in part on Shneiderman (2020, 2022). My concluding remarks follow.

2 Understanding Artificial Intelligence

Understandings of AI vary. Here, I adopt an inclusive approach and regard AI as the systems that are enabled by the machine capabilities described in Russell and Norvig (2016): problem solving, knowledge reasoning and planning, uncertain knowledge and reasoning, learning, communicating (including natural language), perceiving (including computer vision), and acting (robotics). It is important to appreciate the different approaches to AI, as they have their own strengths and limitations (e.g., see Solomon & Davis, 2023). Discussion with academic colleagues, including those from fields outside computing and information systems, has indicated that there may be a lack of appreciation of the nature and issues associated with the different forms of AI.

With the current prominence of generative AI such as ChatGPT (Chat Generative Pre-Trained Transformer), there may be a disregard for forms of AI that were studied earlier in the history of AI—systems that could be referred to as expert systems, knowledge-based systems, decision support systems, logic-based models, or just intelligent systems (e.g., see Gregor & Benbasat, 1999). These systems have not gone away and are commonly in use in different forms in everyday life. Different approaches to AI can complement each other. For example, Wiechetek et al. (2021) investigated both rule-based and machine learning methods for grammar checking on a Sami language, which has a small number of native speakers.

There are important differences between generative AI and logic-based systems, including shortcomings in the ability of machine learning systems to provide explanations and transparency in terms of how output is produced (e.g., see Arrieta et al., 2020; Samek et al., 2017). Further, machine learning systems are reliant on the adequacy of their training data and may exhibit bias and inaccuracies that are hard to detect. On the other hand, machine learning systems may be able to respond quickly to a variety of requests, compared to rule-based systems that need to be reprogrammed by human programmers. Generative AI enables the production of new and creative content such as text, images, music, and videos following prompts from users. The new content is built from existing data using machine learning, often "scraping" the material for the new data from immense datasets. ChatGPT can reportedly access over three hundred billion words, covering all kinds of content on the internet, including news reporting, personal data, policy documents, literary texts, and art (Helberger & Diakopoulos, 2023).

ChatGPT is an example of a "large language model" that uses advanced machine learning. These models "are trained to generate new data, such as text, images, or audio." This "makes them distinct from other AI models … [only] designed to make predictions or classifications" (Hacker et al., 2023, p. 4). ChatGPT

[1] https://aisel.aisnet.org/jais/about.html

was developed by OpenAI and released to the public in November 2022. It integrates multiple technologies, such as deep learning, unsupervised learning, instruction fine-tuning, multitask learning, in-context learning, and reinforcement learning (Wu et al., 2023). ChatGPT is reported to exhibit human-level performance on various professional and academic benchmarks (OpenAI, 2023). Wu et al. (2023) provide comprehensive coverage of the history of ChatGPT, its abilities and potential areas of concern. Other forms of generative AI produce images from text, such as DALL-E-2 of OpenAI [2] and Craiyon, formerly DALL·E mini.[3]

My reaction to using ChatGPT for the first time was that it had astounding capabilities in terms of the amount of information it could gather and structure quickly, its sophistication in the use of the English language, and its ability to remember what I had said earlier in the conversation. It even appeared to show traces of a sense of humor. However, like many others, I soon started to test its capabilities to see where breakdowns might occur.[4] Gary Marcus and others have compiled a collection of errors made by ChatGPT under the heading "The Road to AI We Can Trust."[5] There is vigorous debate as to whether generative AI can lead to what is termed "artificial general intelligence" or "strong AI" (see Chomsky et al. 2023).

Generative AI, based on machine learning, uses inductive probabilistic methods and may be weaker in areas such as logical reasoning, which are the strength of other approaches. It may be that some combination of different approaches will yield improvements in the future, but for now, weaknesses should be recognized. These issues are especially important with AI services such as ChatGPT that are available for public use, where users, even academic users, may have little knowledge of the underlying technologies and their risks.

3 Principles for Responsible AI

Given the characteristics of different forms of AI discussed above, it is important to consider how potential ill-effects can be mitigated, leading to the need for responsible AI.

> *Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and wellbeing in a sustainable world* (Dignum, 2018, p. 1).

The term responsible AI can be used somewhat interchangeably with the terms human-centered AI, ethical AI, and trustworthy AI. Fjeld et al. (2020) carried out a survey of relevant normative principles in 36 documents from the government (e.g., the European Commission), intergovernmental organizations (e.g., the OECD), the private sector (e.g., IBM, Google, Microsoft), civil society organizations (e.g., Amnesty International), and multi-stakeholder organizations (e.g., Beijing Academy of AI). Their comparison of the principles across these documents uncovered a growing consensus around eight key thematic trends, shown in Table 1. Table 1 shows the themes in an order that differs from that in Fjeld et al. (2020). The order has been changed to better show how the themes are linked, corresponding to some extent to the four intersecting perspectives on ethics in information systems and design science research presented in Herwix et al. (2022). The order of the themes proceeds as follows: first, what is "designed in," i.e., linked to designers and the ethical perspectives of design and design-in-practice; second, principles (e.g. accountability) that are more linked to encompassing ethical perspectives of science as a whole; and third, overarching philosophical perspectives, such as deontological normative ethics that stress adherence to moral rules to promote human values.

Each of these themes has associated principles directed at preventing or alleviating harms or risks, and the following discussion points to some of the risks that could occur in the context of publishing in academic journals. Some of the themes are intertwined, in that a particular risk might be addressed under more than one theme. I have exercised personal judgment in positioning examples of risks relating to journal publishing under the theme they seem most pertinent to—others might position them differently. The discussion takes into account both the development and use of AI by research teams in industry and academia and their involvement in the publication process.

[2] https://openai.com/dall-e-2
[3] https://www.craiyon.com/
[4] I continue to find the unreliability of ChatGPT alarming even in casual use. For example, in a request to identify the narrator in the movie "Mars Attacks!" I was given three names in succession, each one obviously not correct. It was only after I commented that the last person identified had been dead for 10 years when the movie was made that ChatGPT admitted that it did not have information on who the narrator was (ChatGPT based on GPT-3.5 architecture. URL: chat.openai.com, accessed June 9, 2023).
[5] https://garymarcus.substack.com/

Table 1. Eight Themes in Principles for Responsible AI (from Fjeld et al., 2020)

1. **Professional responsibility:** Individuals involved in the development and deployment of AI systems play a vital role in the systems' impacts, and they should act with professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for.
2. **Safety and security:** AI systems should be safe, perform as intended (reliable), and also secure, i.e., resistant to being compromised by unauthorized parties.
3. **Fairness and nondiscrimination:** AI systems should be designed and used to maximize fairness and promote inclusivity, with concerns about AI bias already impacting individuals globally.
4. **Privacy:** AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it.
5. **Transparency and explainability:** AI systems should be designed and implemented to allow for oversight, including through the translation of their operations into intelligible outputs.
6. **Human control of technology:** Important decisions should remain subject to human review.
7. **Accountability:** It is important to have mechanisms that ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided.
8. **Promotion of human values:** The ends to which AI is devoted, and the means by which it is implemented, should correspond with humankind's core values and generally promote humanity's well-being

3.1 Professional Responsibility

This theme relates to the individuals and teams who are responsible for the design, development, and deployment of AI systems. Principles include accuracy, responsible design, consideration of long-term effects, multi-stakeholder collaboration, and scientific integrity.

These principles are of direct relevance to researchers who work in the design science research paradigm and develop AI systems that are reported on in publications. Researchers should know "how to do things right," as well as "doing the right thing" (see Herwix et al., 2022).

When authors use AI tools developed by others, they have the professional responsibility to preserve scientific integrity and use the tools only as far as their limitations permit (see Theme 6).

3.2 Safety and Security (Reliability)

The principles here require that an AI system be reliable and do what it is supposed to do before and after deployment, without harming living beings or the environment.

The nature of generative AI means that it employs probabilistic methods and cannot be expected to always give correct answers. OpenAI says in its description of ChatGPT that it "may be inaccurate, untruthful, and otherwise misleading at times" and that "ChatGPT is not connected to the internet, and it can occasionally produce incorrect answers. It has limited knowledge of world and events after 2021 and may also occasionally produce harmful instructions or biased content." [6] These disclaimers are not very helpful in indicating whether some tasks (e.g., calculations, logic) are more likely to result in errors than others and some may dispute what "occasionally" means.

Some authors have noted problems. For example, van Dis et al. (2023, p. 224) report:

> *Next, we asked ChatGPT to summarize a systematic review that two of us authored in JAMA Psychiatry5 on the effectiveness of cognitive behavioural therapy (CBT) for anxiety-related disorders. ChatGPT fabricated a convincing response that contained several factual errors, misrepresentations and wrong data (see Supplementary information, … For example, it said the review was based on 46 studies (it was actually based on 69) and, more worryingly, it exaggerated the effectiveness of CBT.*

Kim (2023, p. 1) reports asking ChatGPT:

> *"Is there any reference for this topic?" Then, it replied that "Yes, there are many references on the effects of streptozotocin-induced diabetes on bone growth patterns in rats. Here are a few examples: … These studies and others suggest that streptozotocin-induced diabetes can have a negative impact on bone growth and development in rats, including the facial bones." I searched whether these*

[6] https://help.openai.com/en/articles/6783457-what-is-chatgpt (accessed June 9, 2023).

> *references are real or fake. Unfortunately, all references are fake including the fake authors. However, its other performance such as editing English grammar was wonderful.*

The degree of reliability can be balanced against the potential severity of the risk. For example, there may be less risk when the AI is used in comparatively low-level tasks, such as grammar checking, and when the outputs are subject to oversight by human users (Theme 6).

3.3 Fairness and Nondiscrimination

This theme includes the principle of addressing risks such as algorithmic bias that can arise when machine learning systems are developed with training sets that include unrepresentative, flawed, or biased data.

An example is that of researchers who used IBM Watson to develop a system to improve cancer care. The system has been criticized for relatively poor performance and one problem noted is that the system was biased towards the conditions specific to the hospital where the system was first developed (Strickland, 2019). Ferrara (2023) discusses the biases in large language models such as ChatGPT and notes that these models inevitably absorb the biases present in the data on which they are trained.

3.4 Privacy

The right to privacy is integral in human rights law. This theme includes principles for consent, control over the use of data, the ability to restrict data processing, the right to rectification, the right to erasure, privacy by design, and recommendations for data protection laws.

A risk with research and journal publishing is that an AI tool could generate responses from many sources and might include data that was gathered without human consent. The data could include individuals' blog posts, product reviews, or comments on an online article. ChatGPT does not ask individuals whether OpenAI can use their data—a clear violation of privacy. The data could be sensitive and used to identify individuals, their family members, or location. Gal (2023) points out that even when data is publicly available, its use could breach what is termed "textual integrity," a fundamental principle of legal considerations of privacy. That is, individuals' information should not be revealed outside of the context in which it was originally produced. Further, OpenAI offers no means for individuals to check whether the company is storing their personal information or to request it be deleted, contrary to legislation such as the European Union's General Data Protection Regulation (GDPR).

3.5 Transparency and Explainability

Principles in this theme include those of transparency and explainability, open source data and algorithms, and notification when interacting with an AI. Transparency and explainability can be a major challenge because of the complexity and opacity of some AI technologies. Arrieta et al. (2020) provide an overview of issues with explainability in machine learning systems and ongoing efforts to improve transparency that AI developers can heed.

ChatGPT usually gives no explanation of how material was generated, attribution for data sources, or any estimate of the degree of accuracy of the particular information supplied. In the excerpt above, Kim (2023) notes that even when ChatGPT was asked for references on a topic, it provided fake answers.

3.6 Human Control of Technology

Principles in this theme include the human review of technology and the ability to opt out of automated decision-making. Given the potential problems noted with generative AI under other themes, including lack of accuracy, bias, and low transparency, it is essential that humans are in control of these technologies, especially in high-risk areas. These ideas are stressed and explained further in Shneiderman (2022), where a key notion is that AI should be "human-centered"—AI is a tool and the human is in charge.

This theme is important for publishing in journals, where high standards of scientific rigor are expected. Authors should be capable of assessing and checking the accuracy of outputs themselves; otherwise, tools should not be relied upon.

The option does seem available for an author to use generative AI as a source of new ideas or in tasks such as reviewing one's own work (see Crawford, 2023; Gunn, 2023) but the latter should be done with care, taking into account the possible loss of intellectual property (see Theme 8).

3.7 Accountability

The accountability theme includes principles for conducting impact assessments for the use of AI and its liabilities and how entities or individuals can be held accountable for AI use outcomes.

A particular concern here is whether an AI can be seen as taking responsibility as an author of published research. Alarms were raised when two articles published in the science and health fields included ChatGPT as a bylined author (Stokel-Walker, 2023). At least some journals hold that as a nonhuman entity—an AI tool cannot take responsibility for the integrity of research work and thus cannot be an author (e.g., see Flanagin et al., 2023).

3.8 Promotion of Human Values

Three principles are included here: human values and human flourishing, access to technology, and leverage to benefit society. Fjeld et al. (2020) note that the principles under this theme were coded directly from explicit references in the source documents to human rights and international instruments of human rights. Concepts mentioned include human dignity and autonomy, "the progress of human civilization" and respect for "justice and the rule of the law" (p. 61).

Some aspects of human well-being have been dealt with separately under other themes—for example, privacy, safety and security, and nondiscrimination. A threat that has not been dealt with in a separate theme and is particularly important in the context of journal publishing is that of "theft," the stealing of intellectual property belonging to others. The protection of intellectual property and copyright issues are matters that are subject to law in many countries. A problem with generative AI that is trained on large public datasets is that the ownership of the intellectual property that is represented in individual items of data is generally not acknowledged.

An example with image generation illustrates how problems with the attribution of sources can arise with generative AI, using the example of the generation of an image to insert in a publication. Figure 1 shows: (a) a text description of a male Superb Fairy Wren in breeding plumage (eBird 2023), (b) an illustrator's image of the wren produced using a reference photo,[7] and (c) an image generated by Craiyon in response to the prompt "Australian Superb Fairy Wren Adult Male Breeding."[8] The image in Figure 1c was one of several produced using Craiyon after variations in the wording of the prompt. Even if an accurate image of the bird is produced by Craiyon and used, there is no way of knowing how close it is to an already published and copyrighted image in order to give correct attribution and seek permission for its use. Further, oversight of the output of the AI is needed to prevent inaccuracy (Theme 6). The human has to rely on their own expertise or cross-referencing of sources to check the accuracy of the drawing. Craiyon's image does not have the black coloration around the eye and through to the back of the head as described in Figure 1a and Craiyon's bird looks suspiciously as if it has three legs.

Loss of intellectual property could also go the other way. Authors should be cautious in reviewing their own work because prompts fed into ChatGPT can be added to its data center, as Samsung found when their employees unintentionally leaked trade secrets (Dobberstein, 2023). This problem is probably not as well-appreciated as it should be. Kim (2023, p. 2) says "If ChatGPT is only used for language editing purposes, then there is *no issue* with using it to prepare scientific articles" (emphasis added), which is a questionable statement.

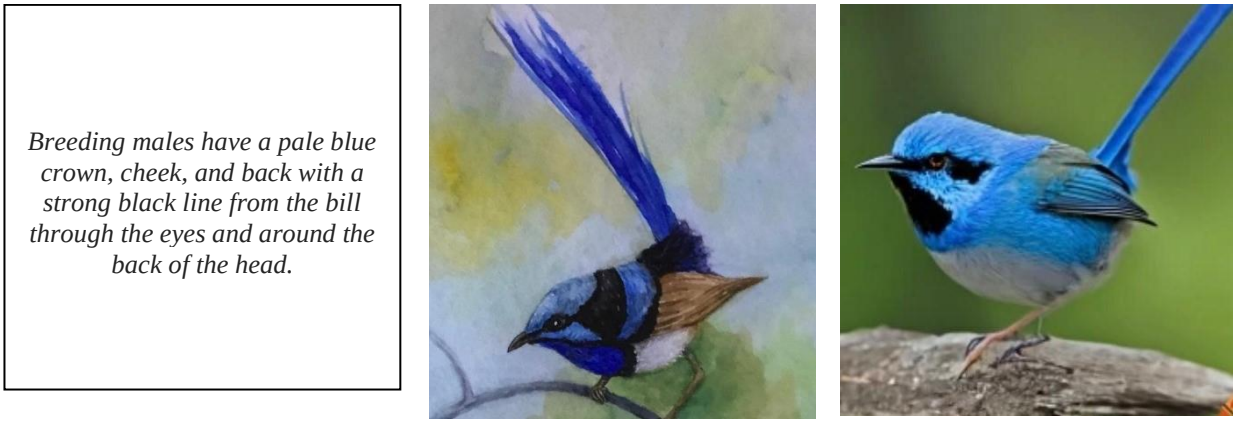

Figure 1. Issues with Accuracy and Intellectual Property with Image Generation: (a) Text Description of the Superb Fairy Wren, (b) Illustrator's Image, (c) Craiyon's Image

[7] Image produced by the author with Simpson and Day (2004, p. 173) as a reference.

[8] Image produced by the author using Crayion, 2023 Craiyon LLC, https://www.craiyon.com/ (accessed June 6, 2023).

4 Principled Use of AI in Academic Journal Publishing

The existence of normative principles, as in "what ought to be done" to address risks**,** does not in itself show how they can be translated into practice, as in "how to do it." To examine how the principles could be implemented in the context of journal publishing, we turn to the framework of governance structures for human-centered AI provided in Shneiderman (2022) and adapt it to the context of academic journal publishing. Shneiderman depicts four nested levels of governance structures, from least to most inclusive:

- **Team:** practices of software engineering teams that enable reliable human-centered AI systems, including audit trails, workflows, verification and validation testing, bias testing, and explainable user interfaces
- **Organization:** a safety culture and management of AI projects with strategies including leadership commitment, appropriate hiring and training, reporting failures and near misses, internal reviews, and following industry standards
- **Industry:** trustworthy certification by external reviews with independent oversight such as auditing firms, insurance companies, NGOs, and civil society
- **Government regulation:** such as the GDPR in Europe

Here Shneiderman's governance levels are adapted for use in the context of journal publishing, with examples of how the principles for responsible AI could be addressed. Again, the selection of countermeasures is indicative rather than complete, given the rapid developments in the field of AI, and personal judgment has been exercised in placing countermeasures at particular governance levels and against particular principles.

4.1 The Researcher-AI Team Level

Researchers (authors) can engage with AI in conducting and publishing research in more than one role. They could be a "developer" of the AI, if using the design science approach, or they could be a "professional user" of an external AI in other approaches, using the terminology of Hacker et al. (2023).

At this level, there is largely a reliance on the personal ethics of the researchers and their knowledge of the nature of risks and how they can be countered. Ingrained personal ethical views may be difficult to change, but exposure to what are regarded as best practices and increased knowledge of issues may be of benefit, as occurs in the teaching of ethics in university IT courses. Thus, a focus on knowledge to encourage professional responsibility (Theme 1) seems appropriate and should, in turn, lead to addressing other themes.

A number of resources are now available to assist. For example, when the researchers are developers, they could be expected to be aware of the practices of software engineering teams that enable reliable human-centered AI systems, as described in Shneiderman (2022). Herwix et al. (2022) discussed varying perspectives on ethics in design science research, including the coverage of underlying perspectives on ethics. Forums devoted to responsible AI could also be consulted to keep abreast of issues: for example, the Human-Centered AI website curated by Ben Shneiderman and Mengnan Du has a comprehensive list of resources.[9] Some codes of ethics from professional bodies directly address risks with AI: for example, The Global Initiative on Ethics of Autonomous and Intelligent Systems by IEEE.[10]

When the researchers are AI users, risks may be related to a lack of knowledge regarding the nature of the AI being used, particularly with generative AI. Again, education and knowledge enhancement could assist.

4.2 The Organizational Level

In this context, organizations are taken to be primarily universities or other research institutions employing researchers. These entities may have their own academic code of ethics for researchers. Although the use of AI by students and issues including plagiarism are topics that are now receiving considerable attention, it is not clear to what extent the use and potential misuse of AI by researchers has been recognized and whether codes for academic integrity have been updated (Eke, 2023).

Universities should make their academic staff aware of issues such as the potential loss of valuable intellectual property if they submit their own work to tools like ChatGPT. As mentioned above, Samsung reportedly found that their employees unintentionally leaked trade secrets when they uploaded source code into ChatGPT (Dobberstein, 2023). Dobberstein gives some details of how OpenAI has attempted to deal with the problem by allowing users to opt in or out of data sharing.

[9] https://hcai.site

[10] https://standards.ieee.org/industry-connections/ec/autonomous-systems/

4.3 The Industry of Academic Publishing

The industry level is taken to include the networks of organizations and associated individuals with an interest in academic publications, including journal publishers and editors. Issues currently of concern to journals include plagiarism, the generation of fake research by "paper mills," and a lack of attribution for others' intellectual property (e.g. see Liebrenz et al., 2023).

The Committee on Publication Ethics (COPE) is an organization at the industry level that "brings together all those involved in scholarly research and its publication to strengthen the network of support, education and debate in publication ethics." It includes "cases with advice, guidance for day-to-day practice, education modules and events on topical issues."[11] For example, one discussion looks at the systematic manipulation of the publishing process via "paper mills"—organizations that produce and sell fraudulent manuscripts, possibly aided by AI tools, that imitate genuine research. COPE promotes the use of tools that can in turn detect these papers, even across journals in order to detect duplicate submissions.[12] However, Homolak (2023) experimented with the use of the AI detection tools ZeroGPT, the OpenAI classifier, and GPTZero, and concluded current methods for accurately detecting AI-generated scientific abstracts are inadequate.

Some journals have initiated policies regarding nonhuman authors and the use of generative AI. *Nature* has adopted a policy for the use of large-scale language models in scientific publications, prohibiting naming such models as credited authors on research papers (Nature, 2023a). *Nature* has also announced that it will not accept the use of generative AI in images and video (Nature, 2023b), for reasons similar to those explained in connection with drawings of the wren in Figure 1.

Flanagin et al. (2023, pp. 637-38) report how JAMA and the JAMA Network journals have updated relevant policies in the journals' "Instructions for Authors" stating:

> ***Author Responsibilities*** *Nonhuman artificial intelligence, language models, machine learning, or similar technologies do not qualify for authorship. If these models or tools are used to create content or assist with writing or manuscript preparation, authors must take responsibility for the integrity of the content generated by these tools. Authors should report the use of artificial intelligence, language models, machine learning, or similar technologies to create content or assist with writing or editing of manuscripts in the Acknowledgment section or the Methods section if this is part of formal research design or methods.*
>
> *This should include a description of the content that was created or edited and the name of the language model or tool, version and extension numbers, and manufacturer. (Note: this does not include basic tools for checking grammar, spelling, references, etc.).*

Further guidelines are given for reproduced and recreated material, image integrity, and relevant policies for reporting the use of statistical analysis software.

Lodge et al. (2023) note similar policies for the *Australasian Journal of Educational Technology* with the addition of a rule for reviewers:

> *AJET Reviewers do not have permission to use generative AI to complete any reviews of AJET articles. Sharing articles under review with third party AI providers for this purpose may contravene authors' intellectual property rights to their work.* (p. 6)

4.4 Government Regulation

The situation at this level currently is one of flux, with changes occurring rapidly. Some countries have already enacted data protection laws, such as the European Union's GDPR, which address issues of privacy, fairness, transparency, and accountability in the computerized storage and accessing of data in general, not for AI tools alone (Goddard, 2017).

Infringement of intellectual property has long been subject to legal recourse and a discussion of generative AI and copyright law can be found in Zirpoli (2023). Zirpoli describes how plaintiffs in the US have filed multiple lawsuits alleging copyright infringement via AI training processes.

Regulatory frameworks for generative AI are under development in a number of countries, with debate and public discourse about the balance between softer approaches such as voluntary principles and standards, and harder policy options through legislation and mandatory requirements. A report by the eSafety Commissioner (2023) gives a summary showing graduated approaches ranging from "voluntary principles and governance frameworks" in India to the "application of existing consumer safety and data regulations and the signing of pledges around self-regulatory principles" in the US and "dedicated AI legislation" in the EU, Canada, South Korea, and Brazil to "intermediate bans on generative AI technology" in Italy. Difficulties in enforcing legislation are noted.

[11] https://publicationethics.org/

[12] https://publicationethics.org/resources/research/paper-mills-research

Table 2. The AI Principled Governance Matrix (AI-PGM) for Journal Publishing

Principle Themes/risks	Governance levels and risk countermeasure examples			
	Researcher/AI team	**Organizational level**	**Industry (journal) level**	**Government regulation**
1. Professional responsibility	• Education and training, professional codes of ethics, knowledge resources.	• Organizational codes of ethics and research conduct.	• Codes for publication ethics (e.g. Committee for Publication Ethics*), journal publication guidelines.	
2. Safety and security				
3. Fairness and nondiscrimination				
4. Privacy				• Data protection laws
5. Transparency and explainability				
6. Human control of technology				
7. Accountability			• Prohibiting non-human authors. Use of tools to detect AI-generated work.	
8. Promotion of human values (including intellectual property rights)		Prohibiting the use of AI for reviewing research work.	• Journal policy on how authors acknowledge the use of AI tools. • Prohibiting the use of images from generative AI. • Prohibiting the use of AI for performing reviews.	• Emerging regulatory frameworks. • Copyright law. • Public discourse.

Note: Arrows indicate that a countermeasure extends to cover other principles.
* See https://publicationethics.org/ and https://publicationethics.org/resources/research/paper-mills-research

5 Concluding Remarks

The aim of this opinion piece is to consider the responsible use of AI in relation to academic journal publishing. It is a difficult subject to address at present, as new forms of generative AI have recently been released and enthusiastically adopted by the public at large as well as the private sector, meaning a very rapidly changing landscape. A wide range of literature is relevant, including many current reports in the gray literature, such as government and consultants' reports, news articles, and preprints. It is inevitable that coverage is selective, and some material will soon be out of date.

However, reflection on what has been presented suggests that some conclusions can be drawn:

- It is important to understand and be clear about what is meant by AI, its different forms, and their characteristics. Some forms are likely to be relatively unproblematic in publishing, such as grammar checkers, while others, such as forms of generative AI, have significant issues associated with their use.
- Examination of the existing literature has shown a high degree of convergence on eight themes for normative principles for responsible AI (see Table 1, taken from Fjeld et al., 2020).
- The framework for governance of human-centered AI presented by Shneiderman (2022) provides a useful structure for considering how to achieve the normative principles of responsible AI in practice. This framework considers practices at the levels of the researcher-AI team, the organization, the industry, and government regulation.

- At the level of the researcher-AI team, there needs to be an emphasis on education and knowledge enhancement and the provision of relevant resources.
- At the level of the organization, it appears that more work needs to be done by research institutions to ensure that their codes of academic conduct are being updated.
- At the industry level, the quick response by some journals (e.g., *Nature*) in providing instructions and policies to follow with respect to generative AI should be of use to authors, reviewers, and researchers generally. The Committee on Publication Ethics (https://publicationethics.org/) provides useful and regularly updated material.
- The level of government regulation is experiencing very rapid change, and it is difficult to see what will eventually come to pass. Existing legislation for copyright and data protection can assist.

The new AI-PGM provides a structured means for examining governance practices in terms of the principles and associated risks in the development and use of AI. The matrix shows how the whole ecosystem of publishing should be considered when looking at the responsible use of AI—not just journal policy itself, but also coordinating knowledge and action across authors, research organizations, and government legislation.

The AI-PGM has been used here in the context of journal publishing. However, it has the potential to be applied in other fields. As the matrix is new, there are opportunities for others to provide commentary and develop it further.

Acknowledgments

Thank you to colleagues David Jones, Aleck Lin, Muralidharan Ramakrishnan, and Ruonan Sun, who provided helpful comments on a draft form of the manuscript, and to editor David Schwartz and the anonymous reviewer who helped in improving the manuscript.

References

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Chomsky, N., Roberts, I. & Watumull, J. (2023, March 10). AI Unravelled: The false promise of ChatGPT. *The New York Times*.

Crawford, J., Cowling, M., Ashton-Hay, S., Kelder, J., Middleton, R., & Wilson, G. S. (2023). Artificial Intelligence and authorship editor policy: ChatGPT, Bard Bing AI, and beyond. *Journal of University Teaching & Learning Practice*, *20*(5). https://doi.org/10.53761/1.20.5.01

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, *20*, 1–3. https://doi.org/10.1007/s10676-018-9450-z

Dobberstein (2023). *Samsung reportedly leaked its own secrets through ChatGPT*. The Register. https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own/

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642.

eBird (2023). Superb Fairy Wren. https://ebird.org/species/supfai1

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, Article 100060.

Ein-Dor, P. (Ed.). (1996). *Artificial intelligence in economics and management: An edited proceedings on the Fourth International Workshop*. Springer Science & Business Media.

eSafety Commissioner (2023). *Tech Trends Position Statement Generative AI*. https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf

Ferrara, E. (2023). *Should ChatGPT be biased? Challenges and risks of bias in large language models.* ArXiv. https://arxiv.org/abs/2304.03738

Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2023). Nonhuman "Authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA*, *329*(8), 637-639.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (Berkman Klein Center Research Publication No. 2020-1). SSRN. https://ssrn.com/abstract=3518482

Gal, U. (2023). *ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned.* The Conversation. https://theconversation.com/chatgpt-is-a-data-privacy-nightmare-if-youve-ever-posted-online-you-ought-to-be-concerned-199283

Geyer, W., Weisz, J., Pinhanez, C. & Daly, E. (2022, March 31). What is human-centred AI? *IBM Research Blog*. https://research.ibm.com/blog/what-is-human-centered-ai

Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, *59*(6), 703-705.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, *23*(4), 497-530.

Gunn, A. (2023). *The age of generative AI in academia: An opinion.* SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4382111

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models*. ArXiv. https://arxiv.org/abs/2302.02337

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1682

Herwix, A., Haj-Bolouri, A., Rossi, M., Tremblay, M. C., Purao, S., & Gregor, S. (2022). Ethics in Information Systems and Design Science Research: Five Perspectives. *Communications of the Association for Information Systems*, *50*(1), 589-616.

Homolak, J. (2023). Exploring the adoption of ChatGPT in academic publishing: Insights and lessons for scientific writing. *Croatian Medical Journal*, 64(3), 205-207.

Kim, S. G. (2023). Using ChatGPT for language editing in scientific articles. *Maxillofacial*

Plastic and Reconstructive Surgery, *45*(1), Article 13.

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet Digital Health*, *5*(3), e105-e106.

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology*, *39*(1), 1-8.

Nature (2023a). *Tools such as ChatGPT threaten transparent science; here are our ground rules for their use* [editorial]. https://www.nature.com/articles/d41586-023-00191-1

Nature (2023b) *Why Nature will not allow the use of generative AI in images and video* [editorial]. https://www.nature.com/articles/d41586-023-01546-4

OpenAI (2023). *Gpt-4 technical report 2023*. Available at https://cdn.openai.com/papers/gpt-4.pdf

Ptaszynski, M., Lempa, P., Masui, F., Kimura, Y., Rzepka, R., Araki, K., ... & Leliwa, G. (2019). Brute-force sentence pattern extortion from harmful messages for cyberbullying detection. *Journal of the Association for Information Systems*, *20*(8), 1075-1127.

Russell, S. & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.

Samek, W., Wiegand, T., & Müller, K. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.* ArXiv. https://arxiv.org/abs/1708.08296

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109-124.

Shneiderman, B. (2022). *Human-centred AI*. Oxford University Press.

Simpson, K. & Day, N. (2004). *Field guide to the birds of Australia.* (7th ed.). Penguin.

Solomon, L. & Davis, N. (2023). *The state of AI governance in Australia*. Human Technology Institute, University of Technology Sydney.

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove [news]. *Nature*, 613, 620-621.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24-31.

van Dis, E., Bollen, J., van Rooij, R., Zuidema, W. & Bockting, C. (2023). ChatGPT: Five priorities for research [comment]. *Nature*, *614*, 224-226.

Wiechetek, L., Pirinen, F., Hämäläinen, M., & Argese, C. (2021). Rules ruling neural networks–neural vs. rule-based grammar checking for a low resource language. *Proceedings of the International Conference on Recent Advances In Natural Language Processing*.

Wu, T. et al. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122-1136.

Zirpoli, C. T. (2023). *Generative artificial intelligence and copyright law*. Congressional Research Service. https://crsreports.congress.gov/product/pdf/LSB/LSB10922

About the Authors

Shirley Gregor is a professor emerita at the Australian National University. Her research interests include artificial intelligence, human-computer interaction, and the philosophy of science and technology. Her work has appeared in leading journals and she has led a number of applied research projects with industry, where she has been able to apply and develop her ideas on design science. Work on e-commerce with the beef industry led to her being made an Officer of the Order of Australia in 2005. She was editor-in-chief for the *Journal of the Association of Information Systems* from 2010-2013. She is a Fellow of both the Australian Computer Society and the Association for Information Systems. She was awarded a DESRIST Lifetime Achievement Award in 2017 for contributions to design science research in information systems and technology.