... ZAOULI N°13, Vol. 4, Mai 2026, pp. 3-38 ISSN : 2788-9343

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

The ethics of artificial intelligence in scientific research: a comparative study of the debates.

Ahmed GHEZAL
Mohamed Boudiaf University - M'sila,
Faculty of Law and Political Science, Algeria.
Email: ahmed.ghezal@univ-msila.dz

Abstract:
The integration of artificial intelligence (AI) into scientific research has sparked one of the most significant ethical debates in contemporary scholarship. Across disciplines, from biomedical science to political science, scientists, journal editors, funding bodies and international organisations are now confronted with a set of common yet contentious questions: What constitutes the responsible use of AI in research? Who is responsible when AI-generated outputs mislead? Where is the line between legitimate assistance and unethical academic conduct drawn? This article examines the rapidly evolving landscape of AI ethics in relation to research integrity, outlining areas of emerging consensus alongside persistent and sometimes unresolved points of contention. Drawing on key ethical frameworks (Floridi et al., 2018; Jobin, Inka & Faina, 2019; Mittelstadt, 2019), notable institutional responses (UNESCO, 2021; COPE, 2024; ICMJE, 2023) and interdisciplinary discussions within political science and computational social science, the article contends that, although there is nominal convergence around core principles such as transparency, accountability, fairness and human oversight, deep normative, structural and geopolitical fissures remain unresolved. The implications for the integrity of political science research are particularly acute given the field's reliance on interpretive judgement, contextual validity and political sensitivity.

Keywords: *AI ethics, research integrity, big language models, transparency, academic authorship, UNESCO.*

Résumé :
L'intégration de l'intelligence artificielle (IA) dans la recherche scientifique a suscité l'un des débats éthiques les plus importants du monde universitaire contemporain. Dans toutes les disciplines, des sciences biomédicales aux sciences politiques, les chercheurs, les rédacteurs en chef de revues, les organismes de financement et les organisations internationales sont désormais confrontés à une série de questions communes mais controversées : qu'est-ce qui constitue une utilisation responsable de l'IA dans la recherche ? Qui est responsable lorsque les résultats générés par l'IA induisent en erreur ? Où se situe la frontière entre une aide légitime et une conduite académique contraire à l'éthique ? Cet article examine le paysage en rapide évolution de l'éthique de l'IA en relation avec l'intégrité de la recherche, en soulignant les domaines où un consensus émerge ainsi que les points de discorde persistants et parfois non résolus. S'appuyant sur des cadres éthiques clés (Floridi et al., 2018 ; Jobin, Inka & Faina, 2019 ; Mittelstadt, 2019), des réponses institutionnelles notables (UNESCO, 2021 ; COPE, 2024 ; ICMJE, 2023) et des discussions interdisciplinaires au sein des sciences politiques et des sciences sociales computationnelles, l'article soutient que, bien qu'il existe une convergence nominale autour de principes fondamentaux tels que la transparence, la responsabilité, l'équité et le contrôle humain, de profondes fissures normatives, structurelles et géopolitiques restent sans solution. Les implications pour l'intégrité de la recherche en sciences politiques sont particulièrement aiguës compte tenu de la dépendance de ce domaine à l'égard du jugement interprétatif, de la validité contextuelle et de la sensibilité politique.

***Mots-clés** : éthique de l'IA, intégrité de la recherche, grands modèles linguistiques, transparence, paternité des travaux universitaires, UNESCO.*

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

Introduction:

The emergence of generative artificial intelligence and the development of large language models (LLMs), which became part of mainstream academic practice following the release of ChatGPT in November 2022, has not raised entirely new ethical questions about science. However, it has amplified existing concerns and emphasised their urgency. The principle of research integrity, which requires honesty, accuracy and transparency when conducting and disseminating scientific research[1], has always been fundamental to academic practice (Resnick, 2011). However, what artificial intelligence has done is subject this principle to new and poorly understood variables, giving rise to forms of misconduct such as AI-generated data falsification, plagiarism via algorithmic tools, fake citations and ambiguous lines of analysis. Organisational ethics documents do not always clearly address this issue (Lu, Zhang, and Chen, 2024).

The academic community reacted quickly, but was divided on specific principles. Within two years of ChatGPT's release, every major publisher, funding body, and international organisation had expressed their position on the use of artificial intelligence in research. By 2024, the Committee on Publication Ethics (COPE), the International Committee of

[1]- David B. Resnik, 'Scientific Research and the Public Trust', Science and Engineering Ethics, 17(3): 399–409. 3 (2011): 399–409. Available at: https://doi.org/10.1007/s11948-010-9210-x.

Medical Journal Editors (ICMJE)[2], Nature, Science, Elsevier and Springer Nature[3] had issued policies and positions that, as this article will demonstrate, were largely similar, with clear differences regarding the limits of permissible use, disclosure requirements and underlying normative justifications[4].

Beyond the world of publishing, the debate has moved to the level of global governance. UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence, which was unanimously adopted by all 193 Member States, is the first truly global standard-setting instrument in this field (UNESCO, 2021)[5].

However, there is still a significant gap between the ambitious principles set out in the Recommendation and the reality of AI deployment in research environments in the Global South, authoritarian states and under-resourced institutions.

---

[2]- "Welcome to COPE", COPE: Committee on Publication Ethics, 26 November 2025, https://publicationethics.org/welcome-cope.

[3]- Jaime A. Teixeira da Silva, 'The ICMJE Recommendations: Challenges in Fortifying Publishing Integrity', Irish Journal of Medical Science, 189(1971-), 1179–81, https://doi.org/10.1007/s11845-020-02227-1. 4 (2020): 1179–81, https://doi.org/10.1007/s11845-020-02227-1.

[4]- Mike Perkins and Jasper Roe, 'Academic Publisher Guidelines on AI Usage: A ChatGPT-Supported Thematic Analysis', preprint, F1000Research, 16 January 2024, https://doi.org/10.12688/f1000research.142411.2.

[5]- 'Recommendation on the Ethics of Artificial Intelligence', UNESCO Digital Library, accessed 10 November 2025, https://unesdoc.unesco.org/ark:/48223/pf0000380455_ara. Scientific, United Nations Educational, Cultural and Scientific Organization, 'Recommendation on the Ethics of Artificial Intelligence', 2021.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

In order to study this issue, the fundamental question was: What new ethical challenges does the use of artificial intelligence pose? This was addressed by identifying a set of foundational documents discussing the ethics of scientific research.

Research methodology:

The research is based on content and thematic analysis. The article and the book are both divided into five sections. The first section provides an overview of the fundamental principles of AI ethics, key researchers, and the associated debates. It then examines... The second section addresses the ethics of artificial intelligence, including issues of integrity in scientific research such as authorship, transparency, reproducibility, and misconduct. This section addresses these issues. The third section addresses the issues raised by the field of political science, where matters of bias, representation, and political sensitivity accumulate within broader ethical challenges. The fourth section finally addresses these issues. It provides a comparative study of fundamental and unresolved normative debates. The conclusion summarises the state of the field and identifies priorities for future institutional and academic attention.

1. First topic: Research ethics in the age of artificial intelligence: principles, frameworks, and researchers.

1.1 Framework of artificial intelligence: Basic principles

The rapid proliferation of AI ethics guidelines, described by Jobin, Inca, and Vajna (2019) as the 'ethical AI boom', has

created a landscape that is impressive[6] in its breadth and troubling in its contradictions. In their pioneering meta-analysis of 84 AI ethics documents from public and private organisations worldwide, Jobin, Inca and Vajna identified five recurring principles: transparency, fairness and equity, do no harm, responsibility and privacy. These principles emerged as the most frequently cited across documents from diverse cultural and legal contexts[7]. Floridi and colleagues (2018, 2019) proposed a unified synthesis of these principles under the acronym ASILOMAR, later simplified to five core values: benevolence, do no harm, autonomy, fairness and explicability.

Table 1: Basic principles in AI ethics

Degree of consensus	Definition	Principle
Very high (almost universal)	The workings of artificial intelligence systems must be understood and disclosed.	Transparency
high	Clear lines of responsibility regarding the outcomes of artificial intelligence	accountability

[6]- Anna Jobin et al., 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1(9), 2019, pp. 389–399.

[7]- Luciano Floridi et al., "AI4People: An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', Minds and Machines 28, no. 4 (2018): 689–707, https://doi.org/10.1007/s11023-018-9482-5

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

Highly principled; practically disputed	The results should not repeat the unfair bias.	Justice/Non-discrimination
high	Artificial intelligence must deliver a net social benefit.	charity
high	Artificial intelligence should not cause harm.	No harm
high	Humans retain ultimate decision-making power.	Independence/Human Control
high	The use of data must respect individual rights.	Privacy
Moderate (growing in the market)	The development of artificial intelligence should not deplete environmental resources.	Sustainability
Lowest; subject to dispute	Artificial intelligence should not deepen social inequalities.	Solidarity

Sources: Jobin, Inca and Faina (2019); Floridi et al. (2018); UNESCO (2021); and Hagendorf (2020).

However, despite this apparent consensus[8], Mittelstadt (2019) presented a counter-argument. In the journal Nature Machine Intelligence, he argued that the apparent consensus around higher principles should not be celebrated prematurely because it masks 'a deep political and normative disagreement'. This is due to the frequent analogy drawn between the ethics of artificial intelligence and the ethics of biomedicine. This analogy is misleading because four of these principles (autonomy, benevolence, do no harm, and justice) have been underpinned by decades of institutional practice. Furthermore, the development of artificial intelligence lacks the structural conditions that made the biomedical framework successful, such as clear and applicable operational principles, a professional history of standards, proven methods for translating principles into practice, and robust accountability mechanisms (Mittelstadt, 2019). Subsequent research on ethics has validated this critique[9]. Hagendorf (2020) demonstrated that, despite verbally committing to established principles, the vast majority of corporate AI ethics documents lacked concrete implementation mechanisms[10]. Green, Hoffman, and Stark (2019) also showed that industry-sponsored ethics documents

---

[8]- Luciano Floridi, 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical', Philosophy & Technology, 32(2), 2019, pp. 185–193. https://doi.org/10.1007/s13347-019-00354-x. Brent Mittelstadt, 'Principles Alone Cannot Guarantee Ethical AI', Nature Machine Intelligence 1, no. 11 (2019): 501–507.

[9]- Brent Mittelstadt, 'Principles Alone Cannot Guarantee Ethical AI'.

[10]- Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines': Publisher Correction. 2020, https://psycnet.apa.org/record/2020-56191-001.

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

effectively function as tools for regulatory proactivity[11]. This mechanism, known as 'ethics whitewashing', delays binding regulation by focusing the discussion on abstract principles and technical solutions while avoiding questions of structural power. These criticisms are not merely academic; they directly impact the ability of AI ethics frameworks to protect the integrity of scientific research or address academic reputation issues.

1.2 UNESCO Framework and Global Governance

The 2021 UNESCO Recommendation on the Ethics of Artificial Intelligence occupies a place within this landscape. Unlike institutional or academic codes[12], it is a multilateral intergovernmental standard, adopted by all 193 Member States following a three-year, multidisciplinary consultation process involving experts from 155 countries (UNESCO, 2021). The experts who oversaw the drafting of this international document identified four fundamental principles that they believe should govern the use of artificial intelligence: respect for human rights and dignity; living in peaceful and interconnected societies; ensuring diversity and inclusion; and maintaining a thriving environment. Eleven sub-issues also branched out from these principles, including

---

11- Daniel Greene et al., 'Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning, 2019, https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/2/.

12- UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (United Nations Educational, Scientific and Cultural Organization: UNESCO, 2021).

data governance, education, culture, gender equality and research.

In the service of research integrity, UNESCO's recommendations emphasised two principles with direct operational implications: transparency and explainability (AI systems must be understandable to users and those affected by them) and human oversight (member states must ensure that AI does not replace ultimate human responsibility and accountability). The Recommendation also mandated ethical impact assessment as a policy tool, recommending the establishment of a global observatory for AI ethics and governance to monitor implementation (UNESCO, 2021; Montreal Institute for AI Ethics, 2021)[13].

Adopting such a recommendation requires a consensus formula, but global governance researchers have noted its limitations. As a UNESCO recommendation rather than a binding treaty, it lacks implementation mechanisms, and its implementation depends entirely on national will. This gap is particularly evident in countries with authoritarian governance, where the use of artificial intelligence for surveillance and political control continues to clash with the human rights frameworks set out in the recommendation.

1.3 Classification of ethical approaches

The UNESCO document and other sources suggest that the academic debate on the ethics of artificial intelligence is organised around three broad theoretical approaches: the principleist approach, which identifies abstract normative principles (Floridi et al., 2018; Jobin et al., 2019)[14]; the

[13]- Renjie Butalid, 'About – The AI Ethics Brief', accessed 3 May 2026, https://brief.montrealethics.ai/about.

[14]- Floridi et al., 'AI4People: An Ethical Framework for a Good AI Society'.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

dependent approach, which evaluates AI systems based on their actual social outcomes; and the rights-based approach, which grounds AI ethics in human rights law[15]. Each approach prioritises different aspects of research integrity: proponents of the principleist approach focus on procedural standards such as disclosure and transparency; proponents of the dependent approach focus on measured harms such as biased outputs and fabricated data; and proponents of the rights-based approach focus on structural questions of power and access, and who bears the costs of AI errors.

Table 2: Diagram of conceptual levels of governance and ethics in artificial intelligence.

Institutional documents and practices	**Level**
recommendation UNESCO (2021) · principles system nations United (2022) agreement council Europe For intelligence Artificial (2024)	Level Global
law intelligence artificial Union European (2024) strategy Union African For intelligence artificial principles intelligence artificial Organization cooperation Economic And development (2019 , (Updated 2023)	Level regional
Charters of organizations such as: COPE COPE , ICMJE , Nature , Sciences , Gama , IEEE	Level Professional

[15]- Jobin et al., 'The Global Landscape of AI Ethics Guidelines'.

Rules Private By specialization (Science) social, Medicine, the law)	
Policies the university requirements Funding (NSF, ERC) instructions office Search	Level institutional
Practice Individual, Disclosure, choice Methodological	level researcher

2. Section Two: Ethics of Artificial Intelligence and Integrity of Scientific Research

2.1 The Status of Research Integrity in the Age of Artificial Intelligence

Traditionally, research integrity has been understood in relation to the FFP triad of prohibiting fabrication, forgery and plagiarism, while supporting transparency, reproducibility and attribution. However, artificial intelligence techniques are challenging all of these dimensions simultaneously and in innovative ways.

Fabrication and forgery have become possible due to the ability of generative AI models to produce outputs that appear plausible yet are factually incorrect, a phenomenon known as 'hallucinations'. Several articles have already been retracted due to authors falling into this trap. For example, a 2023 research paper in Physica Scripta was retracted after reviewers identified AI-generated passages containing forged citations, and legal documents in several jurisdictions have cited non-existent case law generated by ChatGPT[16]. Experts familiar with the workings of AI language models explain that the problem is technical, not accidental: large models are

[16]- Ziyu Chen et al., 'Research Integrity in the Era of Artificial Intelligence: Challenges and Responses', Medicine 103, no. 27 (2024): e38811.

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

trained to produce smooth scripts that meet statistical expectations, not to verify accuracy.

However, for political science researchers who rely on historical sources, government documents, and archival materials, many of which may be underrepresented in training sets, the risk of AI-generated misinformation is openly and blatantly present through the creation of fake, non-original texts by a linguistic model.

Regarding plagiarism, it has become a complicated issue for two reasons. Firstly, AI tools can reproduce texts from training data without attributing them to their sources, which may include copyrighted or unpublished material. Secondly, using AI to generate unattributed texts raises questions about the credibility of academic authorship, as it is a form of intellectual misrepresentation that current plagiarism detection tools cannot identify — AI-generated text may be officially authentic and knowledge-derived[17].

Thirdly, transparency and reproducibility are challenged by the fact that AI analysis lines, especially those using commercial LLM APIs, may change their outputs over time, meaning that research conducted today cannot be reproduced using the same tools six months later. This means that the transparency of training data, model weights and fine-tuning procedures in proprietary systems (OpenAI, Google and Anthropic) means AI-powered research cannot be fully audited by external reviewers, even in principle.

The issue of authorship: an area of unusual consensus

If there is one area where the discourse on AI ethics has achieved genuine institutional convergence, it is the

[17]- Perkins and Roe, 'Academic Publisher Guidelines on AI Usage'.

prohibition on listing AI systems as authors of academic papers. The Committee on Publication Ethics (COPE) has reached this conclusion, as have numerous journals, including Nature, Science, and JAMA, as well as scientific journal platforms such as Elsevier and Springer. Nature, IEEE, Cambridge University Press, and the International Committee of Medical Journal Editors (ICMJE) have reached the same conclusion, albeit for slightly different reasons.[19].

These bodies generally agree that authorship implies accountability, but AI systems cannot be held accountable. As the ICMJE stated in 2023: 'Chatbots (such as ChatGPT) should not be listed as authors because they cannot be held responsible for the accuracy, integrity and originality of the work, and these responsibilities are required for authorship.' Nature's policy adds that the use of AI must be documented in the methodology section, while Science prohibits the attribution of AI-generated content to original work.

However, this consensus has been challenged from a philosophical standpoint. Neil Levy, writing in the Journal of Medical Ethics, argued that the justification for COPE (the organisation for the evaluation of authors in the natural sciences) is philosophically incoherent because liability is a special form of liability. According to COPE, principal investigators routinely list contributors who have done work without assuming legal or ethical responsibility for the whole. Instead, Levy suggested that COPE should either revise its standard for authorship or accept that AI contributions constitute a category requiring a new conceptual framework, rather than exclusion based on classification. However, this philosophical counter-argument has not yet altered institutional practices. Despite its breadth, the consensus on

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

authorship is based on a less stable justification than might be suggested by its almost universal acceptance.

2.2 The issue of authorship: an area of unusual consensus

If there is one area where the discourse on AI ethics has achieved genuine institutional convergence, it is the prohibition on listing AI systems as authors of academic papers. This conclusion has been reached by the Committee on Publication Ethics (COPE), numerous journals (including Nature, Science, and JAMA)[18], and scientific journal platforms (such as Elsevier and Springer). Other organisations that have reached the same conclusion include Nature, IEEE, Cambridge University Press, and the International Committee of Medical Journal Editors (ICMJE)[19], albeit for slightly different reasons.

These bodies generally agree that authorship implies accountability, but AI systems cannot be held accountable. As the ICMJE stated in 2023: 'Chatbots (such as ChatGPT) should not be listed as authors because they cannot be held responsible for the accuracy, integrity, and originality of the work, and these responsibilities are required for authorship. [20]' Nature's policy adds that the use of AI must be

---

[18]- Teixeira da Silva, 'The ICMJE Recommendations'.

[19]-COPE: Committee on Publication Ethics, 'Welcome to COPE'; Ju Yoen Lee, 'Can an Artificial Intelligence Chatbot Be the Author of a Scholarly Article?', Journal of Educational Evaluation for Health Professions 20 (2023), https://synapse.koreamed.org/articles/1516081874.

[20]- 'ICMJE | Recommendations | Preparing a Manuscript for Submission to a Medical Journal', accessed 3 May 2026, https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html.

documented in the methodology section, while Science prohibits the attribution of AI-generated content to original work.

This consensus has been challenged from a philosophical standpoint. Neil Levy, writing in the Journal of Medical Ethics, argued that the justification[21] for COPE (the organisation for the evaluation of authors in the natural sciences) is philosophically incoherent because liability is not actually the standard for such evaluation. Principal investigators routinely list contributors who have done work without assuming legal or ethical responsibility for the whole. Instead, Levy suggested that COPE should either revise its standard for authorship or accept that AI contributions constitute a category requiring a new conceptual framework rather than exclusion based on classification. However, this philosophical counter-argument has not yet altered institutional practices. Despite its breadth, the consensus on authorship is based on a less stable justification than might be suggested by its almost universal acceptance.

[21]- Neil Levy, 'Responsibility is not required for authorship', Journal of Medical Ethics, 51(no. 4 (2025): 230–232.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

Table 3: Policies on artificial intelligence and authors (2023–2024)

Date of update	AI ban on peer review	Disclosure required	Artificial Intelligence Composition	Institution/Publisher
2023 (Updated 2024)	Yes	Yes	forbidden	cup
2023	Yes	Yes (cover letter + research paper)	forbidden	ICMJE
January 2023	Yes (declare use)	Yes (and its inclusion in the methodology section)	forbidden	Nature magazine
January 2023	Yes	Yes	Prohibited (AI text is prohibited as native text)	Science magazine
2023	Yes	Yes , it's listed in the section: No	forbidden	Elsevier portal

		thanks or methodology .		
April 2024	Yes	Yes (thank you)	forbidden	IEEE
2024 (Updated)	Yes	Yes (comprehensive framework)	forbidden	Gamma (JAMA)
March 2023	Yes	Yes	forbidden	Cambridge University Press

Sources: Perkins et al. (2024); Kennesaw State University Research Office (2024); Cobb (2024).

2.3 Disclosure: Consensus on the standard; disagreement on the pattern.

Regarding the inclusion of artificial intelligence and its linguistic models as authors, there is still disagreement about how to disclose the use of AI. JAMA has the most comprehensive framework, with automatic submission screening and detailed requirements for describing the use of AI in research design. By contrast, Nature requires disclosure for some uses, but explicitly exempts copy-editing assistance from disclosure requirements. The European Code of Conduct for Research Integrity (revised 2023) states that researchers must report on their 'results and methods, including the use of external services or artificial intelligence and automated tools', and considers failure to do so a

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

violation of research integrity rather than misconduct per se (Resnick et al., 2024).

The lack of standardisation across institutions creates a fragmented landscape that disadvantages researchers in low-resource settings who lack institutional guidance. It also raises a fundamental concern: if disclosure requirements vary significantly between journals and disciplines, the standard of transparency is undermined, even when formally observed. For political science journals, which have lagged behind the natural and biomedical sciences in formulating AI policies, the practical guidance available to researchers remains inadequate.

2.4 Training data, transparency and the 'black box' problem

Perhaps the most technically challenging issue for research integrity in the age of artificial intelligence is the 'black box problem', which is addressed by the UNESCO Recommendation on 'interpretability'. UNESCO points out that this term refers to AI systems that are 'obscure and difficult to interpret' — systems in which the relationship between inputs and outputs cannot be fully reconstructed by external observers (UNESCO, 2021).

For scientific research, this opacity has several dimensions. Firstly, the training data used to create commercial big language models is often not disclosed, and these models may contain biases reflecting the demographic and linguistic characteristics of internet content (especially English content from the Global North that is time-limited), as well as copyrighted material without proper publication identification. Secondly, there is the issue of model behaviour: the outputs of large models can vary depending on the

version of the model, the API calls and the temperature settings used, which makes it impossible to accurately reproduce the results. Thirdly, fine-tuning and alignment: proprietary models undergo post-training modifications (e.g. human feedback reinforcement learning, RLHF) that shape their outputs in ways which are not entirely transparent, even to the researchers using them. As Lu, Zhang and Chen (2024) observe, 'When AI algorithms are used to process data or generate results, researchers may not fully understand the workings and decision-making processes of these algorithms', which violates the transparency required for research integrity.

3. Third topic: The ethical challenges of artificial intelligence in political science and the fragility of political science research.

3.1 The particular fragility of political science research

Political science occupies a unique position among disciplines affected by artificial intelligence. While the risk of AI-induced hallucinations in natural science research is primarily a matter of factual accuracy, political science research is grounded in interpretive, normative and politically sensitive contexts. Here, AI-generated errors can have far-reaching consequences beyond the academic paper. For example, a fabricated citation in a political sociology study could either reinforce authoritarian rule or challenge political narratives. Similarly, a biased large language model summarising election data could systematically distort findings about voter behaviour, and an AI-generated simulation of public opinion could reinforce the very stereotypes it purports to study.

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

Three particularly prominent categories of challenge in political science require attention from researchers: algorithmic bias, hallucination in interpretive contexts, and simulation validity.

3.2 Algorithmic bias in political research

Large language models that are used in political science research are known to carry biases that cannot easily be mitigated. In the MIT Technology Review, researcher Hekella reported that large language models exhibit systematic political biases traceable to the structure of their training data. A more rigorous study by Exler et al. (2025) found that the political bias of large language models[22] is related to their size, with larger models tending to exhibit a more pronounced leftward bias on Western political scales. This suggests that the scale amplifies, rather than balances, the ideological bias present in the training data[23].

These algorithmic and language biases have direct implications for political science research. Numerous examples illustrate this: researchers using large language models to code political texts, analyse party data, simulate survey responses or summarise legislative debates are susceptible to importing biases that confound their findings. Studies that use large language models as substitutes for human survey participants are particularly prevalent and represent a rapidly growing methodological practice in

[22]- Melissa Heikkilä, 'AI language models are rife with different political biases', MIT Technology Review. Retrieved 5 April 2024.

[23]- David Exler et al., 'Large Means Left: Political Bias in Large Language Models Increases with Their Number of Parameters', Arxiv, preprint, 7 May 2025, https://doi.org/10.48550/Arxiv.2505.04393.

computational political science[24]. Research by Argyle et al. (2023) demonstrated that large language models can accurately reproduce the average responses of demographic groups. However, subsequent work on representational bias revealed that performance deteriorates significantly for non-English-speaking groups, multi-party political systems and non-democratic contexts — precisely the environments of interest to comparative studies specialists. This is particularly pertinent when considering issues relating to the Global South or authoritarian regimes.

Table 4: Classification of risks associated with artificial intelligence in political science research

severity	Methodological field	Description	Risk category
high	Literature review, historical analysis	Artificial intelligence produces plausible but erroneous citations, events, or results.	hallucinations
high	Text analysis, opinion simulation, content coding	LLM outputs are systematically ideologically driven	political bias

[24]- Lisa P. Argyle et al., 'Out of One, Many: Using Language Models to Simulate Human Samples', Political Analysis, 31(3), pp. 3 (2023): 337–351.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

high	Comparative research across countries	Performance deteriorates in non-Western, non-English, and non-democratic contexts.	representation bias
moderate	Current Affairs Analysis	The training cutoff creates methodological blind spots for recent events	Time deduction
high	Any analysis aided by artificial intelligence	Results vary between model versions; pipelines cannot be audited.	Opacity/Non-reproducibility
high	Searching using restricted datasets	Proprietary or sensitive data that goes into commercial AI tools	Breach of confidentiality
Medium-high	Academic writing in all sub-disciplines	AI-generated texts are used without disclosure.	failure of the ratio

Sources: Law et al. (2024); Exler et al. (2025); Argyle et al. (2023); Perspectives in Political Science (2023).

3.3 Artificial intelligence simulation of political behaviour: Promises and Risks

Simulating human political behaviour is one of the fastest-growing applications of large language models in political science, using artificial intelligence to approximate survey responses, voting behaviour or deliberative reasoning without the need for direct data collection. The methodological appeal of simulation lies in its advantages: reduced costs, scalability, and the ability to simulate populations that are difficult to reach. However, ethical and epistemological considerations are also becoming increasingly relevant.

Researchers have found that large language models exhibit 'temporal variability, steerability, and political bias', characteristics that render them unreliable as neutral tools for studying political opinion (Long-Range Study of Large Language Models, 2024). The detection of steerability is particularly troubling: the outputs of large language models can be systematically altered through rapid formulation in ways that an unsuspecting researcher might not detect. This adds variability that is researcher-controlled and mimics sound, systematic experimental manipulation, but lacks its epistemological foundation[25].

A cautionary example is the German Federal Election 2025 study. Researchers found that AI-powered election advice apps (VAAs) designed to 'inform voters objectively' deviated

[25]- Ankit Sabharwal et al., 'VaaS: A Multi-Layer Hallucination Reduction Pipeline for AI-Assisted Science: Production Validation and Prospective Benchmarking', medRxiv, 2026, 2026–03.

Received: 17/06/2025 Accepted: 12/02/2026 Published: 13/05/2026

from stated party positions in 25–50% of cases. They also found that simple, quick questions or prompts produced intense delusions, including fabricated links between political parties and extremist organisations (AI Election 2025 Cautionary Study). This is not merely a technical failure; it demonstrates how AI systems used in politically sensitive research contexts can generate disinformation with potentially surprising political consequences that could affect the democratic process in unexpected ways.

3.4 Generative AI in Academic Writing: The Limits of Integrity

The use of generative artificial intelligence in drafting academic manuscripts is increasing. This includes generating, designing and correcting abstracts, literature reviews, research discussion sections within scientific papers and articles, and even full papers. This has led to the tangible emergence of the issue of research integrity in everyday academic practice. Several attempts have been made in response to clarify the ethical principles and boundaries surrounding the use of artificial intelligence in scientific writing. A framework proposed by researchers at several institutions identifies seven areas for the use of generative AI in scientific research: idea generation; questionnaire design; literature review; data generation; annotation; data enhancement; and academic writing (GenAI Research Ethics Framework, 2025). The common thread across all these areas is the risk that automating cognitively demanding scientific tasks could lead to a decline in the explanatory quality and cognitive responsibility that define rigorous research.

For political science specifically, this anxiety is amplified by the interpretive nature of the discipline. Unlike in clinical research, where an AI-generated methods section might differ only slightly from a human-written one, political analysis involves judgements about context, significance, causality and power that cannot be delegated to a system trained on statistical text patterns without incurring a significant cognitive cost. As Floridi and his colleagues have consistently argued, the ethical deployment of AI in research must maintain human cognitive independence, which is the ability of a researcher to form[26], review and defend judgements based on evidence and logical reasoning.

4. Section Four: A comparative study of points of agreement and disagreement

4.1 From principles to practice: the implementation gap

The most significant point of contention in the ethics of artificial intelligence, acknowledged by scholars across the ideological spectrum, is the gap between stated principles and their institutional implementation[27]. Hagendorf (2020) found, in a systematic analysis of 22 AI ethics guidelines, that most documents failed to translate abstract standards such as transparency and fairness into concrete, actionable requirements[28]. T. Morley et al. (2020) identified similar

---

[26]- Floridi et al., 'AI4People: An Ethical Framework for a Good AI Society'.

[27]- Hagendorff, 'The Ethics of AI Ethics'.

[28]- Jessica Morley et al., 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods, and Research to Translate Principles into Practices', in Ethics, Governance and Policies in Artificial Intelligence, Vol. 144, Ed. Luciano Floridi, Philosophical Studies Series (Springer International Publishing, 2021), https://doi.org/10.1007/978-3-030-81907-1_10.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

difficulties in moving from 'what' to 'how', noting that the fragile consensus around the presented principles and 'foundational' approach did not provide a definitive answer to the deeper question of how to prioritise competing values when they conflict in specific research contexts.

This gap is not merely a technical issue[29]. Rather, as Mittelstadt (2019) argues, it reflects the absence of an institutional infrastructure that gives principles practical force. In the field of medicine, for example, this infrastructure could include an authority that grants professional licences, establishes liability codes and regulates peer accountability, as well as regulatory agencies with genuine enforcement power. While an emerging regulatory landscape is gradually taking shape, most notably in the form of the European Artificial Intelligence Act 2024, which sets out risk-based requirements for AI systems and addresses this gap within the European context, its extraterritorial application to research institutions and publishers outside the EU remains limited.

4.2 Geopolitical and cultural conflicts

Structurally significant disagreements in AI ethics are not about search criteria per se, but rather about the geopolitical and cultural foundations of the emerging global framework. This reflects the convergence of principles documented by Jobin, Inca, and Vajna (2019)[30]. According to Floridi and his colleagues (2018)[31], it essentially stems from a Western liberal

[29]- Mittelstadt, 'Principles Alone Cannot Guarantee Ethical AI'.

[30]- Jobin et al., 'The Global Landscape of AI Ethics Guidelines'.

[31]- Floridi et al., 'AI4People: An Ethical Framework for a Good AI Society'.

philosophical tradition. This is because principles such as individual autonomy, freedom of information, and checks on state power are interpreted differently in political systems where the state is considered the legitimate custodian of information, where collective interests take precedence over individual ones, or where digital sovereignty is deemed a vital national security interest.

For instance, China's AI governance framework prioritises 'AI for social governance', positioning the state as the legitimate overseer of AI development, which differs fundamentally from the UNESCO recommendation that emphasises human rights and multilateralism. The absence of robust, multilateral implementation mechanisms means that the global AI ethics framework effectively operates in parallel with national frameworks that may be systemically inconsistent with its stated values. For cross-border research on AI and political systems — a natural domain of comparative policy — this fragmentation presents ethical and epistemological challenges.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

Points of agreement

- Artificial intelligence systems cannot be listed as authors.
- Disclosure of AI use is an ethical requirement .
- Transparency and human oversight are non-negotiable principles.
- Data falsification and hallucination are forms of misconduct.
- The confidentiality of peer review prohibits the use of artificial intelligence for manuscripts.

Disputed points

- The amount of AI assistance allowed before crossing over
- Misconduct? (Disagreement with drawing lines)
- Who governs the ethical framework? (A geopolitical conflict)
- How should the contributions of artificial intelligence be attributed? (A normative dispute)
- Does "responsibility" require human authorship? (Philosophical)
- How can compliance be enforced across jurisdictions? (Governance)
- Are current detection tools sufficient? (Technical dispute)
- How to balance openness and secrecy in the game with the help of artificial intelligence
- Peer review? (Institutional conflict)

4.3 A critique of 'ethics whitewashing' and the power of industry

In the field of artificial intelligence ethics as a whole, there is a frequent and pressing criticism that has direct implications for the integrity of research. The spread of ethical frameworks that serve corporate and economic interests at the expense of academic considerations is one such criticism. This phenomenon is referred to as 'ethics laundering' (Green, 2019; Whitaker et al., 2018)[32], whereby AI companies and industry-funded bodies strategically disseminate ethical discourse to delay binding regulation, avoid public scrutiny, and frame the debate around technical rather than political solutions.

This criticism resonates particularly strongly in the context of research integrity because the dominant AI systems used in academic research, such as OpenAI's GPT series, Google Gemini and Anthropic's Cloud, are commercial products developed by private entities whose financial interests in conducting research are clear. When these companies produce AI ethics data or support ethics research, a serious question must be asked: Who has the moral authority and power to set the ethical agenda? Are the ethical problems being highlighted the ones that most affect research integrity, or those that are most compatible with continued commercial expansion?

Areas where commercial interests and research integrity interests may diverge include the lack of clarity in training data, the unreproducibility of commercial API outputs, and

---

[32]- Greene et al., 'Better, Nicer, Clearer, Fairer'; Meredith Whittaker et al., 'AI Now Report 2018', AI Now Institute at New York University, New York, 2018, available at: https://www.elindependiente.com/wp-content/uploads/2018/12/Informe-AI-Now.pdf.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

terms of service that may prohibit the use of research outputs for certain purposes, all of which have been relatively overlooked by corporate ethics frameworks.

4.4 Methodological disagreements: When is the use of artificial intelligence (AI) legitimate in scientific research?

Practical disagreements within the social and political sciences concern the legitimate scope of AI's use. There is a general consensus that using large language models to generate entire datasets, fabricate interview responses or produce unattributed literature reviews undermines research integrity. But what about using AI to translate non-English source material? Can it assist with the qualitative coding of large datasets? Or for creating initial drafts of methodological sections that undergo extensive review? Or for identifying potential case studies or literature for comparative research design?

(2020) and the Ethical Framework (2025) [33] suggest that the limit lies at the point at which artificial intelligence replaces human cognitive judgement, rather than assisting it [34]. While this is a useful theoretical standard, it is difficult to apply in practice, particularly for researchers who may not recognise when the use of AI crosses the line from providing legitimate assistance to replacing cognitive processes and undermining knowledge. In the field of political science, unlike in fields such as biomedical research, there are no discipline-specific guidelines to provide an institutional solution to these questions.

---

[33]- Morley et al., 'From What to How'.

[34]- Floridi et al., 'AI4People: An Ethical Framework for a Good AI Society'.

Conclusion:

The landscape of AI ethics, as it relates to the integrity of scientific research, as this article has shown, is an area of asymmetric development: rapid institutionalization at the surface level of principles and declarations, alongside persistent and deep disagreements at the levels of implementation, enforcement, normative basis, and geopolitical legitimacy.

What is agreed upon: The academic community has reached a position that AI systems cannot be authored; that the use of AI must be disclosed; that hallucinations and AI-assisted manufacturing constitute research misconduct; and that human oversight must remain throughout the research process. These agreements are not just nominal; they are underpinned by the institutional policies of major publishers and funding bodies and are based on the global normative framework of the UNESCO Recommendation.

However, methods of disclosure, the implementation of transparency in proprietary systems, the boundaries between legitimate and illegitimate assistance to artificial intelligence, the geopolitical legitimacy of a framework rooted in Western liberal philosophy, the adequacy of current implementation mechanisms and the philosophical rationale for author exclusion remain contested. These are not marginal disagreements; they go to the heart of how research integrity is defined and protected in a world where AI is embedded in every stage of the academic process.

Specifically regarding political science: The discipline's comparative and interpretive traditions, engagement with politically sensitive topics and diverse methodologies make it

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

particularly vulnerable to the ethical challenges posed by AI, ranging from algorithmic bias distorting cross-country comparative results to simulation validity issues undermining opinion research based on large-scale language models. The lack of ethical guidelines specific to the discipline is an institutional gap that academic organisations within the discipline, such as the American Political Science Association (APSA), and leading political journals must urgently address.

References:

- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 'Out of One, Many: Using Language Models to Simulate Human Samples'. Political Analysis, 31(3), 337–51. 3 (2023): 337–51.
- Butalid, Renjie. 'About – The AI Ethics Brief'. Accessed 3 May 2026. https://brief.montrealethics.ai/about.
- Ziyu Chen, Changye Chen, Guozhao Yang et al., 'Research Integrity in the Era of Artificial Intelligence: Challenges and Responses'. Medicine, 103(27), 2024, e38811.
- COPE: Committee on Publication Ethics. 'Welcome to COPE'. 26 November 2025, https://publicationethics.org/welcome-cope.
- Exler, D., Schutera, M., Reischl, M., & Rettenberger, L. "Large Means Left: Political Bias in Large Language Models Increases with Their Number of Parameters". Arxiv:2505.04393. Preprint. 7 May 2025. https://doi.org/10.48550/arXiv.2505.04393.

- Floridi, Luciano. 'Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical'. Philosophy & Technology, 32(2), pp.185–193. https://doi.org/10.1007/s13347-019-00354-x.
- Floridi, Luciano; Cowls, Josh; Beltrametti, Monica; et al. 'AI4People: An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations'. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5. 4 (2018): 689–707. https://doi.org/10.1007/s11023-018-9482-5.
- Daniel Greene, Anna Lauren Hoffmann and Luke Stark. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. 2019. Available at: https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/2/ [Accessed 15 January 2020].
- Hagendorff, Thilo. 'The Ethics of AI Ethics: An Evaluation of Guidelines': Publisher Correction. 2020. https://psycnet.apa.org/record/2020-56191-001.
- Heikkilä, Melissa. 'AI Language Models Are Rife with Different Political Biases'. MIT Technology Review. Retrieved 5 April 2023.
- 'ICMJE | Recommendations | Preparing a Manuscript for Submission to a Medical Journal'. Accessed 3 May 2026. Available at: https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html.
- Jobin, A., Ienca, M., & Vayena, E. 'The Global Landscape of AI Ethics Guidelines'. Nature Machine Intelligence 1, no. 9 (2019): 389–99.

Received: 17/06/2025 **Accepted:** 12/02/2026 **Published:** 13/05/2026

- Lee, Ju Yoen. 'Can an Artificial Intelligence Chatbot Be the Author of a Scholarly Article?' Journal of Educational Evaluation for Health Professions, 20(2023), https://synapse.koreamed.org/articles/1516081874.
- Levy, Neil. 'Responsibility Is Not Required for Authorship'. Journal of Medical Ethics 51, no. 4 (2025): 230–232.
- Brent Mittelstadt. 'Principles Alone Cannot Guarantee Ethical AI'. Nature Machine Intelligence, 1(11), 501–507.
- Morley, Jessica; Floridi, Luciano; Kinsey, Libby; and Elhalal, Anat. 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods, and Research to Translate Principles into Practices'. In Ethics, Governance and Policies in Artificial Intelligence, Vol. 144, edited by Luciano Floridi. Philosophical Studies Series.
- Mike Perkins and Jasper Roe. 'Academic Publisher Guidelines on AI Usage: A ChatGPT-Supported Thematic Analysis'. Preprint. F1000Research, 16 January 2024.
- https://doi.org/10.12688/f1000research.142411.2
- Resnik, David B., 'Scientific Research and the Public Trust'. Science and Engineering Ethics 17, no. 3 (2011): 399–409. https://doi.org/10.1007/s11948-010-9210-x.
- Sabharwal, A., S. M. Patel, A. Carrano, M. Rotman, W. Wierson & S. C. Ekker. 'VaaS is a multi-layer hallucination reduction pipeline for AI-assisted science: Production Validation and Prospective Benchmarking'. medRxiv, 2026, 2026–03.

- Teixeira da Silva, J. A. 'The ICMJE Recommendations: Challenges in Fortifying Publishing Integrity'. Irish Journal of Medical Science (1971–), 189(4), 1179–1181. https://doi.org/10.1007/s11845-020-02227-1. 4 (2020): 1179–81. https://doi.org/10.1007/s11845-020-02227-1.
- United Nations Educational, Scientific and Cultural Organization. 'Recommendation on the Ethics of Artificial Intelligence'. United Nations Educational, 2021.
- Meredith Whittaker, Kate Crawford, Roel Dobbe et al., 'AI Now Report 2018'. AI Now Institute at New York University, New York, 2018.
- https://www.elindependiente.com/wp-content/uploads/2018/12/Informe-AI-Now.pdf