Check for updates

OPEN LETTER

Governing Artificial Intelligence in Scientific Research: The ARCHON Framework — Ethical Requirements Derived from Technical Architecture

[version 1; peer review: 3 approved with reservations]

Jesus Manuel Benitez Baleato

Departamento de Ciencia Política e Socioloxía, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, 15687, Spain

v1 **First published:** 04 Aug 2026, **6**:271
https://doi.org/10.12688/openreseurope.23574.1
Latest published: 04 Aug 2026, **6**:271
https://doi.org/10.12688/openreseurope.23574.1

Abstract

Artificial Intelligence (AI) is rapidly entering scientific research programmes on the assumption that the established ethical frameworks are enough to address failure risks. But that assumption is wrong: AI systems are structurally different from the tools that conventional research ethics was designed to govern. As illustrated by mainstream models, AI systems can be characterized by being stochastic, opaque, resource-intensive, and operate at population scale, meaning that ethical breaches can extend far beyond individual research projects, with potential consequences for public trust in democratic institutions. Taking the case of EU research as reference, this paper proposes the ARCHON framework,a set of ten requirements derived from the technical architecture of AI systems, to facilitate ethical compliance of scientific research for experts and practicioners engaging in roles such as proposal evaluators, project officers, ethics board members, or research administrators. The paper includes a mapping table intended to facilitate the application of the framework to evaluate ethical compliance, both for specific projects and general research programmes. Applied consistently, the ARCHON framework enables substantive rather than merely declarative AI ethical compliance, contributing this way for scientific research policy to continue addressing societal needs while remaining democratically accountable.

Keywords

large language models, research ethics, EU-funded research, interdisciplinary collaboration, reproducibility, data protection, environmental impact, AI governance

Open Peer Review

Approval Status ? ? ?

	1	2	3
version 1 04 Aug 2026	? view	? view	? view

1. **Robert Jungnickel**, University of Seville, Seville, Spain
2. **Valentina Colcelli**, National Research Council of Italy, Florence Research Area, Italy
3. **Rony Medaglia**, Copenhagen Business School, Frederiksberg, Denmark

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Horizon Europe gateway.



This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

Corresponding author: Jesus Manuel Benitez Baleato (jesusmanuel.benitez.baleato@usc.es)

Author roles: Benitez Baleato JM: Conceptualization, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under Marie Skłodowska-Curie grant agreement 101073351.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Benitez Baleato JM. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Benitez Baleato JM. **Governing Artificial Intelligence in Scientific Research: The ARCHON Framework — Ethical Requirements Derived from Technical Architecture [version 1; peer review: 3 approved with reservations]** Open Research Europe 2026, **6**:271 https://doi.org/10.12688/openreseurope.23574.1

First published: 04 Aug 2026, **6**:271 https://doi.org/10.12688/openreseurope.23574.1

The problem

Frameworks designed to facilitate ethical compliance for research are continuously outpaced, and the reason is not just that AI technology moves too fast for regulators to keep up: the reason is that AI models are **structurally different** from the instruments that conventional research ethics was designed to govern. A governance architecture that was built for tractable instruments cannot straightforwardly adapt to instruments whose behaviour is, by design, intractable. Hence, the need for AI ethical frameworks to be elicited by looking at the very architecture of AI in the first place, accepting that artificial intelligence is both a tool and a research object of a fundamentally different nature.

Four architectural properties of AI

Four properties set AI models apart from conventional research instruments.

Stochasticity. The relationship between input and output is mediated by billions of learned parameters whose response cannot be decomposed into an interpretable chain of reasoning. The model generates a statistically probable continuation, not a logical derivation. Even the most advanced interpretability techniques have not yet provided the causal account that scientific accountability requires.[4,5,6,7]

Opacity. This opacity is determined by design and amplified by implementation, because the most widely adopted models are controlled by commercial organisations whose competitive incentives conflict with the transparency that scientific validation demands.[1,3,4]

Resource intensity. Training and serving modern models consumes energy, water, and computational infrastructure, and generates carbon emissions at a scale that imposes costs on communities with no voice in the decisions that produce them.[8–10]

Population-scale operation. Individual methodological choices embedded in a single model can aggregate and multiply into collective consequences for shared societal, environmental, epistemic, and legal resources.[11]

These properties are not accidental. Most contemporary foundation models — including large language models, vision transformers, and audio models — are built on derivations of the transformer architecture introduced by Vaswani et al..[2] The transformer's core mechanism, attention, operates over sequences of tokens (whether words, image patches, or audio frames) and scales with the size of the model and data. This scaling is what produces the four properties identified above: stochasticity emerges from autoregressive sampling, opacity from the distributed nature of attention across billions of parameters, resource intensity from the computational cost of attention, and population-scale operation from the ease of deploying a single trained model across countless applications.

At the same time, the framework does not depend exclusively on the transformer architecture, which is taken as the departing point only. Alternative architectures, such as state-space or diffusion models, could share the same properties through different mechanisms, but the relevant question is whether a given AI system manifests stochasticity, opacity, resource intensity, or population-scale operation, as opposed to which specific architecture it uses. Therefore, while computational language models are taken as the departing reference, the implicatins of the elicited framework can be extended to the broader landscape of AI research as well.

In any case, it is important to emphasize that not all AI systems exhibit all four properties. Traditional models (e.g., random forests, logistic regression, small convolutional networks) may differ in that they lack stochasticity, offer greater interpretability potential, or operate at a more modest scale. However, the systems currently posing the greatest governance challenges for EU-funded research are likely to exhibit these properties collectively. This framework aims to cover such systems, as they represent the most relevant case; for simpler or fully deterministic AI systems, individual requirements may be relaxed or inapplicable, and they should be assessed consistently.

Why AI systems break conventional research ethics

Each of the properties identified above undermines an assumption that conventional research ethics take for granted. Standard frameworks assume that the failure of a scientific instrument can be observed and addressed because they are designed to be reliable and fully auditable, so their nature, together with the established disciplinary practices, allows researchers to identify failure with the required anticipation to prevent error from spreading. AI systems violate this assumption by default, because their stochasticity and opacity are not incidental flaws, but a constitutive feature by design; the possibility of failure moves from being an exception to becoming the default mode of operation, where the internal computations leading to results cannot be inspected to detect it or interpreted to solve the failures. Established

scientific integrity is, in this sense, structurally inverted, because frameworks designed for tools that fail can hardly govern tools requiring failure to work.

The case of verification mechanisms illustrates the case. Replication (also referred to as reproduction in the literature) is the ultimate standard of analytical reliability, by ensuring that the results of an analysis can be reproduced using the same tools and data described in the research. The use of AI systems fundamentally undermines replication practices on three independent grounds: stochastic models are not designed to reproduce identical outputs across runs; the concealment of the data used to generate the models and the process used to train them prevents independent reconstruction of the instrument; and, even if these were available, full re-computation would require an energy-consuming effort that would amplify the very environmental degradation and life-supporting resource scarcity depletion that any ethical framework should be limiting. Replication, therefore, turns into either an unfeasible or unaffordable practice, if acceptable at all, leaving this way scientific validity without its ultima ratio verification mechanism.

Conventional ethical frameworks also assume that research outputs can be contained, meaning that a published finding can enter a field without immediately transforming it, and with time allowing for the discipline to detect and amend any error before it leads to negative consequences. In contrast, the capacity of AI systems to operate at population scale makes that assumption obsolete, because any deployed model or training dataset can immediately become available as public infrastructure, this way reshaping the conditions under which not just subsequent research, but even ordinary life, takes place. In this scenario, no regulatory update schedule can close the identified gaps, because they do not lie between a rule and its next revision: they lie between what the technology *is* and what the governance architecture was designed *for*. What is needed is to deploy governance practice grounded in the very architecture of AI systems, in a way that is aligned with the political values that make scientific research a legitimate institution.

Three principles, ten requirements

The requirements below are organised under three principles, each grounded in the architectural properties of transformer-based foundation models identified above. The three principles supply the reasoning, and the requirements supply what governance actors must require and verify in order to ensure substantive, as opposed to merely declarative, compliance.

Interpretation — keep capable humans in the loop

Because AI models are opaque and stochastic, their outputs require human judgement to interpret: that judgement must be competent in the domain of application.

1. Require involvement of domain expertise from the design stage

The validity of an AI system can only be assessed in the domain of the problem that it was designed to address, regardless of its computational performance. This requires both domain and technical competence to be present in the research project from the very conceptual design stage, as opposed to just being added as an auxiliary or post-hoc review at the end. The case of LLMs exemplifies how AI models generate statistically probable continuations rather than domain-valid inferences and cannot assess the validity of their own outputs — but neither can a reviewer lacking the domain knowledge to do so autonomously. Thus, human oversight is a necessary but not sufficient condition, because the human in the loop must be competent enough, because a non-expert reviewer could not distinguish a fluent hallucination from a correct finding — meaning that formal human involvement without domain competence offers procedural compliance without substantive oversight. But competence alone is not enough if it arrives too late. The framing of a research problem is already a theoretical task: how categories are defined, what counts as a relevant phenomenon, or what the research question assumes about the problem — all of these imply theoretical commitments that can only be reliably assessed by the corresponding discipline, and no amount of output-checking can compensate for the flaws of a misframed concept or research question. Domain competent researchers must take part in the validation of the outputs, but more importantly they must be engaged from the research design stage, contributing to shaping the conceptual architecture of the research from the very outset, including the formulation of categories, the design of gold-standard datasets, the definition of annotation processes and evaluation criteria, and the standards used to determine whether results are valid.

2. Require authorship agreements before writing begins

Written authorship agreements mapping individual contributions should be established before any research is drafted. The referred need to ensure domain expertise leads naturally to interdisciplinary collaboration; however academic fields can differ in so pivotal aspects such as publication norms, citation conventions, and conceptions of what counts as intellectual work. Consequently, what one discipline treats as a core intellectual contribution another may regard as just

auxiliary or technical assistance. Therefore, the requirement for interdisciplinarity creates a specific risk: when domain and technical competences are distributed across the collaborators of a project, disagreements over the recognition of their contributions can derail the research itself, because unresolved claims can damage working relationships, delay or prevent publication, reducing engagement incentives and impose real career costs on collaborators whose contributions are undercounted by the norms of a discipline not their own. In AI research, this risk is notably high due to the conventional association of AI with computer engineering, especially in the case of collaborations involving social sciences and humanities, where academic excellence is measured by criteria that can directly conflict with engineering norms — journal publications over conference proceedings, collective scholarly conversation over individual technical contribution or interpretive depth over quantitative output. Reaching an enforceable agreement on publications before research is deployed and writing begins makes those differences explicit, allowing to address them, protecting both the research and the people doing it; using a system such as the CRediT taxonomy can help with this regard. On a side note, it is increasingly accepted the idea that AI models cannot be listed as authors, as they lack the legal personhood and moral agency required to qualify for authorship[12] and auxiliary AI assistance such as grammar checking, spell correction or style suggestions — does not generally require acknowledgement, but substantial content generation must be explicitly recognized as relevant.

3. Require validation of outputs against human-coded samples

When AI is used as a measurement instrument its validity should be established against well-documented human-coded samples. This is particularly relevant in the case of research on human phenomena, where stratifications of the populations under study should be considered to ensure the instrument measures what it claims across the variation present in the data. For example, language models trained predominantly on English-language, Western text produce systematically distorted measurements when applied to classify attitudes across underrepresented populations; at scale, that distortion propagates through every finding built on it.[4,13] Validation must occur against ecologically representative samples, not only against curated datasets where the phenomenon is overrepresented, and the design of the stratification and the assessment of what counts as a correct classification are themselves tasks that require the domain competence this principle demands. This concern is amplified when the output of the research is intended for use in training or fine-tuning AI models, such as a tagged corpus. A dataset that has been produced by an insufficiently validated classifier inherits any existing biases, and propagates them into every subsequently trained model — potentially at population scale, and with consequences capable of persisting over time, long after the original research project has concluded. In such cases, the validation against stratified human-coded samples is not just merely a methodological safeguard, but also a precondition for scientific validity and responsible publication, and should take into account downstream risk assessment (see requirement 10).

Validity — verify, document, replicate

Since AI models are opaque, stochastic, and generally embedded in commercial systems, researchers must ensure what the model cannot guarantee by itself: reproducibility, verifiability, and robustness.

4. Require explicit computational reliability

AI research should be conducted on infrastructures that are capable of providing sufficient safeguards to ensure reproducibility, data sovereignty, and independence from commercial or foreign decisions. And this requirement must be observed at two differentiated levels. Firstly, at the hardware level, institutional infrastructures should be preferred over private commercial ones because EU law does not require the latter to fully disclose the required information, such as energy consumption or hardware characteristics, which are necessary to enable reproducibility and environmental reporting. Public infrastructures, such as EU High Performance Computing (HPC) facilities, do not carry these limitations and are open for the scientific community. At the software level, copyleft open-source and open standards should be required as possible in order to overcome unnecessary or undocumented dependencies that cannot be fully inspected or replicated. For AI systems specifically, open-weight models (those whose trained weights are publicly available) deployed on local institutional infrastructure should represent the option by default, with alternative options being acceptable only as appropriately justified. The use of commercial black-box systems can compound all these risks simultaneously because market operators can change model behaviour and availability unilaterally and without notice. A model that changes silently between access dates produces results that are not replicable by definition, therefore hampering the peer review process that constitutes the foundation of contemporary science. In any case, using infrastructures that require sending data outside of the EU jurisdiction should be actively avoided as that may constitute unlawful transfer that no contractual safeguard can prevent, as exemplified by the use of US infrastructure under FISA Section 702.[14]

5. Require documentation of all relevant model parameters

All the information determining model outputs must be recorded and made available, including model version identifiers, access dates, hyperparameters, random seeds, and any preprocessing or postprocessing steps that have been applied to inputs or outputs. Consistently, fully open models where not just weights, but also training data (or verifiable provenance) and training pipeline are publicly available should be preferred over models releasing only the weights, without providing the training data, or disclosing the training pipeline, because the latter introduce undocumented factors that cannot be fully captured in supplementary materials regardless of how carefully the parameters have been recorded. For generative models (e.g. large language models or image generation models based on diffusion) documentation must include all prompts, system messages, or sampling parameters. As noted, many AI systems are stochastic, implying that identical inputs may produce different outputs, contrary to the expectations of conventional research tools; this issue compounds with the fact that models accessible only via APIs, as the case of most commercial services, may change without notice across versions, so identical parameters could lead to non identical results at a later date, exemplifying the need to record the exact model version. For analytical tasks that require reproducibility, temperature and equivalent stochastic parameters should be set to zero or near-zero, and deterministic execution must be enabled if the inference framework supports it. Otherwise, non-determinism should be reported as a methodological limitation, including the range of variation observed across repeated runs. Where available, APIs called from reproducible code should be used rather than chats or similar web interfaces, because those can be expected to be affected by additional uncontrolled variability preventing full documentation of parameters.

6. Require verification of outputs against domain ground truth

No AI output should enter the research pipeline, or be published, without having been verified against an appropriate domain validated ground truth. The reason is that AI models can be architecturally indifferent to truth: as exemplified by LLMs, their outputs represent statistical continuations, not the product of logical inference from premises to conclusions supported by empirical evidence.[15,16] This indifference to truth is not a bug but a structural property, and it holds regardless of how fluent, confident, or plausible the output can appear. For discriminative models, whether deterministic or probabilistic, verification can be applied validating with test sets stratified by population, as specified in requirement 3. For generative models, such as large language models, diffusion-based image generators, or code generation models, which exhibit the capacity to produce outputs that are so fluent and plausible that they are false,[17] verification must be claim-level and exhaustive: code against executable tests, statistical claims against raw data, citations against DOI lookup or Scopus, and generated text against the sources purported to be summarized. No category of output should be exempted from this step, and no output type should be assumed to be self-verifying on the basis of computational performance or surface plausibility, consistently with the interpretability requirements. In any case, output verification is not just a condition for scientific integrity, but also for academic publication, because an output that has not been verified against domain ground truth is not a research finding, but rather an unexamined model output, and publishing it as the former is a misrepresentation of the technical value of the work, in addition to a downstream risk (see requirement 10).

7. Require testing for adversarial manipulation

Research pipelines analysing content that could have been designed to influence the analyser, must be tested against adversarial inputs, and residual risks reported as methodological limitations. For example, language models are vulnerable to prompt injection, a technique where the analysed content is modified to contain pseudo-instructions capable of overriding the researcher's prompt[18]; in models trained on image or multimodal data, systematic misclassification can be produced by perturbations imperceptible to human reviewers. These issues can happen not just at execution time, but also at the training stage, seeding training or fine-tuning corpora with content designed to 'poison' data in order to shift the induced associations,[19] a phenomenon particularly relevant for data scraped from the web or supplied by the users. The management of adversarial risks operates preventively and also at execution time. For example, in the case of prompt injection, preventive measures may include input sanitization, the segregation of analysed content from model instructions, and evaluation in sandboxed environments; and for data poisoning, a preference should be established for curated and vetted corpora with verifiable provenance over data scraped from the web or supplied by users; and requirement should be made to audit the data used for training in any case. For detection of both attacks, research should be required to implement cross-validation on adversarial samples and human review of random outputs. These sensitivity analyses should be documented and openly reported, and any residual adversarial risks that cannot be mitigated must be reported as well as part of the study's methodological limitations.

Commons — own the costs and the consequences

Since AI models operate at population scale and consume shared environmental, epistemic, and legal resources, the implication is that individual research choices aggregate into collective consequences that extend beyond the laboratory.

8. Require reporting of environmental costs

Computational research using AI models must minimise and justify its environmental costs, requiring reporting of energy, water and carbon footprint at the minimum to make this requirement verifiable over time. The scalability feature implies that AI models operation imply vast consumption of life-supporting resources such as energy, water, or land, and generates carbon emissions, therefore imposing costs on communities that may have no voice in that process.[20] Evaluating the environmental impact requires data and analysis which need to be explicitly required because this does not take part of the conventional reporting in scientific disciplines. Before initiating any computation involving AI systems, researchers should be required to evaluate whether the question genuinely requires computation at that scale, and to report at least the implied energy consumption, carbon footprint, and water usage alongside other technical parameters — typically in the Methods section. Indeed, established reporting schemes and estimation tools exist for this purpose: CodeCarbon and MLCO2 provide approximations for locally run models, and standardised reporting schemas should be applied where available.[20] The preference for public HPC infrastructures required before is directly aligned with this requirement, because it enables the environmental accountability structurally prevented by commercial platforms, an issue which gets amplified by the fact that EU law does not require them to fully disclose energy consumption or hardware characteristics, making it unfeasible to report environmental costs accurately when using commercial cloud services. Consistently, the use of infrastructures that do not allow to estimate the environmental impact of the required computing would be incompatible with environmentally aware scientific research.

9. Ensure compliance and safeguards before any personal data processing begins

The use of personal data must be explicitly justified and authorized before any processing begins, minimising and safeguarding any necessary processing, with research found to risk privacy violation to be put on hold until ensuring full compliance, or retracted otherwise. The use of personal data introduces high stake risks because it introduces a direct threat to the capacity of individuals to decide autonomously which is required by democratic systems to work properly. In order to manage those risks, research involving personal data requires compliance with institutional and technical safeguards that significantly increase the costs up to questioning the feasibility of the research. Therefore, before resorting to personal data, an evaluation must be made on whether the research question can be answered using non-personal alternatives — domain knowledge should inform what substitutes are viable. Where personal data is unavoidable, data minimisation is required under GDPR Article 5(1)(c)[22], a Data Protection Impact Assessment must be conducted, authorization by the corresponding ethical board should be obtained before processing begins, and monitoring should be maintained until the end of the project. For example, research managers should ensure that the requests submitted to the ethical boards have reliably characterized the sensitive nature of the data, and that institutional and technical safeguards have been put in place, including on digital security and re-identification prevention. With this regard, it is important to emphasize that conventional anonymisation is insufficient, with LLMs amplifying this risk through their capacities to link publicly available sources[21]; that personal data that is publicly available remains personal data; and that compliance with intellectual property law does not imply compliance with data protection law. Transfers of personal data to non-EU jurisdictions are particularly problematic, since they may be considered unlawful under conditions no contractual safeguard can remedy,[14] reinforcing the convenience of the use of computing infrastructures located in the EU.

10. Require downstream risk assessment before publication of classifiers, datasets or models

A documented risk assessment for downstream applications must be completed before publishing any classifier, model, or dataset in order to ensure that any identified risks have been adequately mitigated. As noted, AI models do operate as infrastructure at population scale, and this implies that any released research output (e.g. a dataset or a classifier) can be adopted by diverse actors from commercial platforms to state agencies for purposes other than those intended by the research project, and which cannot be controlled after the research outcome has been released.[11] This dual-use risk is particularly important because the same classifier that has been modeled to support legitimate applications, such as automatic content moderation to support democratic deliberation online, can also serve antagonistic purposes such as illegitimate surveillance, censorship, or digital discrimination. Therefore, research projects must assess any plausible adoption scenarios for their shared data, explicitly addressing any connected risks such as error distribution across populations leading to discrimination, use in authoritarian contexts, or adversarial vulnerabilities at deployment stage. If misuse can be foreseen, licensing restrictions prohibiting surveillance applications should be established, as illustrated by the RAIL licence family templates. If substantial risks cannot be adequately mitigated through technical, legal, or

Table 1. Mapping the ten ARCHON requirements to project-level verification and programme-level architecture.

Requirement	Project-level verification	Programme-level requirement
1. Domain expertise	Domain and technical competence present from design stage	Interdisciplinary authorship framework as condition of project initiation
2. Authorship agreements	CRediT-based written agreement before any paper is drafted	Scientific deliverables circulated to all disciplinary components
3. Output validation	Validation against human-coded samples stratified by population	Effective cross-disciplinary resourcing
4. Computational environment	Open-weight models on EU HPC as default; proprietary justified	Institutional HPC infrastructure and support
5. Reproducibility	All prompts, versions, access dates recorded	Reproducibility documentation as deliverable condition
6. Output verification	Every output verified against domain ground truth	Model documentation in Data Management Plans
7. Adversarial testing	Cross-validation on samples; risks reported as limitations	— (primarily project-officer-level concern)
8. Environmental costs	Energy, carbon, water reported in Methods section	Environmental reporting in deliverables and annual reports
9. Data protection	Written authorisation before processing begins	DPIAs mandated; enforceable authorisation requirement
10. Downstream risks	Impact assessment and RAIL licensing before publication	Institutional review required before publication

licensing measures, then the publication of the research outcomes should be deferred before institutional review is performed, with restricted deployment under direct institutional oversight being potentially more appropriate than open release in those cases.

Mapping project-level and programme-level governance

The ten requirements that have been proposed have been tailored so they can be applied by participants with responsibilities in EU research activities. We illustrate the utility of the framework showing how it can operate at the two extremes of the research cycle: from the programme level, where funding calls are framed, to the project level, where the research is implemented. At **project level**, they specify what should be verified when monitoring a funded research project. At **programme level**, they specify what the funding architecture itself should build in, through evaluation criteria such as deliverable requirements, infrastructure provision, or enforceability conditions. Table 1 maps each of the ten requirements to its expression at both levels. In all cases, compliance with the proposed requirements should be ensured not just at the nominal level, but by checking that sufficient measures have been established to ensure that the requirements can be effectively enforced. For example, commitment to comply with data protection should include practical and measurable activities, such as workshops to familiarise researchers with the regulation, milestones to check that all ethical authorisations have been submitted, and time allocation to check that the authorisation requests provide sufficiently accurate characterisations of the sensitive nature of the data.

Conclusion

AI is already entering EU research faster than the ethical frameworks designed to govern it can adapt. This is especially true for mainstream AI models, characterised by stochasticity, opacity, resource intensity, and population-scale operation —four properties that conventional ethics was never designed to handle. The gap is not a matter of legal frameworks lagging behind fast technical pace, but a structural mismatch between what conventional ethical compliance was built to handle and the very nature of AI systems. Closing that gap requires governance practices that are grounded in the architecture of AI itself while being aligned, at the same time, with the political values at stake.

The ARCHON framework -named after the Greek ἄρχειν to denote the idea of governing AI from within the system, rather than imposing rules from the outside- presented in this paper contributes to bridging that gap by grounding ethical requirements in the technical architecture of the AI systems themselves. Rather than treating compliance as an external constraint, the ARCHON framework looks at ethical governance as an architectural necessity: what must be governed follows directly from how the AI systems work. The proposed set of ten requirements operationalise that approach by

translating the architectural risks into enforceable conditions across the full research lifecycle—proposal evaluation, ongoing monitoring, and post-termination review.

Because AI operates at population scale, ethical failures in research programmes affect entire societies — and with them, public trust in European institutions. EU practitioners (evaluators, officers, ethics board members, administrators) are uniquely placed to prevent such failures: not by multiplying rules, but by requiring governance that matches what AI actually is. A mapping table is provided to facilitate the application of this framework to any EU funded research project working with AI, allowing to quickly identify what should be verified at project level, and what the funding architecture should build in at programme level, so that project-level verification can succeed.

Applied consistently, ARCHON can help EU research to remain both democratically accountable and scientifically innovative. However, the framework is not a panacea: its effectiveness depends on institutional uptake, and future work should test it against emerging AI architectures, such as agentic systems or alternative foundation model architectures. Nonetheless, by grounding governance in technical reality rather than external compliance checklists, ARCHON offers a pathway to strengthen — rather than undermine — the democratic foundations that sustain EU research policy.

Ethics and Consent

Ethical approval and consent were not required for this study. This article presents a policy analysis and does not involve human participants, personal data, or experimental research.

Data Availability

This is a policy analysis and opinion piece; it does not report original empirical findings and therefore involves no dataset generation or analysis.

Acknowledgements

The authors thank the Political Analysis research group (EIP) for their valuable feedback and intellectual support, and also to the Social Data Lab (https://socialdatalab.usc.gal/) for the computational infrastructure, both at the University of Santiago de Compostela. All content presented here represents original research by the authors, and any relevant use of computational methods, including language models, is explicitly described in the relevant section or the correspondent supplementary material, as relevant.

References

1. Burrell J: **How the machine "thinks": Understanding opacity in machine learning algorithms.** *Big Data Soc.* 2016; **3**(1): 1–12.
2. Vaswani A, Shazeer N, Parmar N, *et al.*: **Attention is all you need.** Guyon I, Von Luxburg U, Bengio S, *et al.*, editors. *Advances in Neural Information Processing Systems.* Curran Associates, Inc; 2017; Vol. **30**: pp. 5998–6008.
3. Lipton ZC: **The mythos of model interpretability.** *Queue.* 2018; **16**(3): 31–57.
Publisher Full Text
4. Bender EM, Gebru T, McMillan-Major A, *et al.*: **On the dangers of stochastic parrots: Can language models be too big?** *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'21).* ACM; 2021; pp. 610–623.
5. Elhage N, Nanda N, Olsson C, *et al.*: **A mathematical framework for transformer circuits.** *Transformer Circuits Thread, Anthropic.* 2021.
6. McCoy RT, Yao S, Friedman D, *et al.*: **Embers of autoregression: Understanding large language models through the problem they are trained to solve.** *arXiv:2309.13638.* 2023.
7. Rudin C: **Stop explaining black box machine learning models for high stakes decisions.** *Nat. Mach. Intell.* 2019; **1**: 206–215.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
8. Strubell E, Ganesh A, McCallum A: **Energy and policy considerations for deep learning in NLP.** *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 2019; pp. 3645–3650.
9. Patterson D, Gonzalez J, Le Q, *et al.*: **Carbon emissions and large neural network training.** *arXiv:2104.10350.* 2021.
10. Li P, Yang J, Islam MA, *et al.*: **Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models.** *arXiv: 2304.03271.* 2023.
11. Bommasani R, Hudson DA, Adeli E, *et al.*: **On the opportunities and risks of foundation models.** *arXiv:2108.07258.* 2021.
12. Thorp HH: **ChatGPT is fun, but not an author.** *Science.* 2023; **379**(6630): 313.
Publisher Full Text
13. Shah DS, Schwartz HA, Hovy D: **Predictive biases in natural language processing models: A conceptual framework and overview.** *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics; 2020; pp. 5248–5264.
14. Court of Justice of the European Union: ***Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems* (Case C-311/18).** *Judgment of 16 July 2020. ECLI:EU:C:2020: 559.* 2020.
15. Hicks MT, Humphries J, Slater J: **ChatGPT is bullshit.** *Ethics Inf. Technol.* 2024; **26**: 38.
Publisher Full Text
16. Frankfurt HG: *On Bullshit.* Princeton University Press; 2005.
Publisher Full Text
17. Alkaissi H, McFarlane SI: **Artificial hallucinations in ChatGPT: Implications in scientific writing.** *Cureus.* 2023; **15**(2): e35179.

18. Perez F, Ribeiro I: **Ignore previous prompt: Attack techniques for language models.** *arXiv:2211.09527.* 2022.

19. Wallace E, Zhao T, Feng S, *et al.*: **Concealed data poisoning attacks on NLP models.** *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* ACL; 2021; pp. 139–150.

20. Luccioni AS, Jernite Y, Strubell E: **Power hungry processing: Watts driving the cost of AI deployment?** *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT'24).* ACM; 2024; pp. 85–99.

21. Lermen S, Paleka D, Swanson J, *et al.*: **Large-scale online deanonymization with LLMs.** *arXiv preprint arXiv:2602.16800.* 2026.

22. European Parliament and Council: **Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation).** *Off. J. Eur. Union.* 2016; **L 119**: 1–88.

Open Peer Review

Current Peer Review Status: ? ? ?

Version 1

Reviewer Report 04 September 2026

https://doi.org/10.21956/openreseurope.25521.r78874

© **2026 Medaglia R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Rony Medaglia
Copenhagen Business School, Frederiksberg, Denmark

The Open Letter proposes ARCHON, a framework of ten requirements for governing AI in scientific research, with particular relevance to EU-funded research. It argues that conventional research ethics may not fully address AI systems characterised by stochasticity, opacity, resource intensity and population-scale operation. These characteristics are translated into requirements covering domain expertise, authorship, validation, computational infrastructure, reproducibility, verification, adversarial testing, environmental reporting, data protection and downstream risk assessment. The project-level and programme-level mapping is particularly valuable, as it helps translate the framework into practical governance measures.

The rationale is clear and provides a useful basis for the proposed framework, although it could be strengthened further. The manuscript explains the challenges it seeks to address and why AI may require specific governance mechanisms. However, a fuller discussion of the limitations of the existing EU governance framework would help situate ARCHON more clearly. In particular, engagement with the EU AI Act would be valuable in clarifying whether and how ARCHON complements, extends or overlaps with existing requirements.

Differing perspectives are represented to some extent, but the manuscript could engage more fully with alternative governance approaches and established practices. Doing so would help clarify the framework's distinctive contribution. The relationship between the four architectural properties and the ten requirements could also be explained more explicitly, particularly given the article's acknowledgement that not all AI systems possess all four properties.

The language is generally accessible, and specialist concepts are sufficiently explained.

The recommendations are useful and provide a promising basis for implementation, although some aspects would benefit from further specification. The mapping table provides a strong foundation, while requirements such as exhaustive output verification, environmental reporting, default use of open-weight models and institutional review could be strengthened by clearer applicability criteria, exceptions, responsible actors and proportionality considerations. Further

attention to feasibility for smaller research groups, alongside an explicit mapping of ARCHON against existing EU legal and ethical requirements, would substantially strengthen the letter.

Is the rationale for the Open Letter provided in sufficient detail? (Please consider whether existing challenges in the field are outlined clearly and whether the purpose of the letter is explained)
Partly

Does the article adequately reference differing views and opinions?
Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
Yes

Is the Open Letter written in accessible language? (Please consider whether all subject-specific terms, concepts and abbreviations are explained)
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow? (Please consider whether others in the research community would be able to implement guidelines or recommendations and/or constructively engage in the debate)
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital government

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 22 August 2026

https://doi.org/10.21956/openreseurope.25521.r78866

© 2026 Colcelli V. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Valentina Colcelli
National Research Council of Italy, Florence Research Area, Italy

The issues raised are indeed interesting and relevant. Addressing the growing problem of using generative AI in research activities is very useful.

The technical picture is well linked to the doctrine related to interpretability and ethics in AI (Bender et al., 2021, 'Stochastic Parrots'; Rudin, 2019; Lipton, 2018; Bommasani et al., 2021).

However, the article has a gap in that one of the most important EC policy documents on the main topic is not only not cited, but also not integrated into the table designed to evaluate the governance of the ethical requirements for AI in scientific research. It is important to evaluate the cited document to strengthen the ARCHON. If the author decides not to integrate this EC policy, they must explain why.

Also, the Open Letter does not cite the AI Act, which established specific transparency rules for general-purpose AI (GPAI). Without these specific choices being made clear by the author, the audience cannot understand whether ARCHON adds value to the EU legal and ethical framework or simply restates the same position in different terms.

Another point that the Open Letter should emphasise is that the four properties that can negatively impact research activities from an ethical point of view are addressed very well.

On the other hand, the Open Letter states that requirements 9 (personal data) and 10 (downstream risk assessment before publication of classifiers, datasets or models) are principally based on the population, but these are also applicable to other non-stochastic and non-opaque technologies.

Please take this aspect into consideration as well. There is also a strong recommendation to use open-weight models such as HPC UE as the default option, taking into account the infrastructure costs for small research groups.

Is the rationale for the Open Letter provided in sufficient detail? (Please consider whether existing challenges in the field are outlined clearly and whether the purpose of the letter is explained)

Partly

Does the article adequately reference differing views and opinions?

Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

No

Is the Open Letter written in accessible language? (Please consider whether all subject-specific terms, concepts and abbreviations are explained)

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow? (Please consider whether others in the research community would be able to implement guidelines or recommendations and/or constructively engage in the debate)

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: law and etich

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 21 August 2026

https://doi.org/10.21956/openreseurope.25521.r78875

© 2026 Jungnickel R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Robert Jungnickel
Department of Medical Physiology and Biophysics, Neuroimaging and Brain Networks Group, University of Seville, Seville, Andalusia, Spain

Summary
This article proposes the ARCHON framework, which includes ten ethical requirements for the governance of artificial intelligence in EU-funded scientific research. It is a valuable contribution because it considers ethical requirements in the technological design of AI systems rather than viewing compliance as an external constraint.

I think there are **two main points** in which the article could be improved. These relate to the section entitled "**Improvements for different views and opinions**".

1. Documentation requirements do not include common practices
Requirement 5 of the article only considers technical parameters, such as model version identifiers and hyperparameters. However, this approach is incomplete because it does not mention established practices regarding model behaviour. For example, it does not refer to model cards, which include hyperparameters, or data cards and benchmarks for scientific validity. It should also detail any known failure modes or inappropriate use cases.

2. Consider the European regulatory framework, such as the AI Act
The article suggests a governance framework for AI in EU-funded research. However, it does not consider the EU AI Act that governs AI systems within the EU. This omission is problematic for three reasons.

1. Compliance with the AI Act will be mandatory for many AI systems used in research contexts. A governance framework that ignores mandatory regulations cannot provide practical guidance to researchers or administrators.
2. The AI Act already establishes risk-based requirements for AI systems, including transparency obligations, technical documentation requirements and provisions for high-risk systems. Therefore, the ARCHON framework should clarify how its ten requirements

relate to these obligations.

3. The AI Act contains two provisions (Articles 2(6) and 2(8)) that exempt certain AI research activities from its scope. The ARCHON framework should address these in order to be useful in research practice. For instance, neither of these exemptions removes the need for ethical governance. In fact, they may create a gap that the ARCHON framework could help to address. This could help to ensure that EU-funded research adheres to the appropriate standards, regardless of any regulatory exemptions. Furthermore, compliance with ARCHON requirements could be included in funding agreements as a condition of EU research grants, effectively making them binding.

Is the rationale for the Open Letter provided in sufficient detail? (Please consider whether existing challenges in the field are outlined clearly and whether the purpose of the letter is explained)

Yes

Does the article adequately reference differing views and opinions?

Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language? (Please consider whether all subject-specific terms, concepts and abbreviations are explained)

Yes

Where applicable, are recommendations and next steps explained clearly for others to follow? (Please consider whether others in the research community would be able to implement guidelines or recommendations and/or constructively engage in the debate)

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial intelligence in scientific research, EU Artificial Intelligence Act (AI Act), Digital Services Act, Human-Computer Interaction, Human-Centric Machine Learning, Graph Representation Learning, Complex Networks, Visual Network Analysis, Transformer, Self-Attention, Neuroscience, Neuroimaging

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.