Critical concerns for using LLMs in the (computational) humanities and beyond

Sarah Lang

1. Introduction

When using large language models (LLMs) and other AI tools in their research, humanities scholars must engage with the ethical concerns these systems raise. Yet standard AI ethics remain largely shaped by legal and technical frameworks that prioritise compliance and explainability. These approaches fail to account for the complexities of working with historical data. This contribution outlines key ethical and epistemic concerns surrounding the use of LLMs within computational humanities and related domains, with a particular focus on the implications for historical and humanities data.[1] While the field of AI ethics has become increasingly preoccupied with frameworks such as explainable AI (XAI), this paper draws attention to the limitations of such approaches when adopted in isolation, and it foregrounds the importance of critical, humanities-informed methodologies, including feminist praxis, activist scholarship, and data-centric critique attentive to labour and power.

Explainable AI has gained traction as a technical means of addressing opacity in algorithmic systems.[2] Within the context of AI ethics (Liao, 2020b), its popularity is understandable: providing a post hoc explanation of a model's outputs often appears to enhance transparency and accountability. However, scholars working from critical theory, feminist technology studies, and the digital humanities have questioned the efficacy of such explanations. These critiques argue that XAI can function more as a technical patch than a substantive response to the systemic issues that underlie model behaviour (Rudin, 2019; Klein and D'Ignazio, 2024). Explanations may create the impression of positive change, thereby allowing development to continue without addressing the more fundamental problems embedded in the data or model architecture. Nevertheless, this contri-

1 In this edited collection, there are a number of papers which go in somewhat similar directions. For instance, Meding and Daugs (2026) discuss the reliability versus added value of LLMs for historians. Khutsishvili (2026) is a short paper on epistemic authority. The introduction by the volume editors gives an excellent overview of related issues (cf. Simons et al., 2026).

2 See contribution by Oliver Eberle (2026).

bution does not reject explainability altogether. When used critically, XAI may serve as a valuable diagnostic tool, capable of exposing underlying issues that demand structural correction. The emphasis here is not on discarding technical solutions, but rather on reframing their use within a broader critical practice attentive to root causes rather than surface symptoms.

Given the limitations of explainability, there is growing scholarly attention to training data as the root site of many ethical problems. The flawed nature of most available datasets (Paullada et al., 2021), particularly those involving historical or cultural material, creates challenges for those working in the computational humanities (Lang and Suárez Cronauer, 2026). Concerns range from unclear provenance and embedded power relations to epistemic opacity and skewed representations. These are especially salient in domains like the history, philosophy, and sociology of science (HPSS), where the interpretation and contextualisation of source material are central to scholarly practice.

Other pressing concerns addressed in this paper include the redistribution and devaluation of labour (particularly in the form of ghost work, Gray and Suri, 2019), challenges to research integrity such as hallucinated citations and AI-generated misinformation, and the environmental cost of large-scale model training. These issues are contextualised through references to relevant scholarship, such as Kate Crawford's critique of AI's material footprint (Crawford, 2021) or recent work on "open-washing" LLMs, where transparency is often claimed but rarely truly delivered (Liesenfeld and Dingemanse, 2024). The paper also discusses how methodologies from fields such as critical archival studies, critical data studies, and the critical digital humanities can offer tools such as auditing (Vecchione et al., 2021), which is a particularly effective form of intervention for analyzing and curating datasets. These approaches can help articulate the epistemological stakes of working with LLMs and offer tools to mitigate their ethical shortcomings.

When examining the use of large language models from a critical and ethical standpoint within the history, philosophy, and sociology of science, a number of distinct issues arise. This overview serves as an initial orientation for those wishing to engage with these concerns, whether to reflect more critically on their own use of LLMs or to locate key literature and frameworks. While some suggestions for mitigating specific risks are included, the format does not allow for a comprehensive treatment of each topic.[3]

Importantly, this is not a call for refusal, though refusal remains a valid stance within some feminist and activist circles when technological systems reproduce harm or fail to meet the ethical standards of the communities they purport to serve. Rather, the aim here is to offer a constructive intervention that bridges critical perspectives with technical practice. By doing so, the article addresses an audience that includes technically proficient researchers in the history, philosophy or sociology of science and knowledge engaging in LLM-assisted scholarship or computational humanists who may not be trained in ethics but are seeking accessible and informed frameworks for evaluating and improving their work.

3 More in-depth analyses are currently in preparation by the author, and readers are encouraged to consult forthcoming work for more details: Lang and Suárez Cronauer (2026; forthcoming-2026) and Lang et al. (2025).

The conclusion offers a set of future-oriented reflections. It suggests that fields such as the digital humanities, history of science, and science and technology studies can provide essential contributions to the ethical development of AI, not least through their longstanding engagement with questions of knowledge production, power, and historical context. In particular, it advocates for participatory models of development grounded in solidarity, the feminist ethics of care (Gray and Witt, 2021), and the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics: GIDA, 2021). These frameworks provide both a conceptual and practical foundation for guiding the use of LLMs in a manner that is responsive to the specific needs and values of the humanities. While perfection may remain impossible to achieve in the end, meaningful ethical progress is still both necessary and possible. However, ethical engagement with LLMs in the context of HPSS and the digital humanities must move beyond mere regulatory compliance and technical mitigation. It requires an expanded ethical toolkit, attentive to the specific epistemic, historical, and social stakes of the field. This contribution offers a preliminary guide to that endeavour.

2. AI ethics and the issue of explanation

The ethics of LLMs, particularly as applied in digital humanities and HPSS contexts, remains a developing field.[4] This is not to imply that ethical questions have gone unaddressed; rather, the available literature often appears either overly specialised or not immediately applicable. Much work on AI ethics exists in adjacent disciplines, such as law and medicine where long-standing traditions and regulatory structures have shaped the debate. In these fields, concerns about privacy, personal data, safety, and legal compliance dominate the discourse (e.g., Hildebrandt, 2021). The EU AI Act and similar guidelines are frequently cited as important milestones. However, such documents are primarily oriented towards preventing legal violations or safeguarding individuals from harm, rather than offering guidance for the responsible use of AI in humanities contexts. Their focus, in short, tends to be on preventing illegal wrongdoing, rather than promoting ethical engagement.

Broadly speaking, three disciplinary traditions can be identified within AI ethics: one rooted in computer science (Kuipers, 2020), another in philosophy (Gunkel, 2020), and a third informed by intersectional feminism, critical theory, and activist scholarship (e.g., O'Neil and Gunn, 2020; Gray and Witt, 2021), including areas such as critical digital humanities, critical code and data studies, and archival theory. Each of these traditions brings distinct assumptions and goals. For those unfamiliar with their respective vocabularies and commitments, the differences can be challenging to navigate.

From a computer science perspective, explainable AI (XAI) has emerged as a prominent response to concerns about algorithmic opacity. The aim is to render the internal

4 That is not to say that there is no literature on AI ethics. This would be far from the truth, as the literature is abundant. However, AI ethics is by no means straightforward, and there is even an ethics of AI ethics (Powers and Ganascia, 2020). For an introduction, the following handbooks are an excellent starting point: Dubber et al., 2020; Liao, 2020a; Floridi, 2023.

workings of models more transparent, thereby fostering trust and enabling oversight. Yet such technical solutions are often criticised within critical humanities approaches for failing to address the structural origins of harm (D'Ignazio and Klein, 2020; Klein and D'Ignazio, 2024). As scholars informed by feminist praxis and critical theory have argued, explanation alone may provide a sense of closure or accountability without actually disrupting the logics that produce bias or injustice. In this sense, explainability risks becoming a superficial remedy – a "band-aid" that leaves the core problems intact. In addition, scholars have questioned whether transparency, which is often heralded as an ideal in the context of black box algorithms, is actually the ideal solution many seem to take it to be (Ananny and Crawford, 2016). Nevertheless, explanation remains a necessary, if insufficient, component of ethical AI. It is especially vital for philosophical approaches to machine ethics, which require insight into the reasoning processes of AI systems in order to evaluate moral acts. In the digital humanities, ethical discussions have been less institutionalised and to this day often remain at a surface level, engaging solely with legal issues such as access or privacy. As Berry notes, the field has yet to establish a professional ethics (Berry, 2022a) that could address its specific entanglements with AI and computational infrastructures. This gap is particularly evident given that we use historical data with major concerns about provenance, representation and archival bias.

Humanities scholars dealing with historical datasets simply have a different focus than other disciplines when it comes to ethics (Lang and Suárez Cronauer, 2026). What they seek is not a compliance framework, but a deeper understanding of how to use AI well: how to engage responsibly with tools that hold powerful potential, yet are methodologically alien to traditional humanities practices. In this respect, the existing AI ethics literature does not usually speak to the key concerns of researchers working at the intersection of computational methods and historical inquiry.

Technical fixes, even those motivated by ethical considerations, rarely account for the interpretive complexities and disciplinary norms that shape humanities scholarship. Instead, more fruitful models may be found in audit-based approaches informed by critical theory (Berry, 2022b, 2023). These include dataset audits (Vecchione et al., 2021), critical metadata practices, and frameworks developed within critical archival studies and data feminism (D'Ignazio and Klein, 2020). Such methods aim to uncover the epistemological and political dimensions of data and model construction, moving beyond mere explanation towards structural critique and transformation. While no solution is perfect, especially given the inherently flawed nature of most available datasets, these approaches offer tools for more critical engagement with the problematic legacy of historical data.

3. Auditing to tackle bias at its root

Another central concern in the critical evaluation of large language models within the history, philosophy, and sociology of science pertains to the political orientations and implicit commitments of the actors and institutions that create data and develop AI technologies. The values embedded in AI systems, shaped by the political, economic, or ideological contexts of their creators (Crawford, 2021), profoundly influence the outputs these models produce. Since LLMs are trained on massive, opaque, and uncurated

datasets that incorporate some of the most discriminatory and toxic content available online, their core behaviour reflects these origins (Paullada et al., 2021). Particularly concerning is the use of superficial interventions to manage the outputs of LLMs, such as post hoc filtering mechanisms intended to remove racist or otherwise harmful content. While these filters may suppress problematic outputs at the interface level, the underlying model architecture remains unchanged, and therefore the biases embedded in the training data persist. Data workers involved in this post hoc filtering are constantly exposed to such material, resulting in psychological harm and exploitation, leading scholars to be complicit in ethically problematic industry practices, potentially without being aware of those concerns (Yang et al., 2025). Yet, the fragility of superficial fixes becomes particularly evident when safeguards are bypassed or relaxed, revealing that the models still generate outputs shaped by racist, sexist, or otherwise exclusionary patterns. Feminist scholars have critiqued this form of sanitisation as inadequate (Klein and D'Ignazio, 2024) and potentially precarious in the context of shifting political climates where such filters can be modified or removed. In a critique of discriminatory search engine results, Noble (2018) documented how Google's response to harmful content often involved redirecting queries rather than addressing structural causes – an approach that resonates with current responses to problematic LLM outputs. As in the case of Google's search algorithms, filtering in LLMs does not correct the underlying patterns; it merely reroutes them, allowing the problem to reappear under slightly different conditions. These examples reinforce a broader point: even where technical interventions can serve as useful starting points, they must be accompanied by critical reflection on the historical, political, and epistemological conditions in which both data and technologies are embedded. Without this, even the most advanced systems risk reproducing old exclusions under the guise of innovation. All in all, these dynamics highlight the long-lasting consequences of training LLMs – or any AI for that matter – on datasets whose exact contents are unknown or undisclosed. Their opacity makes rigorous critique or assessment of the data, and by extension the model, exceptionally difficult. Adding to this, current AI development depends on precarious, outsourced, and exploitative labour practices (Gray and Suri, 2019).

For this reason, feminist and critical approaches consistently call attention to training data as the root site of ethical concern. Addressing systemic bias in AI systems requires far more than technical adjustments or user-facing filters. It demands sustained engagement with the datasets that shape model behaviour. As a response, many scholars advocate for dataset (Lang and Suárez Cronauer, forthcoming-2026) and algorithm audits (Vecchione et al., 2021) to expose and interrogate the foundations on which these systems are built. Beyond dataset documentation (e.g., Gebru et al., 2021), algorithm audits have also revealed embedded biases in model outputs (Vecchione et al., 2021). For example, prompting an image-generating model to visualise an inclusive bathroom revealed a stark disability bias: the model was unable to produce a usable representation of a disabled-accessible facility (Manzoor et al., 2025). This not only points to a lack of spatial reasoning capacity within the model architecture but also underscores how bias in training data results in the erasure of marginalised needs and experiences. Such audits highlight the importance of testing model performance in ways that reflect diverse, real-world scenarios, particularly those not centred in dominant cultural narratives. While

such interventions are likely the most effective, it must also be acknowledged that not all researchers may have the capacity to undertake them extensively: Time, expertise, and financial constraints, such as within the framework of grant-funded projects, make it difficult to incorporate comprehensive ethical review or auditing processes unless they are explicitly required and supported by funding bodies.

Scholars frequently rely on inherited or repurposed datasets for practical reasons, but these usually carry the imprint of earlier exclusions and biases. This is especially true when dealing with historical data. The process of digitisation itself frequently amplifies already-dominant voices, as materials that have survived previous curatorial selections tend to be those already privileged by institutional and historical forces. As a result, many early digital humanities projects have ended up reproducing familiar canons, such as works by Goethe or Shakespeare, despite the fact that these topics were arguably already over-researched and of diminishing interest even in traditional, non-digital scholarship. However, audits can also be applied to datasets assembled for humanities research to mitigate the effects of these sometimes unavoidable circumstances (Lang and Suárez Cronauer, forthcoming-2026). If we cannot fundamentally change the datasets we have, we can at least make visible what they contain and what is missing.

4. AI's unsustainability

Environmental concerns further complicate the ethical landscape of LLMs. The energy demands associated with training and deploying large models are considerable and projected to rise sharply in the absence of regulation (Strubell et al., 2019; Bender et al., 2021; Crawford, 2021: chapter "Earth", 2024; Luccioni et al., 2024; Lang et al., 2025). While some have argued that the environmental impact of occasional personal use is negligible (Masley, 2025), particularly when compared to high-emission activities regularly practiced by many scholars such as air travel, this should not justify indiscriminate use. In many humanities research contexts, smaller and more specialised machine learning models may be more appropriate: smaller, non-LLM models are not only easier to fine-tune and more stable, but also significantly less resource-intensive.

For instance, the Data Science Lab at the Staatsbibliothek zu Berlin prefers traditional models over large-scale foundation models (cf. Neudecker, 2023). These choices reflect not only a commitment to environmental, financial and technical sustainability but also frequently provide better results and are more efficient to implement for specialized Humanities use cases. Additionally, large models require high-performance computing infrastructure that many institutions cannot maintain. In contrast, using only what is necessary aligns with both budgetary and ecological responsibility.

Energy consumption is often the primary concern in discussions of sustainability, though its calculation is not straightforward (Lang et al., 2025). While some tools now exist for estimating the energy usage of custom-trained models, many consumer-facing AI services such as ChatGPT do not provide transparent information. Institutions like the German GWDG, for instance, offer high-performance computing resources to researchers, but these, too, still lack tools for estimating or monitoring energy consumption.

As Kate Crawford has pointed out, water usage is another significant but under-discussed issue when it comes to AI's environmental impacts (Crawford, 2024). Many server farms are located in desert regions and require substantial water resources for cooling that might otherwise meet basic human needs. This trade-off raises serious ethical questions about the prioritisation of computational infrastructure over the well-being of local populations.

Electronic waste (e-waste) represents another environmental cost (Strubell et al., 2019). As AI systems evolve, so too must the hardware that supports them. This leads to frequent upgrades and the disposal of equipment that has either become outdated or reached the end of its serviceable lifecycle. Much of this e-waste is exported to the Global South, where it is often dumped without proper recycling or regulatory oversight. The environmental burden of high-end computing thus extends far beyond the institutions that use it. However, even technologically unrelated issues like, quite simply, long-term research data preservation or the digital preservation of digitized cultural heritage collections, are equally affected by this issue.

Researchers must therefore be mindful of the full material footprint of AI usage. Running LLMs at scale, particularly when not necessary, contributes to global ecological degradation. Yet, efficient prompting and careful API use can reduce both financial costs and environmental impact. Since energy seems to correlate with the volume of input and output tokens, optimising these variables may have a tangible ecological benefit in larger scale usage scenarios. Some programming libraries already provide tools for tracking environmental impacts of one's token usage during inference, which may become increasingly important in large-scale projects conducted by libraries or digital humanities centres.[5]

5. Redistributing labour and research integrity

While explainable AI (XAI) should not be overestimated in its capacity to bring about structural improvements to artificial intelligence systems, it remains a valuable diagnostic instrument. It can illuminate how algorithms operate, expose systemic gaps, and support broader efforts in algorithm auditing. For example, recent work employing feature embeddings has revealed that the large language model Claude, when tasked with solving mathematical problems, does not actually perform the mathematical operations it is supposed to perform but rather, manages to get to a correct answer somehow through hardly understandable next-token prediction (Lindsey et al., 2025). More significantly, although Claude produces textbook-style explanations, it does not actually arrive at its results through the application of mathematical rules it could have learned from the maths textbooks in its training data. Instead, it relies on statistical next-token prediction. This discrepancy between the model's generated explanation and its underlying mechanism underscores a critical issue: large language models can

5 However, these usually do not account for costs during training, which are tracked in other scores, such as the AI Energy Score. For more information and a concrete example, see Lang et al. (2025).

fabricate plausible yet misleading accounts of their reasoning processes, thereby raising important concerns about the reliability of self-explanation in machine systems.

Hallucinations (Xiao and Wang, 2021; Singh and Singh, 2025), now widely discussed in both academic and public discourse, pose a threat to research quality and scholarly integrity.[6] While not illegal, the use of inaccurate or misleading AI-generated content in academic work can reinforce harmful ideologies or perpetuate misinformation. As these technologies become more integrated into scholarly workflows, vigilance is required – not only to correct the outputs but to question the infrastructures that produce them.

Transparency remains a major issue in this context. Many AI systems are presented as "open" or accessible while the datasets used to train them remain closed (Liesenfeld and Dingemanse, 2024). This "open-washing" obscures the fact that models are often trained on massive corpora containing unconsented, pirated and harmful content. Kate Crawford's work has shown that this problem predates LLMs: even earlier machine learning systems already relied on uncurated datasets drawn from surveillance, military, or corporate contexts (Crawford, 2021: chapter "Labour"). Once released, such datasets often persist in circulation and continue to shape new models.[7] Even when the flaws of a dataset are well known, the lack of viable alternatives and the sunk cost of development mean that the same corpora continue to be used, and the harms embedded in them are further perpetuated.

A result of this is that even when researchers themselves are not directly working with such datasets, the tools or algorithms they – perhaps even inadvertently – use may be

6 For a more in-depth treatment of hallucinations in the context of this edited volume, see the contribution Meding and Daugs (2026) or Singh and Singh (2025). However, the term *hallucination* has been criticised in various contexts as problematic. First, it psychologises AI, further humanising it and obscuring the fact that this is a technical system. What is called a "hallucination" is not a pathological, psychological delusion but an expected technical behaviour. On a technical level, the term is also misleading, as it assumes that producing incorrect information is a mistake. In reality, these systems are not "making errors" in that sense. All outputs of large language models are the result of mathematical operations and computational programming functions. The model is doing precisely what it is designed to do: predict the most probable next token, given the current context, with a certain degree of randomness. There is no perception involved; it is simply statistics. Anthropomorphic framing not only humanises, psychologises, and pathologises this technical functionality but also implies that it is a malfunction when in fact, this is exactly how the system is intended to operate. The so-called "hallucination" is not a bug; it is an inherent feature of the model's architecture. The term makes it appear to us as though the model should not be doing this, but in reality, next-token prediction is the core behaviour. Similarly, the model's tendency to generalise and omit details is not a malfunction but a built-in characteristic. LLMs are designed to produce knowledge that is as generalisable as possible, which inherently leads to forgetting specific details and favouring central, mainstream knowledge over less standard, more diverse knowledge from the margins. This has the built-in side effect of erasing diversity and nuance, precisely the kinds of details in historical sources that are of greatest interest to historians. An example criticizing the term "hallucination" is: Jason Bell on LinkedIn: https://www.linkedin.com/feed/update/urn:li:activity:7353693695822438400

7 There is a range of publications addressing the lifecycle of AI datasets, including their often flawed creation circumstances and their out-of-control persistence and continued use in AI even after they have been publicly revoked because of their blatant ethical flaws: e.g., Luccioni et al., 2022; Orr and Crawford, 2023; Luccioni and Crawford, 2024; Orr and Crawford, 2024.

shaped by these legacies. Many of these datasets originated in institutional contexts with little accountability or ethical oversight, and their continued use embeds these values into current technologies. For scholars in the humanities especially these entanglements cannot be overlooked.

LLMs tend to produce generalised summaries that obscure or omit specific details, and their architecture favours information from the mainstream rather than from the margins (Peterson, 2025), whose nuance and diversity historians are usually more interested in than in blatant overgeneralisations. Although skilled prompting can help mitigate some of these tendencies, the problem is rooted in the design and training of the models. As such, outputs must be thoroughly verified, especially in scholarly work, and scholars may find that models simply do not produce adequate results in their specialised areas of expertise if these are not well represented in the data the algorithms were trained on.

On top of all that, contrary to promises of labour-saving automation, LLMs often shift the labour of research rather than reducing it. Rather than freeing scholars from repetitive tasks, these systems require them to function as human fact-checkers for unreliable machine outputs. While some tasks may indeed be accelerated, the overall result is frequently an increase in cognitive and editorial labour.

And while it is not illegal to produce substandard academic writing, the proliferation of fraudulent or low-quality papers generated through paper mills represents a serious threat to scholarly integrity and to public trust in science more broadly. While early examples were often easily identified, contemporary high-performing tools make detection significantly more difficult (Else and Van Noorden, 2021; Liverpool, 2023). Academic publishers now face an ongoing and escalating challenge, effectively engaged in an arms race against those producing fake submissions. This dynamic not only undermines scientific credibility but also places additional labour demands on journal editors, reviewers, and the publishing infrastructure as a whole.

The increasing use of AI-assisted writing tools contributes further to the system that generates publication pressure in the first place: While AI tools can support writing workflows in meaningful and ethically acceptable ways (Abernethy, 2024), they also risk accelerating publication rates in ways that reinforce the already problematic "publish or perish" culture in academia (Rawat and Meena, 2014; Johnson et al., 2024; Kendall and Teixeira da Silva, 2024; Hegde, 2025). As researchers are able to produce more text more quickly, expectations around productivity may rise, potentially diminishing the time available for reflection, critical engagement, and scholarly rigour. These developments are not necessarily unique to history, philosophy and sociology of science and knowledge research but should nonetheless be considered within the broader context in which such research now unfolds.

Regulatory and funding institutions must take a more active role in mandating ethical interventions, including dataset audits and representational analyses, as part of project deliverables (cf. Lang and Suárez Cronauer, 2026). Encouragingly, some funding agencies are beginning to request AI ethics statements. However, if such statements remain vague, requiring only that researchers affirm the absence of ethical concerns, then their practical value is minimal. There is also a real risk of "ethics-washing," wherein researchers are incentivised to downplay or overlook issues to streamline approval (Powers

and Ganascia, 2020; Lambrecht and Moreno, 2024). Moreover, these initiatives, through their reliance on virtue ethics, usually assume that good people will intuitively know how to act ethically, downplaying the substantial skill and expertise involved in thoughtfully evaluating the ethics of data practices. As the authors of *Data Feminism* have pointed out, those who have not experienced structural discrimination themselves are often unaware of how their work may negatively affect others – a phenomenon they refer to as the "privilege hazard" (D'Ignazio and Klein, 2020: 28–29; Lang and Suárez Cronauer, 2026). Without the deliberate inclusion of diverse perspectives and affected communities, researchers cannot be expected to reliably detect harms their technologies reproduce or exacerbate. To act more ethically, scholars need tools and strategies. Methods such as Explainable AI (XAI), although critiqued earlier in this article for its limited capacity to address systemic bias, can nevertheless offer valuable starting points for identifying and understanding model behaviour. In the absence of transparency from developers, third-party audits have become increasingly important. Resources informed by critical archival studies – such as data sheets, data papers, and related documentation practices (Gebru et al., 2021) – offer structured ways to reflect the situatedness of datasets and clarify their limitations. These tools provide a means of acknowledging the epistemological and ethical boundaries of a dataset, but they have yet to be adopted at scale within the development of large foundation models.

6. Humanity, solidarity and care for better AI

Amid the challenges outlined in this article, there are also grounds for optimism. Frameworks like data-centric AI present new opportunities. The use of smaller and more efficient architectures in particular may be better suited to many humanities tasks. These models are often easier to fine-tune, more stable, and more transparent than large-scale foundation models, and can be retrained on domain-specific corpora. Although such retraining is not always feasible in humanities contexts due to limited data or computational resources, the possibility remains significant. It is also encouraging to note the increasing visibility and impact of scholars engaged in critical AI ethics, whose work continues to inform more accountable and reflective approaches.

A key aim of this paper has been to introduce readers to these resources and to encourage engagement with existing work in the field. Many useful tools and theoretical frameworks are already available, and increased uptake would strengthen both our practice and our discourse. What is needed is not only more incentive for such engagement, but also greater self-awareness across the field. The digital humanities, for example, already contain a wealth of literature on the ethical implications of data, algorithms, and infrastructure. The challenge now is to mobilise this work effectively and to make it part of standard practice.

Central to this endeavour is the practice of dataset documentation and auditing. Whether applied to one's own data or to third-party datasets, audits can help illuminate representational imbalances and ethical risks. Data papers and dataset metadata provide crucial context about the origins, composition, and limitations of datasets, thus enabling more responsible reuse. For instance, knowing that 50–90% of authors in a given dataset

are male allows researchers to at least state that the dataset will likely reflect a male gaze. While this does not resolve the issue, it creates a baseline for critical reflection and makes visible biases that may otherwise remain implicit. Truly addressing such biases will require filling data gaps, where possible, yet this is especially challenging in historical research where records may be incomplete or unrecoverable. Nevertheless, efforts can be made to mitigate these problems whether through more inclusive digitisation, redesigning research projects to account for diverse perspectives from the outset, or even initiating new collections that avoid replicating past exclusions. In the long term, these steps might even support retraining foundation models on less problematic grounds.

Such efforts cannot be undertaken by individual researchers alone. As Valleriani (2025) has argued, state institutions should take a more active role in developing AI infrastructures, including foundation models. Currently, these models are largely built by private corporations with opaque motivations, and their training data and objectives are often misaligned with scholarly or public values. If large language models are to become core infrastructures for knowledge production, as they already are in many fields, it is imperative that their development is subjected to ethical scrutiny and democratic oversight. The humanities, with their critical and contextualising capacities, are well positioned to contribute to this process not only by critiquing existing systems, but by actively shaping the values and materials that future models are built upon. Importantly, researchers need not be AI specialists to participate in this work. Dataset audits, for example, can be conducted by anyone familiar with the material, and they offer an accessible entry point into AI ethics. By critically examining the datasets they use or create, scholars can contribute meaningfully to the development of more equitable and transparent AI systems. The cumulative effect of such individual actions may help shift the field as a whole.

It is no longer sufficient to ask how we might better *use* LLMs. We must also ask how *we* might contribute to making them better. This includes contributing the intellectual and cultural materials, methodological insights, and ethical reflections that only the humanities can provide. Regulatory institutions must support this process through clearer requirements, enforceable standards, and publicly accountable infrastructures. Creating foundation models whose content is transparent and whose construction is participatory would ensure both community control and alignment with shared ethical aims. Such an approach would represent a decisive alternative to systems designed to serve narrow political or economic interests and would contribute to the development of more equitable and responsible technologies.

A significant yet often overlooked factor in the development of artificial intelligence, computing, and adjacent fields is the absence of solidarity, particularly in the historical and ongoing extractive practices of the tech industry. Many of the ethical and epistemological challenges currently facing these domains can be traced to systemic undervaluation and invisibilisation of labour. For instance, large-scale projects such as Google Books were made possible through the labour of underpaid and insufficiently acknowledged book scanners. Similarly, the infrastructure supporting contemporary AI relies on 'ghost workers' and underpaid contributors from platforms such as Amazon Mechanical Turk (Gray and Suri, 2019).

These patterns are not unique to industry alone. Within the digital humanities and the historical sciences, tasks dismissed as "menial" such as metadata collection or the work of librarians have historically been marginalised, both in terms of compensation and attribution (Lang and Suárez Cronauer, 2026). This systemic disregard for certain forms of labour has long-term consequences. One tangible outcome is the difficulty scholars now face in tracing the origins of specific data or annotations, due to the absence of proper documentation and attribution. The lack of recognition for this invisible labour has contributed to a structural opacity in digital infrastructures, which in turn affects how knowledge is produced, evaluated, and utilised. This demands a renewed commitment to solidarity within both scholarly and technical communities. Valuing diverse forms of expertise and properly crediting all contributors regardless of the perceived status of their labour are essential steps in reshaping the ethics and culture of AI development. While individual efforts may appear limited, collective action across the field holds the potential for meaningful change. A non-extractive, solidaristic model of knowledge production is both necessary and achievable.

Contemporary AI development has largely been driven by unregulated and extractive practices, frequently marked by non-consensual data collection, lack of transparency, and motivations aligned with economic and military interests (Crawford, 2021: chapter "Data"). These practices stand in stark contrast to the ethical commitments we stand for within scholarship and, in particular, the humanities. A more ethical framework would centre solidarity, aiming for collective benefit and co-liberation rather than economic gain for the few. Such an approach resonates with implementing the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics: GIDA, 2021) in the context of AI and a (feminist) ethics of care (Gray and Witt, 2021), which may offer a more constructive ethical foundation than current emphases on explainability. The latter is often ideologically tethered to ideals of rigour and quantification that implicitly devalue humanities knowledge as "soft" or secondary (Riley, 2017). Yet it is precisely this form of expertise that is capable of exposing the limitations of technical solutions. Humanities scholars are well positioned to advance this critical perspective. The foundation of future AI must not be left to unaccountable actors. It is time to build models, literally and ethically, that we can collectively stand behind.[8]

References

Abernethy NJ (2024) Let stochastic parrots squawk: Why academic journals should allow large language models to coauthor articles. *AI Ethics*. DOI: 10.1007/s43681-024-00575-7.

8 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Ananny M and Crawford K (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*. DOI: 10.1177/1461444816676645.

Bender EM, Gebru T, McMillan-Major A and Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. DOI: 10.1145/3442188.3445922.

Berry DM (2022a) AI, ethics, and digital humanities. In: O'Sullivan J (ed.) *The Bloomsbury Handbook to the Digital Humanities*. London: Bloomsbury Academic, pp.445–458. DOI: 10.5040/9781350232143.ch-42.

Berry DM (2022b) Critical digital humanities. In: O'Sullivan J (ed.) *The Bloomsbury Handbook to the Digital Humanities*. London: Bloomsbury Academic, pp.125–136. DOI: 10.5040/9781350232143.ch-12.

Berry DM (2023) Tracing 'toxicity' through code: Towards a method of explainability and interpretability in software. *Digital Humanities Quarterly* 17(2). Available at: https://dhq.digitalhumanities.org/vol/17/2/000706/000706.html (accessed 2025–11-25).

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crawford K (2024) Generative AI's environmental costs are soaring – and mostly secret. *Nature* 626(8000): 693. DOI: 10.1038/d41586-024-00478-x.

D'Ignazio C and Klein LF (2020) *Data Feminism*. Cambridge, MA: MIT Press.

Dubber MD, Pasquale F and Das S (eds) (2020) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780190067397.001.0001.

Eberle O (2026) Grounding AI in humanistic inquiry. Interdisciplinary challenges for evaluation and interpretability. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Else H and Van Noorden R (2021) The fight against fake-paper factories that churn out sham science. *Nature* 591: 516–519. DOI: 10.1038/d41586-021-00733-5.

Floridi L (2023) *The Ethics of Artificial Intelligence – Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H and Crawford K (2021) Datasheets for datasets. *Communications of the ACM* 64(12): 86–92. DOI: 10.1145/3458723.

GIDA (Global Indigenous Data Alliance) (2021) CARE principles of Indigenous data governance. Available at: https://www.gida-global.org/care/ (accessed 2025–11-25).

Gray J and Witt A (2021) A feminist data ethics of care for machine learning: The what, why, who and how. *First Monday* 26(12). DOI: 10.5210/fm.v26i12.11833.

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Eamon Dolan Books.

Gunkel DJ (2020) Perspectives on ethics of AI: Philosophy. In: Dubber MD, Pasquale F and Das S (eds) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, pp.538–553. DOI: 10.1093/oxfordhb/9780190067397.013.35.

Hegde SN (2025) Publish and cherish, not 'publish or perish.' *APIK Journal of Internal Medicine* 13(3): 159–160. DOI: 10.4103/ajim.ajim_39_25.

Hildebrandt M (2021) The issue of bias: The framing powers of machine learning. In: Pelillo M and Scantamburlo T (eds) *Machines We Trust: Perspectives on Dependable AI*. Cambridge, MA: MIT Press. DOI: 10.7551/mitpress/12186.003.0009.

Johnson TF, Simmons BI, Millard J, Strydom T, Danet A, Sweeny AR and Evans LC (2024) Pressure to publish introduces large-language model risks. *Methods in Ecology and Evolution* 15(10): 1771–1773. DOI: 10.1111/2041-210X.14397.

Kendall G and Teixeira da Silva JA (2024) Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing* 37(1): 55–62. DOI: 10.1002/leap.1578.

Khutsishvili K (2026) AI and the scientist. On the fracture of epistemic authority. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Klein L and D'Ignazio C (2024) Data feminism for AI. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 100–112. New York: Association for Computing Machinery. DOI: 10.1145/3630106.3658543.

Kuipers B (2020) Perspectives on ethics of AI: Computer science. In: Dubber MD, Pasquale F and Das S (eds) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, pp.420–441. DOI: 10.1093/oxfordhb/9780190067397.013.27.

Lambrecht F and Moreno M (2024) What is AI ethics? *American Philosophical Quarterly* 61(4): 387–401. DOI: 10.5406/21521123.61.4.07.

Lang S and Suárez Cronauer E (forthcoming-2026) Dataset audits for mitigating data gaps. *Computational Humanities Research Journal*.

Lang S and Suárez Cronauer E (2026) Beyond data feminism: Toward ethical data work in the (digital) humanities. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*. DOI: 10.17175/wp_2026.

Lang S, Pitawanik W, Belouin P, Sevink E, Olszynko-Gryn J, Freeborn A and Benson E (2025) *Quantifying the environmental footprint of curating datasets with LLMs*. Working paper. DOI 10.5281/zenodo.17902821.

Liao SM (ed.) (2020a) *Ethics of Artificial Intelligence*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780190905033.001.0001.

Liao SM (2020b) A short introduction to the ethics of artificial intelligence. In: Liao SM (ed.) *Ethics of Artificial Intelligence*. Oxford: Oxford University Press, pp.1–42. DOI: 10.1093/oso/9780190905033.003.0001.

Liesenfeld A and Dingemanse M (2024) Rethinking open source generative AI: Openwashing and the EU AI Act. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 1774–1787. DOI: 10.1145/3630106.3659005.

Lindsey J, Gurnee W, Ameisen E, Chen B, Pearce A, Turner NL, Citro C *et al.* (2025) On the biology of a large language model. *Transformer Circuits Thread*. Available at: https://transformer-circuits.pub/2025/attribution-graphs/biology.html (accessed 2025–11-25).

Liverpool L (2023) AI intensifies fight against 'paper mills' that churn out fake research. *Nature* 618: 222–223. DOI: 10.1038/d41586-023-01780-w.

Luccioni AS, Corry F, Sridharan H, Ananny M, Schultz J and Crawford K (2022) A framework for deprecating datasets: Standardizing documentation, identification, and

communication. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Luccioni AS and Crawford K (2024) The nine lives of ImageNet: A sociotechnical retrospective of a foundation dataset and the limits of automated essentialism. *Journal of Data-Centric Machine Learning Research*. Available at: https://data.mlr.press/assets/pdf/v01-4.pdf (accessed 2025–11-25).

Luccioni AS, Jernite Y and Strubell E (2024) Power hungry processing: Watts driving the cost of AI deployment? In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 85–99. DOI: 10.1145/3630106.3658542.

Manzoor R, Hussain W and Anjum ML (2025) Out of dataset, out of algorithm, out of mind: A critical evaluation of AI bias against disabled people. *AI & Society* 40: 3941–3951. DOI: 10.1007/s00146-024-02168-8.

Masley A (2025) Why using ChatGPT is not bad for the environment – A cheat sheet. *Substack* (blog), April 2025. Available at: https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about (accessed 2025–11-25).

Meding H and Daugs A (2026) On the use and limitations of large language models in historical scholarship. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Neudecker C (2023) Digital curation and artificial intelligence – Opportunities and risks for cultural heritage institutions. In: Thiel S and Bernhardt J (eds) *AI in Museums: Reflections, Perspectives and Applications*. Bielefeld: transcript Verlag, pp.149–162.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

O'Neil C and Gunn H (2020) Near-term artificial intelligence and the ethical matrix. In: Liao SM (ed.) *Ethics of Artificial Intelligence*. Oxford: Oxford University Press, pp.237–270. DOI: 10.1093/oso/9780190905033.003.0009.

Orr W and Crawford K (2023) The social construction of datasets. *New Media & Society*.

Orr W and Crawford K (2024) Building better datasets: Seven recommendations for responsible design from dataset creators. *Journal of Data-Centric Machine Learning Research*. Available at: https://openreview.net/forum?id=6bd8BrRKTW (accessed 2025–11-25).

Paullada A, Raji ID, Bender EM, Denton E and Hanna A (2021) Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2(11): 100336. DOI: 10.1016/j.patter.2021.100336.

Peterson AJ (2025) AI and the problem of knowledge collapse. *AI & Society* 40: 3249–3269. DOI: 10.1007/s00146-024-02173-x.

Powers TM and Ganascia J-G (2020) The ethics of the ethics of AI. In: Dubber MD, Pasquale F and Das S (eds) *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, pp.26–51. DOI: 10.1093/oxfordhb/9780190067397.013.2.

Rawat S and Meena S (2014) Publish or perish: Where are we heading? *Journal of Research in Medical Sciences* 19(2): 87–89.

Riley D (2017) Rigor/us: Building boundaries and disciplining diversity with standards of merit. *Engineering Studies* 9(3): 249–265. DOI: 10.1080/19378629.2017.1408631.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215.

Singh D and Singh A (2025) Ubiquity of LLM hallucinations across critical domains: A survey. In: Yuan S, Malliaros F and Zheng X (eds) *Trends and Applications in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, vol.15835. Singapore: Springer, pp.115–132. DOI: 10.1007/978-981-96-8197-6_9.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Strubell E, Ganesh A and McCallum A (2019) Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp.3645–3650. Florence: Association for Computational Linguistics. Available at: https://aclanthology.org/P19-1355 (accessed 2025–11-25).

Valleriani M (2025) Large language models that power AI should be publicly owned. *The Guardian*, 26 May. Available at: https://www.theguardian.com/technology/2025/may/26/large-language-models-that-power-ai-should-be-publicly-owned (accessed 2025–11-25).

Vecchione B, Barocas S and Levy K (2021) Algorithmic auditing and social justice: Lessons from the history of audit studies. In: *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. New York: ACM. DOI: 10.1145/3465416.3483294.

Xiao Y and Wang WY (2021) On hallucination and predictive uncertainty in conditional language generation. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, pp.2734–2744. Association for Computational Linguistics. Available at: https://aclanthology.org/2021.eacl-main.236 (accessed 2025–11-25).

Yang T, Strippel C, Keiner A, Baker D, Chávez A, Kauffman K, Pohl M, Sinders C and Miceli M (2025) *Ethics of Data Work: Principles for Academic Data Work Requesters*. Berlin: Weizenbaum Institute. DOI: 10.34669/WI.DP/48.