Challenges in Data Preservation for AI and ML Systems

Emma L. Tonkin [1] and Gregory J. L. Tourte [2]

Abstract: The management and preservation of machine learning (ML) and artificial intelligence (AI) data is increasingly a concern for research institutions, as well as for institutions and industry organisations making use of this type of data and method. This paper summarises key issues in this area, presenting the case that there are significant benefits to the industry in developing best practices and joint standards in this area, and identifying the benefits of this approach, as well as highlighting risks and a current paucity of best practice in the area.

Keywords: data preservation, data management, artificial intelligence, machine learning, best practices

1 Introduction

Data management refers to the practices and principles used to manage data within an organisation, and is a vital area for machine learning. Digital preservation, an associated concept, is a family of practices and activities [Wi17] that includes planning around data retention, making decisions around the future of the data and materials, and acting to maintain appropriate continued access to the materials beyond the initial lifecycle of the material. Digital preservation references the idea of conserving data: for example, once a research project is complete it is necessary to decide what must be done with the data collected. Some administrative data, no longer required following the completion of the project, may be deleted, while other data – required to assure the validity of the experiment, or perhaps required for reanalysis or reproducibility of the dataset – must be retained. More than that, there are concrete benefits to ensuring that data is as visible as possible to other researchers, since – if it is appropriate to do so – data reuse increases the benefit of that data collection and maximises the impact of the work, reflecting our commitment as researchers to honouring the kind contribution of each research participant. The term digital curation [Rh24] refers to processes during the entire lifecycle of data, models and supplementary material, from appraisal (e. g. selection and retention), data management and preservation, through to enrichment or linking of data, and the facilitation of data reuse.

The regulatory environment in which decision-making systems in particular are used and developed raises challenges that often require ongoing access to data and systems to meet the requirements of the field, and consequently, at each stage of the development and

[1] University of Bristol, Digital Health, Bristol, UK,
e.l.tonkin@bristol.ac.uk, https://orcid.org/0000-0001-7405-4982
[2] University of Oxford, Advanced Research Computing, Oxford, UK,
gregory.tourte@it.ox.ac.uk, https://orcid.org/0000-0002-2819-392X

doi:10.18420/inf2024_38

deployment of systems of this kind it is important to ensure that these issues may be addressed. Recent regulatory debate internationally has highlighted the importance of features such as explainability (see e. g. Kuiper et al. [Ku22] and Werder; Ramesh; Zhang [WRZ22]), as well as the ability to trace provenance and characterise areas of validity of machine learning systems. We are, therefore, trending toward an environment in which the ability to accurately characterise artificial intelligence (AI) and machine learning (ML) products and services is increasingly likely to become an obligation, especially for certain more sensitive use cases.

In this paper we attempt to identify challenges encountered when attempting to apply data preservation techniques to large datasets used, created, and/or generated when working on or with AI and machine learning models, and to provide guidelines for developing methods to apply by researchers on their own datasets. Note that we make the assumption that the analysis of which data to gather and collect has already been carried out. There is an entire separate process to go through in order to determine how much data should be collected and is sufficient to carry out the tasks defined in a research project, but this is out of scope for this paper.

2 Key principles and tasks in data management

Problem statement: Decision-making processes surrounding data are focused in part on understanding the data that is held, characterising it and making decisions about it, and in part on understanding what the data's residual value may be, and for which tasks or purposes the data may be used. As a general observation, ML applications typically constitute "distant reads" of datasets isolated from their initial context, which is to say, understanding by aggregation and analysis of large datasets [OKM12], and hence are only as valid as experimental design and benchmarking permits [Li21]. In ML and AI applications in particular, the regulatory environment increasingly implicates specific targets for service providers to meet. In general, data preservation efforts in machine learning might among other things facilitate tasks such as the following: recreation of models from source data; testing of alternative methodologies; removal of data where required (e. g. compliance with GDPR right to be forgotten or right to rectification, see "regulatory background" section below); reproducibility of existing experiments and validations; evaluation of existing datasets or models, such as in new contexts of use or user populations; tertiary reuse of datasets and accompanying resources; and evaluation and characterisation of label sets (ground truths). There is a great deal of data reuse in ML, and considerable thought has gone into the characterisation of resources such as models and data. However, we argue that there is an ongoing need to further develop practices in this area to align with emerging regulatory standards and best practices, as well as developing an interdisciplinary understanding of AI/ML full-lifecycle practices as a form of stewardship, comprising digital preservation and curation processes surrounding these often large, heterogeneous, often human-centred and sensitive datasets and artefacts.

Data reproducibility is key to good academic practice, and machine learning is no exception [Se23]. There are many potential limiters to reproducibility in machine learning, such as:

- **Inherent non-determinism:** Since methods such as neural network typically bring into play various sources of randomness, it is to be expected that different runs of the same task will produce differing outcomes, although these may in aggregate provide statistically similar outcomes. This is inherent to the methods employed, and is likely to be the least problematic reproducibility issue. However, Semmelrock et al. [Se23] note that some methods, such as reinforcement learning, suffer disproportionately from this issue and commonly use frameworks to mitigate the problem.
- **Lack of documentation:** Often studies simply do not provide sufficient information to reproduce the work. This may be a consequence of issues releasing aspects of the work (e. g. source code, data, etc), potentially as a result of regulatory, IP or data protection concerns, or simply lack of time, funding, awareness or incentives to take part.
- **Unclear methodology, or methodological problems:** methodologies can simply be insufficiently clear to follow in their entirety (for example, failure to adequately describe necessary preprocessing steps) or specific flaws may exist in the methodology as documented. For example, issues of data leakage, such as issues that arise from poorly chosen features, dataset separation practices, sampling biases, nonindependence of train and test samples, and so forth, affect computational reproducibility and provide typically overinflated estimates of a model's efficacy [KN23; Se23]
- **Partial datasets and withdrawal of participant consent:** Current efforts in ML often involve the aggregation of extremely large datasets. However, under EU law the data subject has the right to withdraw their consent for data processing at any time; while there are alternative grounds for data processing that may potentially be employed, it is normal for datasets to change over time. This is nothing new: for example, social media datasets are commonly shared in the form of message IDs to sidestep terms of service that exclude direct sharing of compiled datasets, hence as accounts are deleted or set to private, each copy of the dataset varies slightly [Ba21].
- **Unavailability of datasets:** as Semmelrock et al. [Se23] observe, the field of ML/AI in digital health suffers significantly from difficulty accessing datasets, which are often sensitive in nature. Mitigations, such as the availability of platforms on which experiments can be tested without risk of data leakage, tend to be difficult and expensive, requiring significant support from research organisations and funding bodies.
- **Unavailability of versioned resources and services:** Similarly, a reliance on specific versions of services, libraries, frameworks and other resources leads to changes in performance and outcome. This is likely to be particularly visible in the case of proprietary models, such as GPT versions. Ma et al. [Ma24a] give an example of a researcher who finds that following an update, the behaviour of the system changes significantly, and states that participants in their study believed that proprietary LLM companies "violated well-established software update practices without justification". As a consequence, Ma et al. [Ma24a] find that users of LLMs are likely to move to open source LLMs for research purposes.

- **Environment:** ML cloud platforms are often used to provide large-scale environments, yet the use of these platforms does not in itself confer reproducibility: Gundersen; Shamsaliei; Isdahl [GSI22] found that an ML experiment, if performed on different ML platforms, is likely to produce a statistically significant difference in results.
- **Hype and wishful thinking:** as observed by Semmelrock et al. [Se23], nonreplicable results are cited more often than replicable results, and over-optimism is common. AI "techno-optimism" leads to accounts of technical performance and promise that outstrip the evidence [Ma24b], and may result in what Markelius et al. [Ma24b] refer to as "mystification" of AI technologies, obfuscating important practical and ethical concerns about technology and applications, such as reproducibility. These issues, alongside anthropomorphism that often applies to certain applications of AI/ML, may lead us into misjudging the attributes and capabilities of a device [LS19].
- **Ethical concerns:** at times, experiments are documented that are carried out in such a way as to render recreation of the study problematic. For example, critiques of the Stanford "gaydar" Wang; Kosinski [WK18] often focus on the sample, the impact of the purported outcomes and the potential for abuse of the results, e. g. [KB23]. The study is itself arguably unhelpfully framed, as well as presenting data licensing and ethics issues, and so efforts to explore aspects of this work would be unlikely to focus on reproducing the study as described.
- **Assertions beyond the evidence:** Over-promising is a risk in ML/AI, partially as a result of the "techno-optimism" described earlier. Ge et al. [Ge23] argue that inflationary effects in AI lead to significantly overpromising the potential of technologies in applied scenarios, taking the example of measuring vocal impairment in Parkinson's disease, and showing that, though reproducible in its own terms, e. g. using a standardised dataset, the experimental results will not be reflected in real-world applications.

2.1 Regulatory background around data management and preservation in AI/ML

As the complexity of models increases, a lack of transparency about data sources has become commonplace [Da21]. Data contained within training sets may be "of dubious legality" [LLD23], which leads to a broad swathe of potential issues, ranging from the inability to characterise model bias and limitations through inspection of source data through to issues with reproducibility of outcomes and problems with model safety. Model versioning and provisioning are also potential sources of issues with reproducibility, particularly when making use of externally hosted services – for example, Wang et al. [Wa24] explain that "models are dynamic and continue to evolve over time", making use of fixed versions of models to avoid these pitfalls.

AI and ML can be applied across any field, and many involve little or no personal data. However, many currently popular applications of machine learning do involve personal data,

either incidentally (for example, the use of YouTube videos, text from blogs or social media, or images from social media and Instagram) or because the application requires it. For example, targeted advertising intends to draw out associations between information about the user's interests, activities and personal preferences and their response to advertising, in order to optimise the likelihood that the adverts that they are shown result in a website view or a sale. Digital health frequently involves applications of AI/ML to personal data for a wide variety of health-related purposes, such as support for medical diagnosis, or development of tools that provide improved performance or improved user experience for health related tasks such as disease management. When we think of the risks of AI to personal data, however, it is likely that the first case that comes to mind is the large-scale collection and processing of huge amounts of data to power LLMs such as ChatGPT – and the potential for sensitive personal data to be exposed to private companies through the use of services of this kind by professionals such as doctors [Bl24] and judges [Gu24]. The data collection methods of companies such as OpenAI have received much critique, but it is perhaps less widely recognised that many datasets exist that are drawn in a similar manner from resources such as YouTube. As Fried [Fr24] puts it, the perspective of AI companies is that anything publicly available is "fair game". Yet this is far from the reality of the situation. Here, we briefly summarise some features of the GDPR, giving a European perspective on some of the responsibilities of the data processor when processing personal data.

- **Informed consent:** While consent is only one of several grounds on which personal data can be processed, it is often part of the picture. Consent should be freely given (e. g. not forced in order to enable functionality, or the result of a default opt-in, for example), specific, informed and unambiguous.
- **Transparency:** The right to be informed (Articles 13–14 of the GDPR) means that individuals should be able to access clear and transparent information about personal data concerning them, and how it it is collected, used and processed.
- **Right to erasure:** Under Article 17 of the GDPR, individuals have a right to ask for their personal data to be erased.
- **Right to rectification:** Under Article 16 of the GDPR, individuals have a right to ask for their personal data to be rectified. This is relevant to systems that present output as factual data, without providing validation of their output, and which are prone to "hallucinations", such as falsely claiming that a living individual has passed away. OpenAI have stated that they are unable to correct incorrect output [Ru24]
- **Assessing risk and impact of data processing:** No "one-size-fits-all" exists in data processing. One application of ML to a dataset may be quite without risk to the data subject, whereas another may have significant consequences to them. Organisations performing high-risk data processing tasks are obliged to carry out data protection impact assessments, and as a matter of best practice these should be made public.

2.2 Intellectual property risks in generative AI

As well as data protection, many of the potential risk of generative AI in particular relate to intellectual property. It is possible for users of generative AI systems to produce work that, intentionally or unintentionally, reproduces copyrighted material or infringes in some way on an individual's rights. As Appel; Neelbauer; Schweidel [ANS23] assert, generative AI has an intellectual property problem, originating in part from the capability of generative AI to reproduce elements of human performance and ideation (for example, an artist's voice or style), and in part from the often dubiously licensed or unlicensed nature of the source data from which models are trained. When OpenAI's CTO Mira Murati was asked by the Wall Street Journal's Joanna Stern whether their video generation tool Sora was trained on YouTube, Instagram etc, Murati famously responded "I'm actually not sure about that", but suggested that if the videos were publicly available, it was possible. Potential users of many currently popular models are operating in an environment in which the provenance of the training data is unclear.

One risk is that one may be accused of reproducing existing works. Speaking of music generation, Sunray [Su20] points out that there is "an inherently limited musical palette". Automated music generation in general has some risk of reproducing elements of known works, and music generators that rely on what Sunray [Su20] refers to as "a tapestry of upsampled sound" are potentially at greater risk. In a 1986 9th Circuit case in the USA, Fisher v Dees, a rule was established that, if an average audience does not recognise that elements of a work have been appropriated, then the similarity does not signify legally; as such, if it sounds to an average person like another work, then there is a risk that the work may be seen as an appropriation of another's work. A generalisation of this in US law is the "Hand Abstraction Test": similarity of theme or abstract idea is not copyrightable, whereas substantial similarity implies infringement. If the origin or content of the training data is not documented, it is correspondingly more difficult *for the user that the system is supporting, as opposed to a listener with extensive knowledge of the topic* to evaluate whether elements of a work have been reproduced. In general it is currently unclear what risks or liability may result from making use of models and services that potentially lean on material for which no valid licensing for reuse was provided or made available, but at the very least we might imagine that there is an enhanced risk of the model or service eventually becoming unavailable as a result.

Separately, it is also worth noting that the outputs of AI systems are considered uncopyrightable in many jurisdictions. For example, in the EU, copyright on AI-generated data is a topic under active discussion [BS22]. Under EU law, material under copyright is to be "the author's own intellectual creation" [HQ21], so Hugenholtz; Quintais [HQ21] propose that a work must require human intellectual effort and *human* creativity to be copyrightable, meaning that some AI/ML assisted creative processes may be copyrightable, but not all; in any case, the copyright cannot lie with an entity other than a human. The US copyright office has deemed the product of generated AI to be uncopyrightable, since copyright has a

human-authorship requirement (e. g. non-human entities, be they bonobos or ML algorithms, cannot themselves hold copyright). Therefore, at present only aspects of a product that are the result of human input (for example, manual typesetting of AI-generated texts and images) may be copyrighted. Similarly, Australian law requires a human author for copyright ownership, and existing case law shows that AI cannot be named as an inventor for patents. The implications of AI involvement in IP development require careful charting. Returning to decision-making around systems and components that make use of generative AI tools, then, it is important to document the provenance of the data and model as far as possible, so that we ourselves understand where we stand and how we might avail ourselves of the resources we hold. If we do not have this information, it is equally important to recognise the disadvantage at which this places us, and to take this into account when considering retention and reuse of the artefacts involved.

It is useful to distinguish between the uses of AI/ML in research and the inherent differences it brings to the data associated with it. On the one hand, one may only need to use an pre-existing and pre-trained model to run on one's own dataset or generate one's own data. On the other hand, some research may be entirely focused on the development and training of models when data is gathered and collected as input to models. These cases may present different challenges to data preservation. When working with an existing model, where the user has not been given access or input into the choice of data used for development and training, the output of such model may contain copyrighted or sensitive data even though the data from the user appears "safe".

2.3 Information validity: Provenance and hallucination

The tendency of certain models, e. g. LLMS, to "hallucinate" – produce inaccuracies [HHS24] – or produce outcomes that among other things may deviate from user intent or misalign with factual knowledge [Li24], may lead to significant risks for users who rely upon the tool for relevant purposes. For example, if the service is employed as an information retrieval system, a hallucination that might be harmless in other contexts becomes potentially harmful. From a regulatory point of view it is not necessarily a problem for a service to "lie" if it is known to be a liar and handled accordingly, but liability for its errors is likely to lie with the enterprise who presents the service to customers as a factual information retrieval system, e. g. Air Canada's recent loss in court against a passenger who received inaccurate information from a chatbot demonstrates this problem [Gr24], or the 2023 case in which lawyers Schwartz and LoDuca were employed ChatGPT to research a legal case, submitting a legal brief that was found to reference several non-existent cases and failing to verify this when queried, resulting in a fine. Systems which are persuasive but have little regard to the factual nature of assertions present the same danger as any agent, human or otherwise, who prioritises apparent plausibility over factual accuracy [BW23]. A curator seeking to formalise a given application such as legal information retrieval would do well to

supplement the use of such systems with classical information retrieval solutions, and to carefully display the expected reliability of each resource to the user.

2.4 Data auditing: Understanding the data and models we hold

While there are guidelines around the use of tools and processes such as data management plans, for example in the context of Horizon 2020 [Sp21], in the US [BNC21] and in the UK [Ho11] under funding agencies such as EPSRC and others, there are significant challenges involved in encouraging compliance. Bishop; Nobles; Collier [BNC21] cite van Loon et al.'s finding [Lo17] that over half of data management plans do not identify individual(s) responsible for research data management, for example. Shelly; Jackson [SJ18] assesses research data management in 13 Australian research universities and identifies a lack of consistency in practical implementation. Bhardwaj et al. [Bh24] observes in an analysis of evaluation results for 25 ML datasets that "researchers in machine learning, which often emphasizes model development, struggle to apply standard data curation principles". As Steffens et al. [St24] summarises the problem, in the context of a particular area of medical research, there are several important blockers to good RDM, including lack of time, awareness, incentives and funding for its implementation; a lack of understanding of the legal and regulatory aspects of the task; and a need to "better identify meaningful and actionable data among the increasing volume and complexity of data being acquired". This task, the triaging of data, grey literature, software artefacts and so forth, may need to be carried out post hoc in the form of an auditing process.

2.5 Data curation and audit in ML/AL

The task of performing a review of the data and software held is necessarily much more significant in an environment in which little metadata is held or in which processes and processing are applied ad hoc. One example of an approach that may be applied is the Data Asset Framework, which seeks to: identify data assets; explain how the data are stored, managed, shared, and reused; identify risks (e. g. misuse, data loss, irretrievability); understand researchers' attitudes and feed into future decisions about data management. The task is challenging: the DAF (Data Asset Framework) [JRR08] finds on the basis of pilot studies that it is "not feasible to be comprehensive" when attempting to inventory data, and that low rates of engagement may be expected. Audit criteria for training datasets and trained models that may be of interest in the context of ML research in particular include not only those points identified above, but also issues of licensing, ownership, database rights, sensitivity, transparency, and ethics. There has been considerable work put into the development of appropriate metadata, such as model cards describing trained ML models [Mi19], and data cards describing datasets [PZK22]. The issue is therefore not one of a foundational lack of understanding of the data to be collected, but one of capturing the appropriate data at the right time: it is difficult, for example, to establish the intended aims and appropriate use cases

for a dataset without detailed knowledge of the information one is holding, including the circumstances of its collection or capture, and it may be extremely difficult to establish its regulatory and consent status after the fact.

To summarise, as Paullada et al. [Pa21] observe and we have briefly discussed above, ML datasets are commonly bogged down by questions of licensing, intellectual property, data protection, the presence or absence of informed consent, transparency and ethics. It may be extremely impractical to thoroughly grapple with the task of auditing activity in this field when beginning *in medias res*. As the joke goes: "If I were you, I wouldn't start from here".

3 Discussion and Conclusion

Processes of digital curation, such as data appraisal and audit, data management and preservation, and data enrichment and long-term availability for reuse, are clearly relevant to AI/ML tasks. Some areas, such as dataset and model description, have extensive histories in the literature, and existing best practices in this area are ripe to become part of digital curation best practices. From a regulatory perspective, however, there are significant areas of concern surrounding which there is further work to do, such as the many issues with potential leakage of sensitive data from datasets and models, which have only briefly been covered in this paper, and the areas of concern surrounding dataset transparency and implementation of the data subject's rights under the GDPR (e. g. "right to be forgotten" and "right to rectification"). The availability of datasets for reuse raises the potential for issues such as inference of sensitive personal data without the consent of the data subject, and hence the dataset curator must grapple with the potential for harm linked to the reuse of these resources.

Much of the existing discourse surrounding safety in AI/ML relates to the archetypical large, wealthy organisation running extremely large and extensive models, rather than to research institutions and groups or small organisations or startups, and we argue that this is another example of "hype and wishful thinking", and that the majority of AI/ML datasets, beyond these few very large organisations, are in the care of digital curators, staff, or students with relatively few resources. Hence, we suggest that there is a need for digital curation guidance and resources tailored for the broader community.

In this paper we have sought to identify the different type of data one may use or generate when working with AI/ML, the multiple challenges paused for each type when applying data management methods to such data, as well as the different challenges around data preservation once projects are done. We have attempted to provide guidelines as to the different questions that should be asked and considered when when dealing with such workflows: data provenance and ownership; data enrichment; data sensitivity; copyright, intellectual properties and other legal issues; and data auditing and curation.

References

[ANS23] Appel, G.; Neelbauer, J.; Schweidel, D. A.: Generative AI Has an Intellectual Property Problem, 2023.

[Ba21] Banda, J. M. et al.: A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. Epidemiologia 2 (3), pp. 315–324, 2021.

[Bh24] Bhardwaj, E. et al.: Machine learning data practices through a data curation lens: An evaluation framework. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24, Association for Computing Machinery, Rio de Janeiro, Brazil, pp. 1055–1067, 2024.

[Bl24] Blease, C.: Open AI meets open notes: surveillance capitalism, patient privacy and online record access. Journal of Medical Ethics 50 (2), pp. 84–89, 2024.

[BNC21] Bishop, B. W.; Nobles, R.; Collier, H.: Research Integrity Officers' Responsibilities and Perspectives on Data Management Plan Compliance and Evaluation. Journal of Research Administration 52 (1), pp. 76–101, 2021.

[BS22] Babbar, S.; Stuart Aguiar, A. C.: Artificial intelligence and EU copyright law: a net of authorship claims, 2022, URL: https://www.nottingham.ac.uk/research/groups/commercial-law-centre/blog/artificial-intelligence-and-eu-copyright-law.aspx, visited on: 05/13/2024.

[BW23] Burtell, M.; Woodside, T.: Artificial Influence: An Analysis Of AI-Driven Persuasion, 2023.

[Da21] Daneshjou, R. et al.: Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. JAMA dermatology 157 (11), pp. 1362–1369, 2021.

[Fr24] Fried, I.: For AI firms, anything "public" is fair game, 2024, URL: https://www.axios.com/2024/04/05/open-ai-training-data-public-available-meaning, visited on: 05/14/2024.

[Ge23] Ge, W. et al.: Has machine learning over-promised in healthcare?: A critical analysis and a proposal for improved evaluation, with evidence from Parkinson's disease. Artificial Intelligence in Medicine 139, p. 102524, 2023.

[Gr24] Gross, G.: Air Canada chatbot error underscores AI's enterprise liability danger, 2024.

[GSI22] Gundersen, O. E.; Shamsaliei, S.; Isdahl, R. J.: Do machine learning platforms provide out-of-the-box reproducibility? Future Generation Computer Systems 126, pp. 34–47, 2022.

[Gu24] Gutiérrez, J. D.: Critical appraisal of large language models in judicial decision-making. In: Handbook on Public Policy and Artificial Intelligence. Edward Elgar Publishing, pp. 323–338, 2024.

[HHS24] Hicks, M. T.; Humphries, J.; Slater, J.: ChatGPT is bullshit. Ethics and Information Technology 26 (2), p. 38, 2024.

[Ho11] Horton, L. et al.: Data Management Recommendations for Research Centres and Programmes. 2011.

[HQ21] Hugenholtz, P. B.; Quintais, J. P.: Copyright and artificial creation: does EU copyright law protect AI-assisted output? IIC-International Review of Intellectual Property and Competition Law 52 (9), pp. 1190–1216, 2021.

[JRR08] Jones, S.; Ross, S.; Ruusalepp, R.: The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions. In: 5th International iPRES Conference: Joined Up and Working: Tools and Methods for Digital Preservation. London, England, 2008.

[KB23] Kerrigan, P.; Barry, M.: Algorithms, artificial intelligence and machine learning for gender and sexual minorities. Routledge Handbook of Sexuality, Gender, Health and Rights, 2023.

[KN23] Kapoor, S.; Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4 (9), 2023.

[Ku22] Kuiper, O. et al.: Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. In: Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10–12, 2021, Revised Selected Papers 33. Springer, pp. 105–119, 2022.

[Li21] Liao, T. et al.: Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). 2021.

[Li24] Liu, F. et al.: Exploring and evaluating hallucinations in llm-powered code generation. arXiv preprint arXiv:2404.00971, 2024.

[LLD23] Liesenfeld, A.; Lopez, A.; Dingemanse, M.: Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In: Proceedings of the 5th International Conference on Conversational User Interfaces. CUI '23, Association for Computing Machinery, Eindhoven, Netherlands, 2023.

[Lo17] van Loon, J. E. et al.: Quality evaluation of data management plans at a research university. IFLA Journal 43 (1), pp. 98–104, 2017.

[LS19] Leong, B.; Selinger, E.: Robot eyes wide shut: Understanding dishonest anthropomorphism. In: Proceedings of the conference on fairness, accountability, and transparency. Pp. 299–308, 2019.

[Ma24a] Ma, Z. et al.: Schrödinger's Update: User Perceptions of Uncertainties in Proprietary Large Language Model Updates. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. Pp. 1–9, 2024.

[Ma24b] Markelius, A. et al.: The mechanisms of AI hype and its planetary and social costs. AI and Ethics, pp. 1–16, 2024.

[Mi19] Mitchell, M. et al.: Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. Pp. 220–229, 2019.

[OKM12] Oelke, D.; Kokkinakis, D.; Malm, M.: Advanced visual analytics methods for literature analysis. In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Pp. 35–44, 2012.

[Pa21] Paullada, A. et al.: Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns 2 (11), 2021.

[PZK22] Pushkarna, M.; Zaldivar, A.; Kjartansson, O.: Data cards: Purposeful and transparent dataset documentation for responsible ai. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Pp. 1776–1826, 2022.

[Rh24] Rhee, H. L.: A New Lifecycle Model Enabling Optimal Digital Curation. Journal of Librarianship and Information Science 56 (1), pp. 241–266, 2024.

[Ru24] Ruschemeier, H.: Generative AI and Data Protection. 2024.

[Se23] Semmelrock, H. et al.: Reproducibility in Machine Learning-Driven Research. arXiv preprint, 2023.

[SJ18] Shelly, M.; Jackson, M.: Research data management compliance: is there a bigger role for university libraries? Journal of the Australian Library and Information Association 67 (4), pp. 394–410, 2018.

[Sp21] Spichtinger, D.: Data management plans in Horizon 2020: what beneficiaries think and what we can learn from their experience. Open Research Europe 1, 2021.

[St24] Steffens, S. et al.: The challenges of research data management in cardiovascular science: a DGK and DZHK position paper—executive summary. Clinical Research in Cardiology 113 (5), pp. 672–679, 2024.

[Su20] Sunray, E.: Sounds of science: Copyright infringement in AI music generator outputs. Cath. UJL & Tech 29 (2), p. 185, 2020.

[Wa24] Wang, B. et al.: DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models, 2024.

[Wi17] Wilson, T. C.: Rethinking digital preservation: definitions, models, and requirements. Digital Library Perspectives 33 (2), pp. 128–136, 2017.

[WK18] Wang, Y.; Kosinski, M.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology 114 (2), p. 246, 2018.

[WRZ22] Werder, K.; Ramesh, B.; Zhang, R.: Establishing data provenance for responsible artificial intelligence systems. ACM Transactions on Management Information Systems (TMIS) 13 (2), pp. 1–23, 2022.