Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

THE POLITICS OF MACHINE-LEARNING EVALUATION: FROM LAB TO INDUSTRY

Anna Schjøtt Hansen
University of Amsterdam

Dieuwertje Luitse
University of Amsterdam

Introduction

Artificial Intelligence (AI) applications are today implemented across various societal sectors, ranging from health care and security to taking part in shaping the media environment we encounter online. In the last decade there has been a significant shift in the field of AI, as the development of AI applications is no longer confined to the laboratory, but rather widely used and tested in and on societies (Whitaker et al., 2018; Seaver, 2019: Marres and Stark, 2020; Sheik et al., 2023; van der Vlist et al. 2024). With this rapid industrialisation of AI, there is an increased need to understand the implications of both the development and deployment of these systems. While critical scholars have started to scrutinize different components of AI development, such as dataset construction and annotation practices (Miceli et al., 2021; Miceli and Posada, 2021; Paullada et al., 2021; Orr and Crawford, 2023), the study of evaluative practices in AI has received limited attention. A few studies have highlighted the importance of benchmarking practices and how these methods become integral in establishing the validity of the system and its success, which then enables widespread application (Jaton, 2017; 2023; Raji et al., 2021; Grill, 2022). This paper presents a research agenda that outlines how to study machine-learning evaluation practices that move beyond the laboratory into industry applications and standardised validation practices. Based on emerging research and illustrative empirical examples from recent fieldwork, we argue to study machine-learning evaluation as a sociotechnical and political phenomenon that requires multi-level scrutiny. Therefore, we provide three analytical entry points for future research that address the political dynamics of (1) standardised validation infrastructures, (2) the circulation of evaluation methods and (3) the situated enactment of evaluation in practice.

Suggested Citation (APA): Hansen, A. S., Luitse, D. (2024, October). *The Politics of Machine-Learning Evaluation: From Lab to Industry*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

Evaluation is a foundational part of the development of AI systems because it requires "reflection on what the objectives of an effective model involve, in addition to conscious decision making on how to best represent these desired outcomes in evaluation metrics, data and methodology" (Raji et al., 2021, p. 3). In practice, this often involves developing metrics that can best approximate and measure whether the system achieves the anticipated behaviour (Aradau and Blanke, 2021; Raji et al., 2021). As also highlighted by Thomas and Uminski, (2020): 'what most current AI approaches do is to optimize metrics' (p. 1). Taken together, this illustrates how evaluation is a highly political practice, where the decisions will often favour 'good enough' optimisation methodologies (Amoore, 2020; Thomas and Uminsky, 2021). The political nature of 'good enough' optimisations has been highlighted by authors, pointing out how this enables the construction of validity (Raji et al., 2021) and the production of certainty through accuracy measurements (Grill, 2022). In addition, Thakkar et al., 2022 highlight how the construction of machine-learning datasets is also guided by questions of 'good-enough', while Jaton (2017) has focused on the key role of so-called 'ground truths,' in both training and evaluating AI applications through randomized tests to assess their performance.

However, these studies have been predominately concerned with the established 'scientific' evaluative practices of AI research and development and do not consider that AI applications and their evaluative methods are increasingly becoming industrialised. In light of these developments, this paper aims to expand the current research on evaluation by building on emerging insights from the cited critical enquires and illustrative empirical examples from ongoing fieldwork. For the latter, we draw on examples from a recent ethnographic enquiry within the BBC's data science department (conducted by author 1) and ongoing fieldwork on the development of AI applications at a Dutch medical centre (conducted by author 2).

Taking a Critical AI Studies and Science and Technology Studies (STS) research approach, we outline how to study what we call industrialised accounts of AI evaluation beyond the laboratory. Concretely, we present three analytical entry points that address the sociotechnical and political nature of AI evaluation. These entry points focus on different levels of AI evaluation in industry settings that include wider attempts of standardised evaluation infrastructures, the circulation of evaluative methods and the situated practice of evaluating emerging AI applications within localised contexts. In what follows, we briefly discuss these entry points and substantiate them with empirical illustrations from existing research and our ongoing fieldwork.

The Political Dynamics of Standardised Validation Infrastructures

The first entry point engages with the emergence of wider attempts at standardisation of evaluation in the AI industry via large-scale infrastructural agents that shape AI development and evaluation regimes. One example of such an infrastructure is the AI competition platform Kaggle, which facilitates the evaluation of new models across domains. Such competition platforms participate in setting the conditions for evaluation by providing a pre-specified list of metrics for competitions hosted on their platforms. As these competition platforms are gaining increased popularity this homogenises the types of tasks and associated metrics that are used to develop new AI applications, while at the same time concentrating power amongst a small group of actors (Luitse et al., 2024).

Political Dynamics of Circulating Methods of Evaluation

With the second entry point, we wish to bring attention to the role of specific methods of evaluation that circulate. For example, via ‘best practice’ metrics for certain tasks or the functionalities of AI development platforms and development frameworks, such as Sagemaker or TensorFlow. Building on the work of Selbst et al. (2019), we argue for the need to study the ‘portability’ of metrics into different domains and how that shapes the development and evaluation of AI applications. Luchs et al. (2023), for example, illustrate how machine learning frameworks such as TensorFlow produce ‘step-by-step' approaches to AI development including methodologies for iterative and gradual optimisation of the model (i.e. evaluation). In addition, developers at the Dutch medical centre, highlighted that the commonly used developer framework PyTorch comes with specific machine-learning pipeline requirements, which again shapes how the application is developed and evaluated.

The Political Dynamics of Situated Enactments of Evaluation

With the third entry point, we follow Jaton and Sormani’s (2023) recent call for situated accounts of AI development and highlight the need to study the localised enactments of evaluation. Such concrete case studies (Flyvbjerg, 2011) can help illustrate the tensions between metrics and contextual understandings of what makes the system ‘good enough’ to be deployed. Schjøtt and Birkbak (2022), for example, illustrate that the failure of a recommender system for news could not solely be ascribed to its lack of providing accurate predictions, but rather the opposite. Instead, it was the editor’s need to protect the public interest that ultimately led to the decision not to deploy it on the news sites. The fieldwork at the BBC also illustrated how many local methods of

evaluation were put in place, such as visualisation tools, to enable qualitative feedback from editorial staff. However, these tools also configure the development of AI systems by making certain features of the system salient, while also abstracting away other ways of understanding the system.

References

Amoore L (2020) Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. Duke University Press.

Aradau C and Blanke T (2021) Algorithmic Surveillance and the Political Life of Error. *Journal for the History of Knowledge* 2(1). 1: 10–10.

Flyvbjerg B (2011) 'Case Study', in N.K. Denzin and Y.S. Lincoln (eds) *The SAGE Handbook of Qualitative Research*. Thousand Oaks: SAGE, pp. 301–316.

Grill G (2022) Constructing Certainty in Machine Learning: On the performativity of testing and its hold on the future. OSF Preprints. Available from: osf.io/zekqv.

Jaton F (2017) We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811–840.

Jaton F and Sormani P (2023) Enabling 'AI'? The situated production of commensurabilities. *Social Studies of Science* 53(5): 625–634.

Marres N and Stark D (2020) Put to the test: For a new sociology of testing. *The British Journal of Sociology* 71(3): 423–443.

Miceli M and Posada J (2021) Wisdom for the Crowd: Discoursive Power in Annotation Instructions for Computer Vision. arXiv:2105.10990. arXiv. Available at: http://arxiv.org/abs/2105.10990 (accessed 26 February 2024).

Miceli M, Yang T, Naudts L, et al. (2021) Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada, 3 March 2021, pp. 161–172. ACM. Available at: https://dl.acm.org/doi/10.1145/3442188.3445880

Luchs I, Apprich C and Broersma M (2023) Learning machine learning: On the political economy of big tech's online AI courses. *Big Data & Society* 10(1). DOI: 10.1177/20539517231153806 .

Luitse D, Blanke T, and Poell T. (2024). AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, 1–22. https://doi.org/10.1080/1369118X.2024.2334393

Orr W and Crawford K (2023) The social construction of datasets: On the practices, processes and challenges of dataset creation for machine learning. SocArXiv. Available from: osf.io/preprints/socarxiv/8c9uh.

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

Raji D, Denton E, Bender EM, et al. (2021) AI and the Everything in the Whole Wide World Benchmark. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (eds J Vanschoren and S Yeung), 2021. Curran. Available at: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Seaver N (2019) Knowing Algorithms. In: *Knowing Algorithms*. Princeton University Press, pp. 412–422. Available at: https://www.degruyter.com/document/doi/10.1515/9780691190600-028/html.

Selbst AD, Boyd D, Friedler SA, et al. (2019) Fairness and Abstraction in Sociotechnical Systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA, 29 January 2019, pp. 59–68. ACM. DOI: 10.1145/3287560.3287598.

Sheikh H, Prins C and Schrijvers E (2023) AI Is Leaving the Lab and Entering Society. In: Sheikh H, Prins C, and Schrijvers E (eds) *Mission AI: The New System Technology*. Research for Policy. Cham: Springer International Publishing, pp. 43–84. Available at: https://doi.org/10.1007/978-3-031-21448-6_3 (accessed 28 February 2024).

Schjøtt A and Birkbak A (2022) You cannot make the algorithm do something it does not want to do: exploring the agency and power of AI systems in practice. In: *EASST 2022*, 2022. Available at: https://easst2022.org/programpreliminary7.asp

Thakkar D, Ismail A, Kumar P, et al. (2022) When is Machine Learning Data Good?: Valuing in Public Health Datafication. In: *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA, 29 April 2022, pp. 1–16. ACM. Available at: https://dl.acm.org/doi/10.1145/3491102.3501868 .

Thomas RL and Uminsky D (2022) Reliance on metrics is a fundamental challenge for AI. *Patterns* 3(5). Elsevier. DOI: 10.1016/j.patter.2022.100476

van der Vlist F, Helmond A, and Ferrari F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241232630

Whittaker M, Crawford K, Dobbe R, et al. (2018) AI Now Report 2018. Available at: https://ainowinstitute.org/aiareport2018.pdf