AI Research is not Magic, it has to be Reproducible and Responsible: Challenges in the AI field from the Perspective of its PhD Students

Andrea Hrckova[1[0000−0001−9312−6451]], Jennifer Renoux[2[0000−0002−2385−9470]], Rafael Tolosana Calasanz[3[0000−0003−3057−6273]], Daniela Chuda[1,4[0000−0002−3873−9308]], Martin Tamajka[1[0000−0002−0107−6459]], and Jakub Simko[1[0000−0003−0239−4237]]

[1] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
andrea.hrckova@kinit.sk
[2] Center for Applied Autonomous Sensor Systems, Örebro University, Sweden
jennifer.renoux@oru.se
[3] I3A, University of Zaragoza, Spain
rafaelt@unizar.es
[4] Slovak University of Technology in Bratislava, Slovakia
daniela.chuda@kinit.sk

Abstract. Unlocking the full societal potential of artificial intelligence demands a fundamental shift towards responsible and reproducible research. Understanding that PhD students are pivotal in conducting and reproducing experiments, we investigated the challenges of 28 AI PhD candidates from 13 European countries. We identify three critical areas where current practices fall short: (1) the findability and quality of AI resources such as datasets, models, and experiments; (2) the difficulties in replicating the experiments in AI papers; (3) and the lack of trustworthiness and interdisciplinarity. After uncovering some of the underlying reasons behind the challenges, we propose a combination of social and technical recommendations to overcome the identified challenges and foster a more transparent and reliable AI research ecosystem. Socially, we recommend the general adoption of reproducibility initiatives in AI conferences and journals, as well as improved interdisciplinary scientific collaboration, especially in data governance practices. On the technical front, we call for enhanced tools to better support versioning control of datasets and code, and a computing infrastructure that facilitates the sharing and discovery of AI resources, as well as the sharing, execution, and verification of experiments.

1 Introduction

AI research is facing a reproducibility crisis [7]. Worse, AI systems and methods are increasingly being used to perform research in other fields, and concerns arise that it may lead to another major crisis in science in general, as scientists may use non-reproducible and non-responsible AI systems in an ill-informed way [6].

In this work, we use the term *reproducibility* as *obtaining consistent results using the same input data; computational steps, methods and code; and analysis conditions* [18]. Responsible AI is the practice of developing and using AI systems in a way that benefits society while minimizing the risk of negative consequences. The goal is to employ AI in a safe, legal, trustworthy and ethical way [1]. This paper focuses on these pillars of responsible AI: privacy and data governance, fairness (to researchers as AI stakeholders), sustainability and reliability [3]. We also consider reproducibility as a foundational component of responsible AI as its absence prevents verification of model behavior, fairness, or ethical compliance, thereby undermining broader responsible AI goals. Several factors fuel AI's reproducibility crisis and impede responsible AI development. A major concern is data quality, which significantly impacts model performance [27]. Some researchers point out that data challenges are often overlooked in machine learning, criticizing inadequate practices in data annotation and documentation [20]. While the importance of data quality is widely acknowledged, and best practices and recommendations for scientific data management are summarized in the FAIR principles [28], challenges related to data quality and reproducibility persist, as documented in [13] and supported by our own research. To shed light on the persistent issues with reproducible and responsible AI research, we propose a complementary approach to previously conducted analyses [4, 11]: investigating the ground-level practices and hurdles encountered by PhD candidates. These early-career researchers are pivotal in conducting and reproducing experiments, making them uniquely exposed to reproducibility hurdles. Their point of view is therefore invaluable for understanding the true causes and consequences of reproducibility crisis. Our study of 28 European AI PhD students uses qualitative methods from information science. Prior work on information interaction (or seeking behavior) has examined the needs of scientists [12], PhD students [23], and computer scientists [5]. While these studies focused on specific groups and tool recommendations, our exploratory and interdisciplinary approach offers broader insights for AI, leading to general recommendations on reproducibility and responsible AI. The main contributions of this paper are:

1. We explore and express the issues encountered by European PhD students in AI field, many of which are particularly related to reproducible and responsible AI. We identify the main sources of difficulties, which include, but are not limited to, the quality of AI resources (datasets, code, and models).

2. We formulate recommendations to address these issues, highlighting, among others, (1) the need for adoption of reproducibility practices on every level, particularly in AI journals and conferences, (2) the need for improved interdisciplinarity in AI, especially greater participation of domain experts, data and information specialists, ethicists, and legal professionals (notably in data governance practices).

2 Methodology

2.1 Study design and methods

This study employed an exploratory qualitative design, based an existing model of 7 stages [15], to investigate challenges faced by AI PhD students. This research process helped us manage the uncertainty and ambiguities that arise from a qualitative and exploratory investigation. The overarching research question in the *origination stage* of the research was: *'What are the biggest challenges that PhD students face when conducting AI research?'*. The first, fourth, and fifth author led this initial and the next phase. During the *orientation stage*, preliminary research led the team to focus on identifying technological recommendations for AI researchers. At this point, the planned interview duration was also shortened from four hours to 1.5-2 hours. We collected data via semi-structured, in-depth focus group interviews during the *exploration stage*. The first and fourth author drew from a set of 48 open-ended questions, selectively posing them based on the respondents' expertise. The full list of questions can be found under this link [5]) The *elucidation stage* focused on analyzing the collected data through manual content analysis. We employed an inductive coding process, which yielded 32 initial codes for problems in the AI field. Following iterative refinement, five unsaturated codes were excluded and nine were merged, culminating in 19 final categories that were all used in mind map visualizations[6]. The first author conducted the initial coding, with validation by the fourth and fifth author. The *consolidation stage* involved brainstorming sessions with all authors to create stronger, more consistent connections between our findings and existing state-of-the-art research, particularly concerning reproducible and responsible AI research. Ultimately, this phase aimed to develop a coherent and easily understandable final outcome. In the *reflection stage*, we took a broader view of the results, aiming to contextualize the findings within a larger picture of AI research challenges. During this phase, the study's inductive results were also controlled by the first and second author to enhance the validity of the findings. During the *culmination stage*, the authors formulated recommendations, recognizing that implementing change within established research institutions is far more complex than applying the technological solutions initially proposed in the research's earlier stages.

2.2 Participants

Our study involved 28 PhD students (19 males, 9 females) representing 13 European countries, including Slovakia, Greece, Germany, Spain, Sweden, Ireland, France, Bulgaria, Belgium, Norway, Finland, the Czech Republic, and Ukraine. We conducted 11 semi-structured focus group interviews from February to June 2023. One focus group involved up to four participants from different AI fields.

[5] https://zenodo.org/records/15920758
[6] Mind map visualizations were made in Canva.com

Participants were recruited through European project partners without any financial compensation. Most of interviews took place online via Google Meet in English. One interview was conducted face-to-face in the native language, and a single session with two interviewees was completed in written form upon their request. Participants demonstrated diverse expertise, covering ten AI focus areas such as machine learning (including neural networks, deep learning, federated learning, and accelerators), natural language processing (NLP), explainable AI, ethical AI, computer vision, recommenders, robotics, multimodal processing, and human activity recognition. Their research also extended into four additional domains: AI security (cyber-attacks), medical data processing, AI in business, and AI in energy and green environments (specifically, time-series analysis). The cohort included 16 advanced-stage and 12 early-stage PhD students.

3 Findings

Most of the challenges that we identified concern the quality of available AI resources, including datasets, code, and models; therefore, we grouped them into these three categories. These identified difficulties consumed most of the time of these researchers and significantly prevented them from being able to replicate the results in papers. As respondent P1 stated:

> *"Sometimes you don´t understand how they gained such results in the paper and how to replicate it."*

Yet, replication studies are the type of research in which these early researchers are most involved. Sections 3.1, 3.2, and 3.3 explore these issues in more detail. However, it is worth noting that we also uncovered another set of challenges that PhD students encounter during AI research, particularly concerning the absence of interdisciplinary collaboration, issues with information retrieval, concerns regarding the lack of human involvement, and a lack of motivation to share research findings with the public. These challenges were named "Other challenges regarding AI research process" and are explored in Section 3.4.

3.1 Quality and findability of datasets

The foremost challenge frequently discussed in our interviews was the quality and utility of datasets required for the training of AI models (6 groups out of 11, see Fig. 1). This challenge stems partly from the significant time investment required to curate datasets annotated by human annotators. Some respondents (6 groups out of 11) noted difficulties in accessing experts, low agreement between annotators (2 groups out of 11), or the necessity to involve themselves or the not-so-motivated students to annotate (2 groups out of 11). However, the main concern was the privacy issue when trying to publish a dataset (6 groups out of 11). Respondent P2 illustrates this problem:

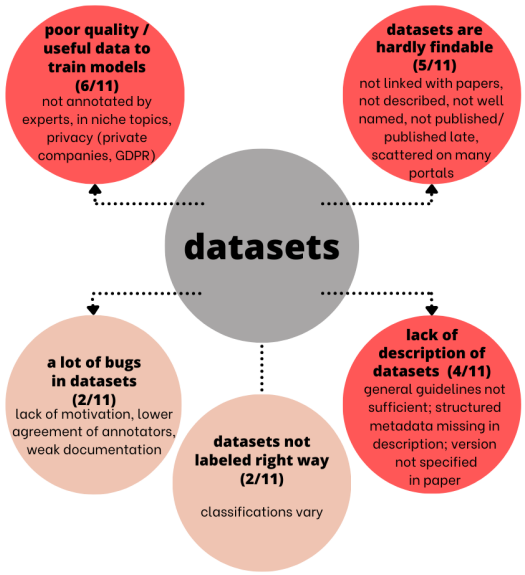

Fig. 1. Dataset-related challenges of AI resources as reported by research participants. The darker the red, the more often was the problem mentioned during the focus groups. The numbers in brackets indicate the count of focus groups out of the total 11 groups that agreed on such issues.

> *"There is just one public dataset of breast cancer that was properly labeled by experts with data about patients and everybody is using it... Obtaining approvals is difficult from patients."*

The challenge of accessing quality datasets extends beyond the medical domain. Several respondents encountered obstacles due to privacy restrictions imposed by their companies, preventing them from publishing papers despite having ample data from their workplace. This restriction poses a significant hurdle, as publishing without accompanying data is problematic. Our respondents expressed skepticism about anonymization as a solution, citing doubts about its effectiveness and the impracticality of individually informing each affected individual. Five groups agreed that locating good datasets can be a challenge, even when they do exist. Many datasets lack direct links to associated papers for various reasons, such as delayed publication or non-publication. Consequently, these datasets are often dispersed on multiple data storage platforms. Additionally, authors may not always consider the discoverability of their datasets when publishing, neglecting to provide clear and descriptive names or descriptions. Clear and comprehensive descriptions in introduction, metadata, and classification are essential to ensure that a dataset is not only easily discoverable, but also allows researchers to assess its relevance to their specific requirements. For a PhD student in AI, the process of repeatedly downloading and opening avail-

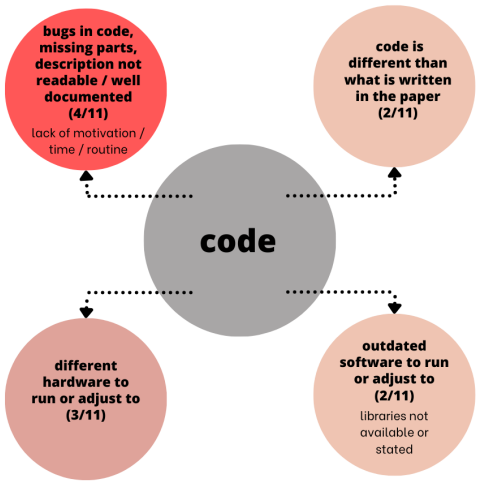

Fig. 2. Code-related challenges of AI resources as reported by research participants.

able datasets to discern their contents consumes a significant amount of time. An example of useful metadata to save their time would include specifying the version of the dataset used in the paper, given the multitude of available dataset versions. Although general guidelines can help to describe datasets to some extent, respondents (for example, those in the security sector) expressed concern that they may not be directly applicable in specific domains.

In addition to that, four groups of PhD students out of 11 expressed their frustration about numerous bugs present and not documented in the available datasets, exemplified by respondent P3:

> *"Papers do not mention data drifts, there is usually no information about the data preparation."*

Apparently, the issue is shifting from availability to quality of the dataset. An exception was seen in niche topics that still suffer from a shortage of data for training models.

3.2 The quality of code

Quality issues also arise concerning the code (Fig. 2). Early career researchers typically have a practice of publishing code (just one student mentioned the unavailable code as a problem). Our respondents were driven by inner motivations such as a sense of reciprocity or awareness of reproducibility concerns. As respondent P4 pointed out:

> *"If a failure is identified, it is good for science, even if it is embarrassing for an individual. "*

Nonetheless, this does not mean that the code published is clean or well documented. That is the reason why several students (4 groups out of 11) complained about serious bugs in the code or missing parts of the code. Respondent P5 highlights an additional unpleasant problem - a mismatch between the code and the paper, a concern echoed in two focus groups:

> *"The biggest problem comes, if you are trying to make the paper work. Also with access to the code you get specific results. Sometimes the code is different than what is written in the paper and you have to do double work. "*

At the same time, respondents in these groups acknowledged that they lack motivation to publish meticulously documented code, as it is not a mandatory aspect of the review process and requires significant time investment, necessitating a consistent routine for continuous documentation during code writing and editing. In addition to that, these respondents emphasized that having some code available is better than having none at all. Nevertheless, researchers would appreciate at least some information on the program versions, the required hardware, and the libraries that were used when the code was originally run. This information would save researchers a lot of time, as they quite often encounter challenges related to incorrect hardware (3 groups out of 11) or software configurations (2 groups out of 11) necessary to run and adapt the code. As noted, research papers often lack the space to address such issues.

3.3 Benchmarking and quality of models

Some AI models face comparable challenges with respect to replicability and reproducibility as code (Fig. 3). Three groups of PhD candidates highlighted discovering significant flaws in the models presented in the papers, such as absent or inaccurate hyperparameters, along with inadequate documentation. These shortcomings, compounded by rigidly defined hyperparameters, hindered the utility of published models beyond their intended paper applications (even 5 groups of respondents out of 11 reported this problem). Respondent P5 aptly illustrated the issue by questioning the quality of certain published AI models:

> *"It even happened that model was leaking labels, they were entering the input of the model. And that was not raised as an issue, the paper is still there. You cannot trust the code blindly; otherwise, you repeat the same mistakes."*

Despite the existence of Papers with Code, which 6 out of 11 groups of students found helpful in some cases, effectively utilizing benchmarking tools remains a challenge, as pointed out by 2 groups of PhD students. This difficulty arises from the inherent difficulty of comparing different models, compounded by the vague nature of benchmarks. Consequently, results from automated benchmarking lack reliability and interpretability, as each model or user may define their parameters and occasionally engage in deceptive practices, such as training on testing data or comparing with the weakest model.

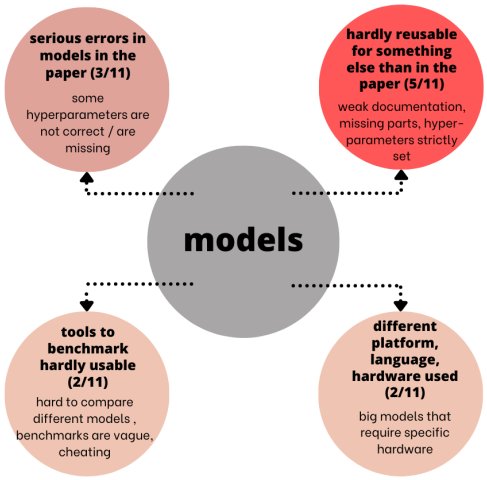

Fig. 3. Model-related challenges of AI resources as reported by research participants.

3.4 Other challenges regarding AI research process

Our approach allowed us to identify bottlenecks throughout the research process, addressing not only reproducibility challenges but also a spectrum of issues encountered by PhD students. These include technical challenges and managing expectations related to both AI and PhD research (both mentioned in 3 groups out of 11), information retrieval, and problems with publication and dissemination (both mentioned in 5 groups out of 11). While not the primary focus of this paper, these issues are significant aspects of the research journey. (Fig. 4).

These challenges often require collaboration with others. Despite some mistrust in online intra-disciplinary communication, we recognized the need for a scientific interdisciplinary discourse involving experts well-versed in the application domain. Researchers who initiated their research based on their own ideas expressed a desire for increased opportunities to exchange and compare problems while brainstorming with experts from various fields. When such cooperation occurred, it facilitated the advancement of their research. However, researcher P6 encountered difficulties in finding suitable communication channels:

> *"I store many ideas in my personal document and it is difficult to find someone [outside the field] to communicate with."*

We consider another significant concern in AI research to be the lack of human involvement. This concern is closely tied to trustworthiness, primarily regarding accountability requirements. It also intersects with stakeholder participation, as outlined in the "Diversity, Non-discrimination and Fairness" requirement within ALTAI (Assessment List for Trustworthy Artificial Intelligence) [2]. Young AI researchers seemed to overlook ethical assessments and rarely involve

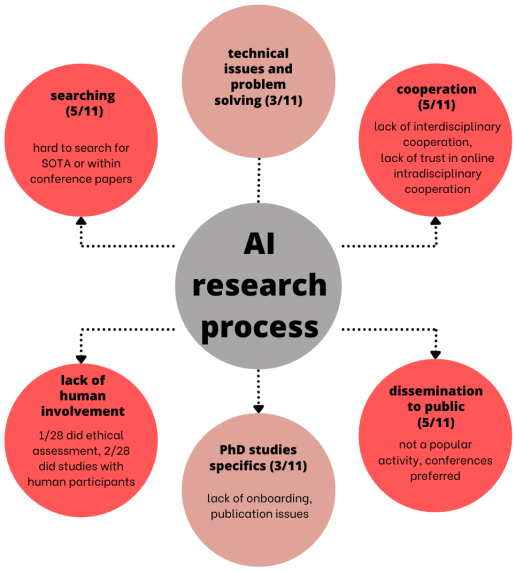

Fig. 4. Problematic parts of AI research process as reported by research participants.

human participants in their studies. *De facto*, none of the AI researchers conducted any additional ethical assessments beyond what was required for the ethical sections of the journals. This issue was corroborated by the experience of an AI ethics specialist that has been helping the AI community with these kind of assessments (P8):

> *"I was surprised by the lack of care of computer scientists about the ethical issues and information about data."*

The limited understanding of ethical assessments and a lack of standardized frameworks contribute to this issue, as noted by an AI ethics expert who assists researchers with crucial questions regarding data collection, ownership, and storage. Nevertheless, rigorous and standardized ethical assessments are not merely a technical requirement but a fundamental humanistic imperative in AI development. Furthermore, human participant engagement in AI research was minimal, with only two out of 28 respondents investigating stakeholder requirements, despite many focusing on areas where human involvement would be beneficial. Involving end-users and domain experts early in the AI development lifecycle (e.g., through co-design or participatory design) ensures that AI solutions are relevant, usable, trustworthy, less biased, and address real-world problems from a human perspective, rather than being purely technology-driven.

4 Perspectives and recommendations

From our findings, it appears that three main areas of improvements are possible for better AI research: (1) Discoverability and quality of AI resources. (1) Reproducibility of experiments (1) Trustworthiness and interdisciplinarity

4.1 Discoverability and quality of AI resources

Locating AI resources, such as code, datasets, or models, can be challenging. Our findings highlight that the links between papers and AI resources are often broken once a paper is accepted, hindering long-term accessibility and utility. Existing platforms such as Papers with Code[7] while widely used (researchers in 6 groups out of 11 mentioned using it occasionally), have proven insufficient. Despite their theoretical aim to alleviate these issues, we observed they often fall short due to a lack of quality assurance policies for the resources they host. Beyond the discoverability problem, the AI resources that our respondents examined consistently exhibited significant quality issues. We argue these problems stem from two primary causes: (A) insufficient support for researchers to perform necessary AI resource curation, and (B) a lack of incentive or recognition for this crucial work. Researchers could greatly benefit from research data curation services, which encompass essential tasks like data cleaning, validation, metadata provision, and consultations, similar to those already integrated into established institutional repositories [14]. The imperative for proper documentation of AI resources have been addressed by numerous papers and initiatives, including but not limited to the "Datasheets for Datasets" [10], and Model Cards for Model Reporting[16]. As the practices and needs highly depend on the domain, papers summarizing the optimal data curation practices for machine learning also emerge, for example in security [24] or medicine [9]. However, as noted by [22], there is still much work to be done. Institutionalizing collaborations among domain experts, lawyers, ethicists, and data/code curators is vital. Research organizations must provide guidelines and encourage standardized documentation practices, which will not only enhance the quality of AI resources but also improve the reproducibility of experiments (see Recommendation 3).

Recommendation 1

The AI research community needs to foster interdisciplinary collaboration for the governance of AI resources.

4.2 Reproducibility of experiments

Our findings show that early-career researchers frequently encounter significant hurdles in reproducing AI experiments. This issue stems from a combination of technical problems, such as bugs in code or datasets and discrepancies between

[7] https://paperswithcode.com/

papers and associated resources, as well as social factors like inadequate peer review and difficulties with accessing or configuring hardware and software. To overcome these challenges, the AI community should draw inspiration from fields like physics by adopting stricter reproducibility policies.

Recommendation 2

The AI research community should embrace reproducibility practices widely and enforce stricter reproducibility policies in journals and conferences.

A small number of conferences and journals in computer science have taken steps to embrace reproducibility initiatives *in practice*. IEEE Transactions on Parallel and Distributed Systems [8] is the first IEEE journal to pilot a reproducibility initiative. Conferences in computer science with reproducibility initiatives include ASPLOS [9], ACM SIGPLAN [10], ACM SIGMOD [11], or SC [12], and more recently in AI, MLSYS [13], or NeurIPS [14]. These initiatives aim to enhance research reproducibility primarily through reproducibility evaluations. Reproducibility evaluations involve dedicated tracks for accepted papers. During these evaluations, authors of accepted papers submit a computational resource containing all necessary components for replicating their experiments, including datasets, code, and scripts. These efforts go beyond simple code and data sharing, including mechanisms for experiments packaging (e.g., using containers or virtual machines to manage software dependencies), dataset curation and documentation, and thorough process documentation. Successful evaluations often result in the awarding of reproducibility badges, recognized by publishers like ACM or IEEE, which signify the availability of resources and the successful reproduction of results by reviewers. Beyond conference and journal initiatives, individual research organizations can also play a vital role. For instance, one robotics laboratory of one of our respondents implemented mandatory coding guidelines and employed dedicated staff member to maintain and debug code repositories, significantly improving reproducibility and collaboration within the lab by enabling efficient reuse of code packages.

Recommendation 3

Research institutions and laboratories should set up guidelines and practices for their researchers and provide adequate resources (time and human) to ensure quality of AI resources and reproducibility of experiments.

[8] https://www.computer.org/csdl/journal/td/write-for-us/104303
[9] https://www.asplos-conference.org/
[10] https://www.sigplan.org/
[11] https://sigmod.org/
[12] https://sc24.supercomputing.org/
[13] https://mlsys.org/
[14] https://neurips.cc/

Despite the growing adoption of reproducibility initiatives, several technological challenges remain, particularly concerning hardware dependencies. Reproducing certain AI experiments, like those involving constraint programming, often demands highly specific hardware that current editorial policies don't support. Another significant hurdle is the substantial time and energy required to compile, install, deploy, and execute experiments. This poses a problem for reviewers facing strict deadlines and raises concerns about the energy consumption if authors pre-run experiments prior to reviewers for verification. A promising solution to these challenges is a cloud federation [26, 8]. This concept allows AI researchers to efficiently share both AI resources and computational resources from various providers. Central to this approach is a metadata catalogue, designed using FAIR principles (Findability, Accessibility, Interoperability, Reusability), which facilitates navigation among datasets, AI models, papers, and experiments. Furthermore, existing tools and services used by reproducibility initiatives are often not well-suited for AI, imposing a significant burden on researchers. For example, while code is typically uploaded to version control systems like GitHub and paired with Zenodo for DOIs, GitHub lacks support for different datasets versions. This forces authors to use additional tools and services, such as open dataset repositories, to manage dataset versions and DOIs. There's a clear need for new, AI-specific tools that better align with the unique requirements of AI research.

Recommendation 4

The AI research community should investigate the possibility to create a cloud federation for AI systems and suitable tools for control versioning of AI experiments, including datasets, models and code altogether.

4.3 Trustworthiness and interdisciplinarity

The third issue that our research uncovered goes beyond the technical aspects of AI research but more related to interpersonal and institutional relations, and the lack of interdisciplinarity in the field. First, we observed that early-career researchers rarely include end-users or human participants in their experiments, despite many of them working on topics that would benefit from such practices. The research of PhD students is often very techno-centered: the goal is to produce a working algorithm or system, and the analysis of factors other than performance is often forgotten (for instance user experience, inclusion, ...). Event though methods exist and are considered best practices in other fields, such as Codesign [25], they are rarely included in the field of AI research. This may be a direct consequence of the second issue that our research highlighted: the lack of interdisciplinary collaboration for early-career researchers. Early-career researchers are most often left alone (or with their supervisors) to conduct their research and lack the means to contact or discuss with other researchers, especially outside their institutions or main field of research. Our research also indicates that existing public virtual communities are often not conducive to idea

sharing among researchers due to concerns about idea theft. Similarly to what we recommended for quality and reproducibility (Recommendation 3), institutions have a role to play in encouraging multidisciplinarity. One starting point would be for them to diversify their research staff, for instance with lawyers, ethicists, domain experts, UX or HCI specialists, or information scientists. For institutions in which this diversity already exists (for instance, universities often have several schools that employ researchers in these different domains), they should ensure robust institutional support for interdisciplinary meetings, collaboration and hands-on research endeavors.

Recommendation 5

Research institutions should support interdisciplinary collaboration by ensuring a large diversity of research staff and providing the means for researchers to collaborate efficiently.

5 Conclusions and discussion

Good research is collective and multidisciplinary, building on prior work to advance knowledge. In order to produce good research efficiently, this prior work must be usable - discoverable, reproducible, and responsible for researchers. Our study of AI PhD students shows this is not yet the case. Despite growing awareness and initiatives, more effort is needed to improve AI research practices. Efforts must be both technological—developing platforms for discoverability and reproducibility—and societal—encouraging multidisciplinarity and trustworthiness. To this extent, we proposed five recommendations for the AI research community and institutions to consider in order to improve the research process. We acknowledge systemic barriers that hinder adoption of our recommendations and note that addressing them lies beyond this paper's scope. We believe overcoming such barriers requires collective action, particularly from institutions with the power and resources to support and incentivize researchers. Our study is limited to European PhD students and a moderate sample size, restricting quantification. Broader validation is needed to assess these challenges globally. Notably, reproducibility challenges in AI mirror those in fields like psychology and neuroscience, which have responded with reforms such as preregistration, data and code sharing, and large-scale replications [19, 17]. AI research, however, studies artifacts created by researchers themselves, suggesting repeatability should be easier since models, code, and data can be shared [11]. Yet, as our study and prior work show, sharing alone is insufficient; PhD students highlight further efforts needed to reach this baseline. Finally, while some work has studied the impact of reproducibility (e.g., citation counts [21, 29]), our work highlights the effects of irreproducibility on the AI community—an underexplored area deserving more attention.

Acknowledgments. This work was supported by the European Commission via Horizon projects AI4Europe (Grant 101070000) and Low Resource Artificial Intelli-

gence (Grant 101136646); by the Government of Aragon (Group Ref. T64_20R, COSMOS research group); and as part of project PID2020-113037RB-I00, funded by MICIU/AEI/10.13039/501100011033. Although the ethics committee was not yet operational at the interviewing institution, the research adhered to all applicable national and international ethical guidelines and regulations for studies involving human participants. Informed consent was obtained and all personal data was anonymized to protect sensitive information and ensure interviewee anonymity; therefore, interview data cannot be openly published. The authors are solely responsible for all content and ideas, with AI tools (ChatGPT and Gemini) used only for language refinement. We also thank Miroslav Blstak for recording and note-taking of the interviews.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Building a responsible AI: How to manage the AI ethics debate, https://www.iso.org/artificial-intelligence/responsible-ai-ethics
2. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future (Jul 2020), http://bit.ly/4nruwkL
3. Artificial intelligence index report 2024. Tech. rep., Stanford University (2024)
4. Albertoni, R., Colantonio, S., Skrzypczynski, P., Stefanowski, J.: Reproducibility of machine learning: Terminology, recommendations and open issues. CoRR **abs/2302.12691** (2023)
5. Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., Jacucci, G.: Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. Proceedings of the American Society for Information Science and Technology **50**(1), 1–11 (2013)
6. Ball, P.: Is ai leading to a reproducibility crisis in science? Nature **624**(7990), 22–25 (2023)
7. Cockburn, A., Dragicevic, P., Besançon, L., Gutwin, C.: Threats of a replication crisis in empirical computer science. Commun. ACM **63**(8), 70–79 (jul 2020)
8. Craig, A., Assis, M., Bittencourt, L.F., Nativi, S., Tolosana-Calasanz, R.: Big iron, big data, and big identity. New Frontiers in High Performance Computing and Big Data **30**, 139 (2017)
9. Diaz, O., Kushibar, K., Osuala, R., Linardos, A., Garrucho, L., Igual, L., Radeva, P., Prior, F., Gkontra, P., Lekadir, K.: Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Physica Medica **83**, 25–37 (Mar 2021)
10. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
11. Gundersen, O.E., Kjensmo, S.: State of the art: Reproducibility in artificial intelligence. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
12. Hemminger, B.M., Lu, D., Vaughan, K., Adams, S.J.: Information seeking behavior of academic scientists. Journal of the American Society for Information Science and Technology **58**(14), 2205–2225 (2007)

13. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9) (2023)
14. Lee, D.J., Stvilia, B.: Practices of research data curation in institutional repositories: A qualitative view from repository staff. PLOS ONE **12**(3), e0173987 (2017), publisher: Public Library of Science
15. Mansourian, Y.: Exploratory nature of, and uncertainty tolerance in, qualitative research. New Library World **109**(5/6), 273–286 (2008)
16. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 220–229. FAT* '19, Association for Computing Machinery, New York, NY, USA (Jan 2019)
17. Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.: A manifesto for reproducible science. Nature Human Behaviour **1**(1), 0021 (2017)
18. National Academies of Sciences, Engineering, and Medicine and others: Understanding reproducibility and replicability. Reproducibility and Replicability in Science pp. 39–54 (2019)
19. Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015)
20. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns **2**(11), 100336 (2021)
21. Raff, E.: Does the market of citations reward reproducible work? In: Proceedings of the 2023 ACM Conference on Reproducibility and Replicability. pp. 89–96 (2023)
22. Rogers, A.: Changing the World by Changing the Data (May 2021), http://arxiv.org/abs/2105.13947, arXiv:2105.13947 [cs]
23. Steinerová, J.: Methodological Literacy of Doctoral Students – An Emerging Model. In: Worldwide Commonalities and Challenges in Information Literacy Research and Practice. pp. 148–154. Springer, Cham (2013), iSSN: 1865-0937
24. Tran, N., Chen, H., Bhuyan, J., Ding, J.: Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection. IEEE Access **10**, 121900–121923 (2022), conference Name: IEEE Access
25. Trischler, J., Pervan, S.J., Kelly, S.J., Scott, D.R.: The value of codesign: The effect of customer involvement in service design teams. Journal of Service Research **21**(1), 75–100 (2018)
26. Villegas, D., Bobroff, N., Rodero, I., Delgado, J., Liu, Y., Devarakonda, A., Fong, L., Sadjadi, S.M., Parashar, M.: Cloud federation in a layered service model. Journal of Computer and System Sciences **78**(5), 1330–1344 (2012)
27. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Loy, C.C.: The Devil of Face Recognition Is in the Noise. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 780–795. Springer International Publishing, Cham (2018)
28. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
29. Winter, S., Timperley, C.S., Hermann, B., Cito, J., Bell, J., Hilton, M., Beyer, D.: A retrospective study of one decade of artifact evaluations. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 145–156 (2022)