Promoting Reproducibility of Research Results in International Events (Report from the 4^{th} RRPR)

B. Kerautret[1(✉)], K. Kirchheim[2], D. Lopresti[3], P. Ngo[4], and P. Tomaszewska[5]

[1] Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, 69676 Bron, France
bertrand.kerautret@univ-lyon2.fr
[2] Otto-von-Guericke-University, Magdeburg, Germany
konstantin.kirchheim@ovgu.de
[3] Lehigh University, Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu
[4] Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, France
hoai-diem-phuc.ngo@loria.fr
[5] Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland
paulina.tomaszewska3.dokt@pw.edu.pl

Abstract. Following the fourth edition of the workshop on Reproducible Research in Pattern Recognition (RRPR) at the International Conference on Pattern Recognition (ICPR), this paper reports the main discussions that were held during and after the workshop. In particular, the integration of reproducible research inside an international conference was the first main axis of reflection. Further discussions addressed the ways of initiating or imposing reproducible research, as well as the problem of performance comparisons of published research papers that emerges due to the fact that the reported results are often based on different implementations and datasets.

1 Introduction

Open science practices, such as sharing data and code, are important in computer science for advancing the field as a whole and, in particular, the reproducibility axis. They can help to increase the transparency and reproducibility of research. Encouraging researchers to adopt these practices can increase the impact and credibility of their research. Relying on the advances of new platforms supporting reproducible research [1], in this paper we explore some solutions to the challenges of improving reproducibility in computer science.

Following the RRPR workshop, two main axis were defined in order to guide the discussions on reproducible research (RR). The first point considered the methods to promote and incentivize RR at international conferences (Sect. 2). This part includes new proposals on how to address the topic of RR followed by some reflections on potential effects and impact measures of the current solutions. The second axis was more oriented toward motivating RR on platforms

B. Kerautret et al. (Eds.): RRPR 2022, LNCS 14068, pp. 111–123, 2023.
https://doi.org/10.1007/978-3-031-40773-4_9

like *CodeOcean* and on the key reproducibility issues in computer science communities (Sect. 3).

In the rest of the paper, the reproducibility term will refer to the definition given by the Association for Computing Machinery (ACM) [2] which is the capacity to obtain results by a *"person or team other than the authors, using, in part, artifacts provided by the authors"*. From the same reference, replicability refers to results obtained by *"a person or team other than the authors, without the use of author-supplied artifacts"*.

2 Addressing RR at International Conferences

A starting point for the discussion on this topic was to analyze how reproducible research is currently being addressed at some international conferences in the computer science field. This was inspired by the study of Raff [3] who analysed over 255 papers in machine learning, and reported that only 63.5% could be replicated successfully (or reproduced, under the ACM definition [2]). In the following, we review different proposals that have been made to integrate RR in international conferences and events.

2.1 Recent Proposals

NeurIPS Checklist for RR. We started the analysis by identifying current practices addressing the topic of RR using the example of one of the largest international conferences in Artificial Intelligence – Neural Information Processing Systems (NeurIPS). In the case of NeurIPS, authors are asked to attach a *Paper Checklist* [4] during submission.

This requirement is an outcome of the NeurIPS 2019 Reproducibility program where a *Machine Learning Reproducibility Checklist* [5] was proposed (the revised version is available in [6]). This checklist covers various aspects that affect the reproducibility of the results, *e.g.,* model and algorithm descriptions (including complexity analysis), and the theoretical claims (including proofs) when applicable. It is also recommended to make the dataset publicly available together with its description and share details on the design choices made during training (like hyperparameters chosen, number of runs or dataset used).

In addition to the information on the checklist, it could prove beneficial to disclose the total training time, since resource requirements can drastically limit the number of individuals or institutions that can reproduce experiments (and there are also attendant environmental concerns due to the energy usage and carbon emissions) [7]. Moreover, information about the required dependencies to run the code should be provided, as well as information about the pretrained models (preferably including trained model weights).

Following an author feedback survey reported in [5], the NeurIPS general chairs proposed a simple checklist designed to help authors in assessing their research from the reproducibility standpoint by analyzing aspects defined in the *Machine Learning Reproducibility Checklist*. The authors are asked to fill out the checklist by answering *yes*, *no*, or *n/a*. Note that selecting an option other than

yes does not necessarily entail the rejection of the paper. The checklist should rather initiate self-reflection. Moreover, it helps to identify the limitations of the contribution.

Other Methods of Checking the Requirements of RR. Conferences such as Artificial Intelligence and Statistics (AISTATS) or the European Conference on Computer Vision (ECCV) use a different approach to promote reproducibility of research. During submission, the authors are asked whether they will release the code and the datasets used in the paper. If they agree, the editors check whether the code/dataset are provided with the camera-ready version; for example, as an attachment, or in the form of a link to a public repository. If authors who previously agreed do not publish the code/dataset at the camera-ready stage, the article will not be accepted for publication. This policy imposes control over promises that are not kept by authors.

However, this verification may be rather limited because it is questionable whether the quality of the check is sufficient, since verifying the mere existence of source code/dataset does not guarantee the reproducibility of the results reported in the paper. Furthermore, since the code repositories remain under the control of the authors, they can be altered or removed after publication of the paper. However, still, this initiative should be considered positive, since it incentivizes authors to make the source code publicly available.

Another initiative is the Machine Learning Reproducibility Challenge which consists of an annual call over five successive years for reproducibility reports on papers published at eleven top Machine Learning conferences (including for instance NeurIPS, ICML or CVPR) or top journals [8]. Based on the RR definitions mentioned in the introduction [2], other approaches give incentives for authors to prioritize RR by offering awards or recognition for papers that are checked to be reproducible or replicable [9].

2.2 New Ideas on Promoting RR at International Conferences

To address the limitations of simple checks by editorial teams on whether a code/dataset was publicly shared, the discussion group would like to highlight the idea of introducing a special submission channel oriented to reproducible research. Here, the key principle is to ask authors interested in participating to complement their research submission with a technical description. Such a document should provide details on how the results from the submitted paper can be reproduced. More precisely, it should contain information on the requirements, dependencies, installation procedure, sets of parameters and instructions on how to run the code to reproduce the results described in the main paper. In exchange for the effort to provide such a document, the idea is to construct for authors an online demonstration that could be used to test the proposed model behaviour on a broader spectrum of use cases – not only on the ones presented in the main paper. In the discussion group, we analyzed the schedule for organizing such a special channel. We believe that it could be undertaken as follows:

1. **Preliminary call for Technical Reproduction Instructions Document (TRID).** At this first stage, a document should be uploaded before the deadline of the main conference paper (*e.g.,* one month before). It contains the key instructions on how to run the code of the proposed method/study. The provided instructions can help reviewers test the reproducibility of the research results. Moreover, they can also be considered a starting point for the construction of an online demonstration to be described next. Since the TRID is only a simple document containing technical instructions, it could be submitted before the main paper deadline, even if the algorithms in the paper might be modified later.
2. **Online demonstration and RR report.** In the second step, starting just after the TRID deadline, while the authors are finalising their main paper, the RR reviewing channel team will construct an online demonstration based on the submitted TRID. Future readers of the paper will benefit from such a demonstration which allows testing the method on various inputs without any installation or time-consuming processes. As described in the next point, the online demonstration could be also potentially passed on to reviewers.
   Note that the online demonstration construction could be realized by Master's students with Computer Science backgrounds who will be members of the RR reviewing channel team. For this, one could imagine that the conference chairs would cooperate with various university partners that are identified by the organizers before the event. Ideally, this identification would take place when a team applies to organize the conference. To manage the technical aspects of creating the demonstration, an online demonstration construction system can be used [10]. When an idea proves successful, we can also think about the deployment of the open-source demonstration system [11] in other infrastructures. For this task, the main point of attention should be a hosting system that ensures continuity over time and is accessible to future editions of the conference. Obviously, contributions by Master's students to the process should be recognized, *e.g.,* via a certificate where the details of the contribution are listed. In cases where constructing an online demonstration are too difficult or complex, a simpler alternative could be to require a simple report of code reproduction (RR report).
3. **Open access to reviewer board.** With a deadline that is one month earlier, the demonstration could be passed on to reviewers during the middle of the main paper reviewing phase. The online demonstrations or RR report would first be made available to the authors. After the authors have reviewed the prepared files, we could ask for permission to share them with the reviewers. Thanks to the earlier TRID submission deadline, it would be possible for authors to mention in the paper the limits of the proposed method that were identified using the demonstration.

An overview of this proposal is depicted in Fig. 1. The idea could be beneficial to students, researchers and reviewers. Thanks to the referencing system, each student who participates in the demonstation/RR report preparation process could receive proof of their research experience and will have demonstrated skills

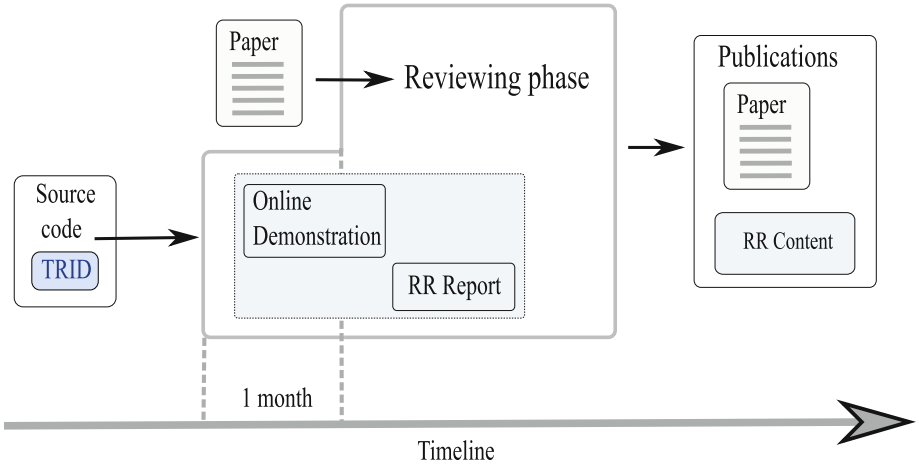


Fig. 1. Example of the proposed procedure to encourage the incorporation of the RR idea in conferences.

and interest in the topic. In addition, students will become familiar with the demonstration system and are likely to be more inclined to incorporate it into their future research. For authors, they would receive a proof of reproduction by another (hopefully objective) team which can be considered by reviewers in making their acceptance decisions. In particular, it will be more difficult for authors to miss limits of their method that they may not have noticed before. Finally, this can also be beneficial to reviewers, since they can now focus on the content of the paper instead of trying to reproduce the supplementary material.

Some questions can be asked about the potential impact of the required submission effort on the authors' willingness to submit their work to conferences that have such a requirement. To motivate them, the procedure for the RR submission should be as simple as possible for authors; such as, for the step of TRID, only a simple description would be required. This could be a formatted document similar to the content found in a Readme file on a project repository indicating the steps needed to install and run the code.

It may well be that the authors could potentially be more interested in devoting their time to work on another publication rather than to care about reproducibility issues. This could be linked to the "publish or perish" paradigm where there is pressure to publish as much as possible, as quickly as possible. In contrast, there is a slow science movement where more time is allocated to research and trying ideas that are innovative, and therefore with higher risk of failure. As a result, the papers are less frequently written. In the slow science movement, there will be more time to work on one topic and develop demonstrations. However, an approach like this that may reduce the number of papers and the number of conferences may be unpopular with conference organizers.

In most cases, building an online demonstration will not require a special effort on the part of authors who have already written their own code. It is possible that the authors will not want to publicize their code because it is not "clean enough" to be shared or due to intellectual property concerns. In such cases, the advantage of an online demonstration is that the source code can be kept hidden from the public if it is implemented as a ready-to-use API with a graphical interface. Moreover, online demonstration platforms are constantly evolving and now can be based on a private repository through the use of *Docker* images[1].

The idea of including the Master's students in an RR processes related to the conference is not necessarily something simple. In particular the organizing team needs to coordinate with supervisors of the Master's programs to ensure student motivation and availability. An alternative solution would be to ask people submitting their work to invest their time in creating the RR report. This way, the additional effort would be shared by everyone which would create incentives to make reproducibility as easy as possible.

2.3 Impact of Efforts Encouraging RR in Conferences

When analyzing examples of conferences that encourage RR and ways to evaluate its impact, a number of questions come to mind. Such information could help strengthen the appeal of RR to other international events. In particular, we raise the following questions:

(a) Have these steps been successful in either increasing the visibility of RR and/or the confidence of those who later cite the published papers that they are reproducible? This evaluation may require defining a new measure, as the number of citations by itself does not necessarily correlate with reproducibility. One could also consider download metrics of source code and data, or the success rate of reproducing research, *etc.*
(b) Are authors, reviewers and conference organizers generally satisfied with the processes? In other words, do they consider the required effort worth it? Feedback on the reproducibility process from authors and reviewers (*e.g.,* the ease of sharing data and code, the usefulness of the reproducibility materials, and the impact of reproducibility on their research, *etc.*) are important for improving the quality of RR and increasing its impact. Some partial answers can be found in the NeurIPS 2019 Reproducibility program report [5].
(c) It is important to note that some conferences are in high demand and have very low acceptance rates which means that they can require authors to do just about anything in the hopes of getting published. But holding such power over authors is a double-edged sword. If the new requirements make it harder to get the work published, who benefits from the new rules and who is hurt by them? It is probably too simplistic to say "good researchers benefit and bad researchers are hurt". Less selective conferences will find it difficult

[1] https://hub.docker.com/.

to request similar amounts of added effort from authors who may become even more likely to try for higher-rated conferences given the time investment required. The proposal described in Sect. 2.2 based on the TRID submission could be a way to attract authors who are curiosity-driven rather than those whose only aim is to publish. In the former case, researchers are interested in broadening awareness of their work, so reproducibility is important to them. In the latter case, researchers would mainly base their decision where to submit based on the acceptance rate of the conference.

(d) It is natural to ask whether any conferences have studied the aforementioned questions and reported the results. Making such an analysis seems to be imperative if changing the way a conference works is motivated by the idea of "improving" it. In the report of the NeurIPS event [5], it is stated that the perception of usefulness of the reproducibility checklist was analysed from the reviewers' point of view. However, it turned out that 34% of the reviewers find the checklist useful while the others either do not read it or find it not useful. Another interesting point is the fact that reviewers who find the reproducibility checklist useful tend to give better reviewing scores and better acceptance rates. The authors of the study also mention that the number of submissions continues to increase (from 40% between 2019 and 2020)[2] which suggests that the checklist does not impact the perception of the conference and the reviewing process negatively. Moreover, the number of authors willing to submit their code increased from 50% to 75%. Other questions remain open from the point of view of authors, but the existing study reports no negative consequence of incorporating the idea of a checklist into the submission process, which may encourage further steps in promoting RR at conferences.

Publishing the analysis and examining the impact of reproducibility requirements, as in the previous example, are good steps forward and should be encouraged across all conferences that have added requirements to ensure research reproducibility. Beyond this, further investigation and questioning are needed about the criteria that will help us understand whether these efforts are successful and whether the extra work by authors and reviewers is justifiable. Here we list some points that should be considered when evaluating the investment of time and the impact on the community through emphasizing RR in international events:

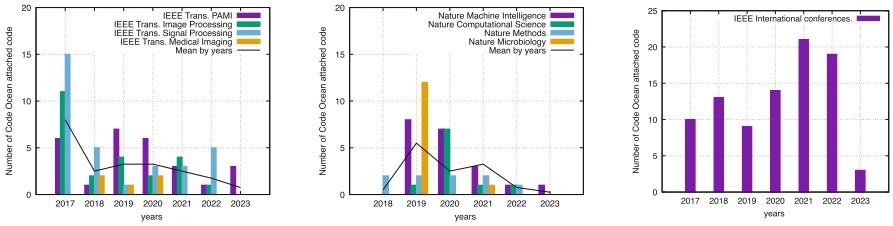
- How do RR practices affect the speed and efficiency of research in computer science?
- Does RR allow higher quality research papers because more time can be devoted to the parts of the paper that advance its new ideas?
- What are the potential benefits to authors for investing in RR? How would such data be captured and reported?
- How to properly measure and report about the failure rate when reproducing previous results?

[2] Source: https://github.com/lixin4ever/Conference-Acceptance-Rate (accessed on 2 April 2023).

- Is RR beneficial to graduate students? RR may allow graduate students to progress faster in their work and to produce stronger PhD dissertations because they will not have to spend as much time reimplementing past work. But, on the other hand, if reproducibility is "easier", they may lose an important learning opportunity for better understanding the work of others.
- Will we see faster progress toward major goals in the field? This would have to be defined and may not be easy to estimate. For example, we might ask whether there are fewer papers claiming small improvements on the state of the art, and more papers claiming significant improvements.

3 Focus on Motivating RR

Following the previous analysis of existing and new strategies for including RR in scientific events, the discussion here considers the motivations behind the use of the platforms facilitating RR and related issues.



(a) IEEE Journal (b) Nature (c) IEEE Conference

Fig. 2. Evolution of *Code Ocean* attachments to papers published in *IEEE* reference journals (a), four *Nature* journals (b) and *IEEE* international conferences (c).([3]Extracted using filters in code ocean platform https://codeocean.com/explore?page=1&filter=withArticle in 26 March 2023.)

3.1 Recent Initiatives

Incentive or Obligation in RR. One point to consider relates to the choice between two approaches – motivation vs. obligation – for including a description of the RR measures undertaken in the published research. The previous example of the RR checklist at NeurIPS is an intermediate state. It asks for answers to the questions in the checklist, but lets authors state that some measures do not apply in their case and does not directly imply rejection of the paper. As mentioned in the previous section, authors and reviewers have found the use of a checklist to be a positive, and the number of submissions seems not to be negatively impacted.

Including RR Platforms in Journal Publication. IEEE also encourages authors to follow RR practices through publicly sharing code and data to facilitate use by others and confirmation of reproducibility. IEEE offers the option of releasing code online using the *Code Ocean* platform launched in 2017 (see [1] for a review on reproducible research platforms). Figure 2 shows the evolution over time of *Code Ocean* attachments mentioned. The number of code submissions to this platform appears stable or decreasing both in the case of *IEEE* and *Nature* journals (Figs. 2a and 2b). It remains relatively low in some cases; for instance, for *IEEE PAMI*, the highest rate of papers containing a *Code Ocean* demonstration was reached in 2019 at only 4.1% (considering 12 issues, each of them containing an average of 14 papers). On the other hand, the number of code capsules associated with publications in international IEEE conferences increased by over 68% from the period 2017–2019 to 2020–2022. Even if the aforementioned trend can be explained by other factors (like specific information campaigns, better knowledge of the *Code Ocean* platform), this suggests that authors are more and more willing to provide demonstrations of their work in international conference publications. This point can be an argument to justify the implementation of the scheme proposed in the previous section to encourage the inclusion of demonstrations in the publication pipeline.

Calls for Demonstration at Conferences. Initiatives like the Call for Demonstrations of IJCAI events appears to be a new way of encouraging RR with links between theory and practice [12]. This call gives authors the opportunity to publish a showcase of research results.

3.2 Issues for Research Result Comparisons

When a new method is proposed, its performance and properties are usually compared with current state-of-the-art methods from the field. Sometimes, this requires reimplementing previous methods as the code was not provided by the original authors. Such additional work is not really considered during the submission and review process. However, it could be of interest to the community. In such cases, maybe the replication could be considered to have value in itself and be recognised by a specific label indicating a post-hoc contribution to already-published work by publicly sharing the code for the replication. Such efforts, if encouraged, would likely require less of a time investment than proposing full-scale replication to a journal like ReScience [13] or IPOL [14].

Training Data. A lack of access to the data for which the results were reported makes reproducibility very difficult. Sometimes, authors cannot release data due to privacy and licensing concerns. This most often occurs when medical or health data is involved in the research, or data is confidential due to company policy (legal issues). In these cases the following solutions are possible:

1. In the paper, the performance of the proposed method could be reported not only on the private data but also on public date so that at least a part of the results can be reproduced. The possible drawback of this solution is that there might not be public data of a similar type, and using other data with a significant distribution shift would require employing a different method than the one being proposed.
2. In cases where it is not possible to share data, *e.g.,* due to legal issues, and there is no similar public data, the authors could give limited access to the private data to a specific certified, third-party entity that would be responsible for checking whether the results can be reproduced. To the best of our knowledge, there is currently no such well-established entity that can certify reproducibility (something similar to the Reproducible Label at the RRPR workshop). Such a certified entity would have to sign Non-Disclosure Agreements (NDA) to ensure that the confidential data will be used in a clearly defined way (whatever is necessary to perform the task) without sharing it with third parties. This idea follows the notion of cybersecurity testing by Red Teams, a service is often outsourced. This policy is more complex than the first one, and would only have to be considered if the first option does not apply.
3. In some cases, a hybrid solution may work, *e.g.,* when medical datasets cannot be fully publicly available. Researchers who would like to use the data must apply for access and sign an agreement specifying Terms of Use (sometimes there is also a requirement to be affiliated with a university). In some cases, the procedure is more complex and the applicant has to additionally complete an online course regarding the Terms of Use. This procedure is applied in the case of the MIMIC dataset [15].

Trained Model. Another way to improve the reproducibility of research is to not only provide the source code, but also the weights and other parameters of the trained models. Such models can then serve as foundation models [16] to facilitate further research of those who would like to build on the original work. Transfer learning is a related notion that is climate-friendly, as it decreases the number of computations during the model training, and therefore reduces carbon emissions which can be substantial for large models [7].

Processing data in such a way could be risky since it may be possible to infer private information. This could also happen due to model inversion attacks [17] that can potentially recover the training data. More recently, some entities have refused to publish model weights, noting ethical and safety concerns [18,19].[3]

3.3 Strengthening Reproducibility: From Publications to Teaching

Open to Coupled Publications. There is no doubt that reproducibility requires extra work on the part of authors and reviewers. In a world where

[3] However, for GPT-4 [19], OpenAI published the evaluation code, which makes comparison with their claimed results easy.

researchers aim to maximize the number of their publications, the extra effort required to follow RR best practices could be a barrier. A possible solution would be to provide additional publication opportunities so that authors receive more recognition depending on the additional work that they do. For instance, in addition to the main conference paper, the authors might be invited to submit another paper providing reproducibility details.

This paradigm would allow work corresponding to a single paper to get publication credit as two papers under a new model: a scientific paper and a paper focusing on its reproducibility. The two separate papers would both be submitted and reviewed at the same time, and if both are accepted, the authors would get two publications – double credit – to account for the extra work they have done. Such an approach is comparable to the "companion paper" initiative we proposed to the authors of accepted ICPR 2022 papers (and also repeated later for the ACM ICMR Reproducibility Track [9]). Such a solution would require some organizational adjustments to account for the extra work done by reviewers and conference chairs. The approach that has already been adopted in some cases is to increase the size of the program committee to match the additional work, perhaps dividing the responsibilities between reviewers who have the expertise to handle scientific submissions versus reproducibility submissions.

To avoid possible confusion, this new paradigm would also require a clear explanation in the Call for Papers so that authors understand the role played by the reproducibility companion paper and how both parts of the submission will be reviewed.

Reproducing Scientific Results and Teaching. Today, even though common licenses such as GPL and BSD require it, it is common for authors to provide their code without instructions on how to run it, or to omit key details (*e.g.,* running parameters, hardware or software configurations, *etc.*). Sometimes, the code is not provided at all. This can make the reproduction task difficult or even impossible.

In many scientific papers, it is often necessary to reproduce the results of other work (*e.g.,* when comparing the performance of different methods, or applying an existing method to a new problem area, *etc.*). While it can be an excellent educational experience to involve students, including them in an attempt to reproduce past work could raise the following issue. In the case that the reproduction attempt fails, it may be hard to distinguish the degree to which the failure is caused by the lack of reproducibility of the paper, or the competence of the *reproducer*. To address this point, an incentive structure can be created that motivates the reproducing party to put serious effort into the reproduction attempt. To implement this, a possible strategy could be to encourage multiple groups (that do not know each other) to independently reproduce the results. In the end, analyzing whether the majority of groups succeeded in their attempts could provide a broader picture of the reproducibility of a paper compared to a single yes-no answer. Groups whose reproducibility assessments fail to agree with the majority could potentially receive a negative score (in an educational

scenario). This incentives students to put effort into the reproduction, since they must assume that others they are competing against will do the same. This straightforward approach for incorporating this idea into teaching could be a starting point for discussions leading to a more sophisticated approach. If all of the groups fail to reproduce the results, it may mean that the code (if provided) is not sufficiently complete, or that the paper is poorly written, at least from the standpoint of reproducibility. The latter reason may be due to the complexity of a particular topic where a high level of in-depth expertise is required to understand the publications. It may also be the result of tight page limits for the paper and any supplementary materials, which does not now allow sufficient space to fully describe the method. If none of the groups succeeds in reproducing the results, it could be interesting to study how this kind of information spreads throughout the research community. We assume that this situation already arises in practice, but that it does not receive enough attention.

Note that there will be work that cannot be reproduced with contemporary means by others since it would require too much computing power, time, money and expertise – for example, large language models like ChatGPT [20], or foundation models more generally. Here, the question arises of how to verify the reproducibility of these very recent solutions. We believe this class of extreme models will require the development of new methodologies.

4 Conclusion

Resulting from the discussion sessions organized during and after the RRPR workshop, this report addresses various questions on the integration of reproducible research in international events, and on motivating authors to apply RR good practices. Perspectives we discussed include the creation of new RR submission channels providing ways of integrating RR in the future. From the analysis of other recent initiatives designed to encourage RR, we anticipate growing degrees of success of these and other proposals to promote reproducible research in upcoming events in our community.

References

1. Colom, M., Kerautret, B., Krähenbühl, A.: An Overview of Platforms for Reproducible Research and Augmented Publications. In: Kerautret, B., Colom, M., Lopresti, D., Monasse, P., Talbot, H. (eds.) RRPR 2018. LNCS, vol. 11455, pp. 25–39. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23987-9_2
2. Artifact review and badging, 2020. Revised August 24. https://www.acm.org/publications/policies/artifact-review-and-badging-current. Accessed October 14
3. Raff, E.: A step toward quantifying independently reproducible machine learning research. In: Advances in Neural Information Processing Systems. Curran Associates Inc, (2019)
4. NeurIPS 2022 Paper Checklist Guidelines. https://neurips.cc/Conferences/2022/PaperInformation/PaperChecklist (accessed in 26 February 2023)

5. Pineau, J.: Improving Reproducibility in Machine Learning Research (a Report from the NeurIPS 2019 Reproducibility Program). J. Mach. Learn. Res., **22**(1) (2022). Publisher: JMLR.org
6. The Machine Learning Reproducibility Checklist. https://www.cs.mcgill.ca/jpineau/ReproducibilityChecklist.pdf. Accessed 13 Mar 2023
7. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650 (2019)
8. ML Reproducibility Challenge 2022, 2022. https://paperswithcode.com/rc2022,. Accessed 4 Mar 2023
9. ICMR reproducibility, 2023. https://icmr-reproducibility.github.io/website/cfp2023/, Accessed 4 Mar 2023
10. Arévalo, M., Escobar, C., Monasse, P., Monzón, N., Colom, M.: The IPOL Demo System: A Scalable Architecture of Microservices for Reproducible Research. In: Kerautret, B., Colom, M., Monasse, P. (eds.) RRPR 2016. LNCS, vol. 10214, pp. 3–16. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56414-2_1
11. IPOL demo system development. https://github.com/ipol-journal/ipolDevel. Accessed 26 Feb 2023
12. Call for demonstrations of the IJCAI international conference. https://github.com/ipol-journal/ipolDevelhttps://ijcai-23.org/call-for-demos. Accessed 1 Apr 2023
13. Rougier, N.P., Hinsen, K.: ReScience C: A Journal for Reproducible Replications in Computational Science. In: Kerautret, B., Colom, M., Lopresti, D., Monasse, P., Talbot, H. (eds.) RRPR 2018. LNCS, vol. 11455, pp. 150–156. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23987-9_14
14. Colom, M., Kerautret, B., Limare, N., Monasse, P., and Jean-Michel Morel. IPOL: a new journal for fully reproducible research; analysis of four years development. In: Badra, M., Boukerche, A.,mUrien, P., eds 7th International Conference on New Technologies, Mobility and Security, NTMS 2015, Paris, France, July 27–29, 2015, pp. 1–5. IEEE (2015)
15. Johnson, A., Bulgarelli, L.: Tom Pollard. Leo Anthony Celi, and Roger Mark. MIMIC-IV, Steven Horng (2021)
16. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
17. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, pp. 1322–1333, New York, NY, USA (2015). Association for Computing Machinery
18. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
19. OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
20. van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L.: ChatGPT: five priorities for research. Nature **614**(7947), 224–226 (2023)