Towards Reproducible Artificial Intelligence: Roles and Responsibilities of Researchers and Publishers

Anita de Waard, Sweitze Roffel, Catriona Fennel, Sergios Petridis,
Thom Pijnenburg, Efthymios Tsakonas, Rinke Hoekstra, George Tsatsaronis
Elsevier BV
Radarweg 29,
1043 NX, Amsterdam, The Netherlands

Abstract

Over the past few years, we have seen a rapid growth in the field of artificial intelligence and its applications, such as machine learning, computer vision and natural language processing. This means that the issue of ensuring reproducible published scientific results is now becoming even more crucial for the AI community. With increased frequency, our community is experiencing critical reports of research results that cannot be reproduced. Some recent studies tried to shed light on the main reasons for this lack of reproducibility. A primary criticism is the lack of proper documentation and an appropriate explanation of the methods and the experimental setups and conditions in AI publications. In this paper, we share some of our own experiments in this arena, and present our opinion on the roles and responsibilities that both researchers and publishers have in the process of increasing reproducibility in AI. We suggest some potential directions for the community to consider, and identify some possible best practices that we can undertake, to improve the reproducibility of future research in artificial intelligence.

Introduction

Modern scientific research, especially in fields such as artificial intelligence, is a living ecosystem: code, data, protocols, methods, hyperparameters, analysis and evaluation methodologies are all elements of computational research. They are interconnected, and all are needed to fully describe what was done and why, and what conclusions can be drawn from a specific piece work. To provide access to research in a rigorous and reproducible way, researchers need to document, archive and share these various components, e.g., by adopting new best practices and guidelines for authors (Peng 2011), or by allowing more detailed descriptions of an environments to support this narrative (Brinckman et al. 2019). Clearly, these efforts need to be supported by an environment that can fully enable reproducible AI research.

This is evermore important, since this research is increasing at a tremendous rate. According to a recent report that maps and analyzes the landscape of the field of AI (Elsevier 2019)[1], the volume of AI research publications has been growing by $12,9\%$ annually, over the last 5 years. A significant portion of these publications is published as conference proceedings; in fact, over 70% or recent corporate AI research in the United States has been published in proceedings. Pre-prints in *arXiv*[2] in core AI categories have grown $37,4\%$ annually over the same period of time, especially in *Machine Learning* and *Computer Vision and Pattern Recognition*, as Figure 1 shows.

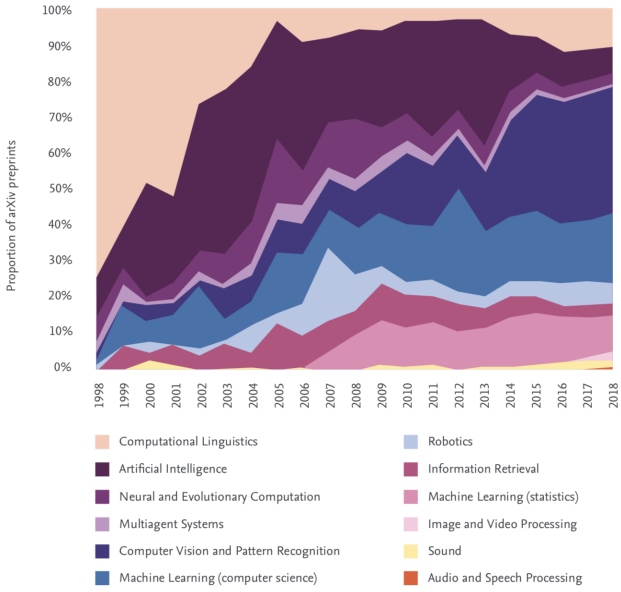

Figure 1: Proportion of *arXiv* preprints submitted in core AI categories, per category, 1998-2017; Source: *arXiv*.

A *Scopus*[3] study conducted in the framework of the same report shows that the trends in publications are growing faster in the fields of *Machine Learning*, *Neural Networks*,

[1] https://www.elsevier.com/research-intelligence/resource-library/ai-report

[2] https://arxiv.org/

[3] https://www.scopus.com/

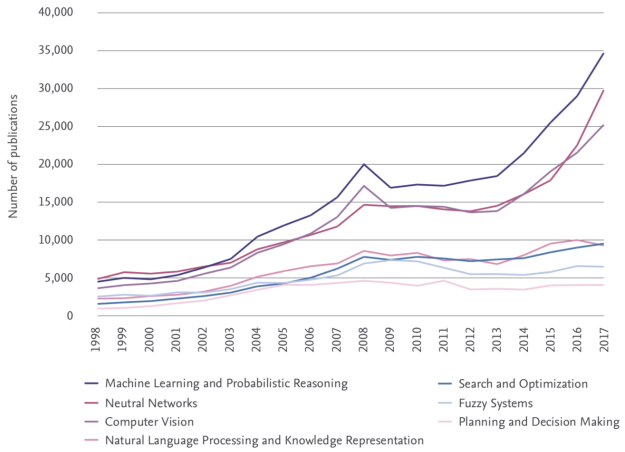

Figure 2: Annual number of AI publications by keyword co-occurrence cluster (all document types), 1998-2017; sources: *Scopus* and *Elsevier* clustering.

Computer Vision and *Natural Language Processing and Knowledge Representation*, as shown in Figure 2. Though the larger volume in AI publications lies in conference proceedings and pre-prints, the growth trends are similar for journals, suggesting that support for reproducible AI research should be available for all types of publications.

As a leading science publisher, we are engaged in supporting reproducible research in experimental and computational sciences; at the same time, as a data analytics company, we employ AI researchers who conduct research, and wish to do so in a reproducible way. In this paper, we describe what an evolved publication model can look like, for both of these roles: as a publisher, and as a community of researchers.

We will commence with providing a definition of reproducibility that is threefold, namely experiment, method or data reproducible. Then, we will discuss all three aspects of reproducibility for our work as a publisher, and next, as a research organisation.

Definition of Reproducibility in AI

To begin with, we want to provide our definition of AI, since there are many different definitions given in different domains. Following the work of Goodman et al. (2016) and Gundersen and Kjensmo (2018) we will define three levels of reproducibility:

R1: Experiment Reproducible (or '*Repeatable*') This means one would draw the same conclusions from either an independent replication of a study or a reanalysis of the original study, i.e., the exact implementation of the AI method on the same data produces the same results. To achieve this, the researcher must document the AI Method, the Data used to conduct the experiment, and the Experiment itself (including the source code for the AI method and the experimental setup)

R2: Data Reproducible (or '*Replicable*') This means one would obtain the same results from an independent study with procedures as closely matched to the original study as possible, i.e., an alternative implementation of the AI method executed on the same data would produce consistent results. To achieve this, the researcher must document the AI Method and the Data.

R3: Method Reproducible (or '*Reproducible*') This means one needs to provide sufficient detail about procedures so that the same procedures can be repeated on different data, providing consistent results. To achieve this, the researcher must document the AI Method. In (Sandve et al. 2013), the authors formulate this requirement as follows: "*As a minimal requirement, you should at least be able to reproduce the results yourself.*".

Reproducibility Responsibilities of Publishers

As publishers, our work in Reproducibility has been shaped in thought and action by Marcus Munafo and colleagues 2017 Manifesto for reproducible science [4]. Munafo et al. proposed a series of measures to improve the efficiency and robustness of research by targeting specific threats to reproducible science. As an influential stakeholder in the research workflow, Elsevier journals take their responsibilities to nurture and incentivize reproducibility very seriously[5]. The broad, practical, evidence-based and actionable Munafo manifesto proved a perfect "gold standard" for Elsevier to benchmark its existing Reproducibility program and create a roadmap for future initiatives. The manifesto's categories-methods, reporting and dissemination, reproducibility, evaluation and incentives - allowed us to structure and prioritise our program to comprehensively address the threats to reproducibility throughout the research workflow and system.

This section is organised according to the three aspects of Reproducibility mentioned above, which can be formulated as Repeatability, Replicability and Reproducibility. As Gundersen et al. point out (2018), this means identification of three aspects: Methods, Data, and the full Experiment (including the full research environment). We will discuss a number of our efforts to improve reproducibility in science along these three axes, in turn.

Methods

Next to the general Methods and Protocols sections provided in our 2,800 journals, the Cell Press family of journals have been working on expanding and structuring the coverage of Methods, Materials, and Protocols using a formal structure dubbed 'STAR Methods', which stands for 'Structured, Transparent, Accessible Reporting'[6]. For motivation and more details, see also the Editorial[7]. Recently, this approach has been expanded by offering a machine-learning generated Key Resources Table[8]. We hope to expand these

[4] See https://www.nature.com/articles/s41562-016-0021

[5] See https://www.elsevier.com/connect/how-elsevier-is-breaking-down-barriers-to-reproducibility

[6] https://www.cell.com/star-methods

[7] https://www.cell.com/cell/fulltext/S0092-8674(16)31072-8

[8] See https://www.elsevier.com/connect/editors-update/cant-see-the-method-for-the-madness-reach-for-the-star

efforts to enable improved extraction of structured methodology, which is generated from the paper automatically, but curated by the author.

Data

The Semantic Web community has traditionally been very active in advocating the publication of well described research data; teaming up with bio- and medical informaticians to launch the FAIR data initiative (Wilkinson et al. 2016)[9]: research data should be findable, accessible, interoperable, and reusable. The FAIR initiative has now been adopted across other disciplines, is embraced by government organisations such as the EU,[10] and the G20[11] as well as industry.

The challenge of publishing research data has been met head on by other data-intensive disciplines in the social sciences as well as the life sciences and resulted in both infrastructural – free to use data archives such as *Mendeley Data*, *Figshare*, *Dataverse* – as well as institutional embedding – data management clauses are now required by almost all major funding organizations. As one of the leaders in various data sharing efforts, 1800 Elsevier journals offer transparent research data-sharing policies and facilitate data sharing using the TOP Guidelines[12]. Working within the wider Research Data community, Elsevier team members have helped to pioneer the *Force11 Data Citation Implementation*[13] as well as efforts in the Earth and Space sciences to establish a cross-publisher set of author guidelines that mandate data sharing and the use of Data Availability Statements [14].

To help validate, review, disseminate and archive these artefacts, Elsevier has created new article types for software code, data and other digital research outputs [15]. These new elements can be published by existing journals, or by dedicated journals, for example journals that exclusively review and publish scientific software like "Software Impacts - ISSN:2665-9638 [16], SoftwareX ISSN: 2352-7110 [17] and the venerable Computer Physics Communications - ISSN: 0010-4655 [18] and together with community journals like The Journal of Open Source Software [19]. These dedicated software journals support not only the principles of the force 11 software citation working group [20]. They also aid the efforts of the Software Sustainability Institute's motto of:"Better software = better research" [21], and help to provide scientific publication outlets for the people in research groups "who write code, not papers". This "new" persona in academia is also known as the Research Software Engineer [22].

Experiments

The pinnacle of reproducibility is to enable the full reproduction of entire computational experiments, including the motivation behind the experiment, the runtime environment, and all parameters and settings used for a computational experiment.

In 2011 Elsevier formed the Executable Paper Grand Challenge to address the problem that computer science research results can be difficult to reproduce. Vital blocks of information needed to replicate such results – for example, software, code, large data sets—are typically unavailable within the context of a scholarly publication. The Executable Paper Grand Challenge created an opportunity for scientists to design solutions that capture this information and provide a platform whereby this data can be verified and manipulated. At the 2011 International Conference on Computational Science (ICCS) on June 2 in Singapore 3 winners were presented from over 70 submissions (Gabriel and Capone 2011) and number of winning systems were piloted in an Elsevier Special issue to test scalability of these proposed solutions [23].

To explore what such fully repeatable research would look like, our Computer Science program has also been following early efforts in the Database community who have been testing and learning techniques since 2008[24] (Manolescu et al. 2008). Since this groundbreaking work, efforts to support reproducible science have only grown in the database community, and include current work that award a prize for the most reproducible paper[25] and offering badges for fully reproducible research[26]. the Inspired by these efforts in the DB community, from 2015 on-wards Elsevier Publishers helped its journal editors to launch a new Reproducibility Section in the Information Systems Journal (Chirigati et al. 2016).

This new journal section facilitates computational validation of results presented in manuscript submissions by asking selected authors to capture their entire experimental environment together with all the software, code, data (the journal recommends *ReproZip* and *Docker* to fully package and visualize the experiment), and make this available to the journals reviewers and readers via *Mendeley Data*, so, another lab redoes the same experiment, with dame method

[9] See https://www.go-fair.org/fair-principles/

[10] See https://www.dtls.nl/2016/04/20/european-commission-allocates-e2-billion-to-make-research-data-fair/

[11] See https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967

[12] See https://www.elsevier.com/connect/editors-update/supporting-openness,-transparency-and-sharing

[13] https://www.nature.com/articles/sdata2018259

[14] http://www.copdess.org/enabling-fair-data-project/author-guidelines/

[15] https://www.elsevier.com/authors/author-resources/research-elements

[16] https://www.journals.elsevier.com/software-impacts

[17] https://www.journals.elsevier.com/softwarex

[18] https://www.journals.elsevier.com/computer-physics-communications/

[19] https://joss.theoj.org/

[20] https://www.force11.org/group/software-citation-working-group

[21] https://www.software.ac.uk/about

[22] https://rse.ac.uk/who/

[23] https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-announces-winners-of-the-executable-paper-grand-challenge

[24] See http://db-reproducibility.seas.harvard.edu/past-efforts/index.html for an overview

[25] https://sigmod.org/sigmod-awards/sigmod-most-reproducible-paper-award/

[26] http://db-reproducibility.seas.harvard.edu/papers/#SIGMOD2018

and same data. Even with all the data, software and computational environment at hand, actually replicating research from another lab is non-trivial, and therefore the reviewers work with the original authors and systematically document their entire replication experience for public consumption. This replication guide to replication is also shared, and Information Sciences published its first "invited reproducibility paper" in January 2016 (Wolke et al. 2016), together with all data and software needed (Wolke 2015) to replicate and thereby validate the results and claims presented in the original research paper (Wolke et al. 2015), enabling any other lab to further build on this trusted, replicated, open science.

Next to these efforts, Elsevier journals welcome the submission of preprints, and successful pilots have proven the viability of publishing peer review reports and conducting full-scale replication studies[27].

Reproducibility Responsibilities of Researchers

Bearing in mind the three elements for reproducible research: clear description of methodology, definition of raw materials, and availability of executable implementation of algorithms and other analyses, there are standards and technologies researchers may adopt to ensure their work satisfies these reproducibility criteria. Borrowing best practices from the field of software engineering around packaging and distribution of software has the potential to drastically improve the reproducibility of data and code-centric research.

Methods

The first element can be addressed by rigorous documentation of the complete experimental process. In the context of *NeurIPS 2019*, the Reproducibility chairs drew up a checklist[28] capturing the essential elements that manuscripts need to contain in order to be regarded reproducible, such as descriptions of mathematical tools, algorithms or models, descriptions of the data collection process, and generation of training, validation and test data samples, clear definitions of statistics and metrics used to report results and more. Whereas this sounds straightforward, steps that are obvious to the authors are often left out, making reproducibility attempts much harder.

Experiments

In a large segment of AI research, experiments are relying on data used to train and test models. Therefore, reproducibility of the experiments includes being specific on the exact data used to derive the models. In a dynamic/collaborative environment, data can evolve in the context of workflows involving pre-processing or cleansing. Flows of incoming data can further complicate reproducibility of experimental results due to data drifting phenomena. The researcher should then make sure that data is versioned in a similar way as code. Tools that support versioning data as well as synchronising their version with the code version, such as `git lfs`[29] or `dvc`[30] can assist in this direction. Sharing the data with the community however may manifest as a different kind of challenge due to licence restrictions or privacy concerns.

Reproducibility pivots on the ability to reconstruct the complete set of experimental conditions in which the research has been performed, yet this is often made challenging by lacking documentation or incomplete dependency management. Without a complete set of instructions, reverse engineering the exact conditions is an impossible job. As AI research is leaning heavily on statistical techniques, deceptively small details matter, like setting the seed for a random number generator, which variations could already lead to differences in trained models and their predictions.

A contributing factor is the dependency of AI research on external software or libraries, which forms an intricate web of interdependent packages. Scientific computing ecosystems have been enriched by open source projects providing tools to perform linear algebra, or providing convenient interfaces for the programming of deep learning models. Although these tools have made AI research better accessible than every before, their popularity can lead to the development of applications with a complicated dependency structure that needs to be made explicit by the researcher.

In established and emerging technologies the community could find tools to address the technological burden of replicating software-centric research. One principle that could be a step towards this goal is containerization. Containerization technologies such as `Docker`[31] allow the researcher to define an explicit set of instructions to recreate the experimental conditions for their research that will behave consistently regardless of the specifics of the host operating system. Containers wrap the application as a single executable package of software that contains all code, supporting configuration files, and dependencies required for it to run. The benefits of containerization go beyond portability, as its ability to interface with container orchestration technologies like `Kubernetes`[32] allows to easily spawn scalable model training jobs or deploy AI applications as web services.

One obvious drawback of adding technology and tooling to research outputs, is its increase in complexity, and counter productively, an increase in opacity. While docker containers are often distributed as opaque objects, the research community could mandate a similar level of rigorous descriptiveness in the container definitions, as it requires in the methods descriptions in the research paper (as in *NeurIPS 2019*[28]) .

Data

Similar principles that apply for code, could also apply for data. Software repositories are already widely adopted in

[27] https://www.elsevier.com/social-sciences-and-humanities/business-management-and-accounting/journals/virtual-special-issue-on-replication-studies

[28] https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

[29] https://git-lfs.github.com/

[30] https://dvc.org/

[31] https://www.docker.com/

[32] urlhttps://kubernetes.io/

AI research, but a full integration of these digital supplements with the research paper is still lacking. Papers often reference their code and data through a url as a supplementary material living on an online repository like `Github` or a file server managed internally. This essential resource is too loosely coupled to the manuscript, and this connection can be broken by various factors, that it begs for innovation on the publishers front to consider code and data as key resources in the publication of AI research. A potential direction could be the ability of researchers to actually publish both data and software that will have persistent unique identifiers, so that researchers could link and cite; similarly, such a solution could enable software, as well as data versioning. In this light, the efforts by the *Force11 Software Citation Working Group* [33] promise to offer an easy to implement solution that can garner adoption across different stakeholder communities.

All these practices, tools and technologies stimulate the development of sustainable applications for AI systems, and address many of the points described by Sculley et al. (2015) pertaining to the hidden technical debt of machine learning software.

In Conclusion

Although many of the infrastructure and technological challenges to reproducibility in AI research have been solved in principle, there is still important work to be done to obtain agreement and adoption of common tools and practices, to address the growing need for scrutiny in this rapidly expanding field.

The three aspects of reproducibility (repeatability, reusability and reproducibility) described above, lead to a recommendation to store, share and cite all the Methods, Data and full Experimental environments for computational research. In this paper, we provided some examples of ways in which we support these principles in practice, both as publishers and as researchers.

But key questions remain. How do we incentivize researchers to make their work reproducible? Should data and code used in AI research become primary citizens in research submissions? Code currently enters the publication chain as supplementary material, but should it be more at the forefront? And how do we connect various parts of the value chain and ensure that the (non-neglible!) efforts to make research reproducible pay off, when the time comes to showcase your research, and reap its rewards?

For these questions to be answered, it will be key that all parties involved (researchers, publishers, software repositories, industry, government and academia) work together to provide an ecosystem that is open, equitable and accessible, yet allows proper attribution to work done by developers, researchers and curators.

References

Brinckman, A.; Chard, K.; Gaffney, N.; Hategan, M.; Jones, M. B.; Kowalik, K.; Kulasekaran, S.; Ludäscher, B.; Mecum, B. D.; Nabrzyski, J.; Stodden, V.; Taylor, I. J.; Turk, M. J.; and Turner, K. 2019. Computing environments for reproducibility: Capturing the "whole tale". *Future Generation Comp. Syst.* 94:854–867.

Chirigati, F.; Capone, R.; Rampin, R.; Freire, J.; and Shasha, D. 2016. A collaborative approach to computational reproducibility. *Information Systems* 59:95 – 97.

Elsevier. 2019. Artificial intelligence: How knowledge is created, transferred and used. Artificial intelligence report, Research Intelligence Elsevier.

Gabriel, A., and Capone, R. 2011. Executable paper grand challenge workshop. *Procedia Computer Science* 4:577 – 578. Proceedings of the International Conference on Computational Science, ICCS 2011.

Goodman, S. N.; Fanelli, D.; and Ioannidis, J. P. A. 2016. What does research reproducibility mean? *Science Translational Medicine* 8(341):341ps12–341ps12.

Gundersen, O. E., and Kjensmo, S. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, 1644–1651.

Manolescu, I.; Afanasiev, L.; Arion, A.; Dittrich, J.; Manegold, S.; Polyzotis, N.; Schnaitter, K.; Senellart, P.; Zoupanos, S.; and Shasha, D. 2008. The repeatability experiment of sigmod 2008. *SIGMOD Rec.* 37(1):39–45.

Peng, R. 2011. Reproducible research in computational science. *Science* 334(6060):1226–1227.

Sandve, G. K.; Nekrutenko, A.; Taylor, J.; and Hovig, E. 2013. Ten simple rules for reproducible computational research. *PLOS Computational Biology* 9(10):1–4.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.; and Dennison, D. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2503–2511.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J. .; da Silva Santos, L. B.; and Bourne, P. E. e. a. 2016. Comment: The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3.

Wolke, A.; Tsend-Ayush, B.; Pfeiffer, C.; and Bichler, M. 2015. More than bin packing: Dynamic resource allocation strategies in cloud data centers. *Information Systems* 52:83 – 95. Special Issue on Selected Papers from SISAP 2013.

Wolke, A.; Bichler, M.; Chirigati, F.; and Steeves, V. 2016. Reproducible experiments on dynamic resource allocation in cloud data centers. *Information Systems* 59:98 – 101.

Wolke, A. 2015. Reproducible experiments on dynamic resource allocation in cloud data centers. *Mendeley Data*.

[33] https://www.force11.org/group/software-citation-implementation-working-group