On Reproducible AI

Towards reproducible research, open science, and digital scholarship in AI publications

Odd Erik Gundersen, Yolanda Gil and David W. Aha

Abstract

Background: Artificial intelligence, like any science, must rely on reproducible experiments to validate results. **Objective:** To give practical and pragmatic recommendations for how to document AI research so that results are reproducible. **Method:** Our analysis of the literature shows that AI publications currently fall short of providing enough documentation to facilitate reproducibility. Our suggested best practices are based on a framework for reproducibility and recommendations for best practices given by scientific organizations, scholars, and publishers. **Results:** We have made a reproducibility checklist based on our investigation and described how every item in the checklist can be documented by authors and examined by reviewers. **Conclusion:** We encourage authors and reviewers to use the suggested best practices and author checklist when considering submissions for AAAI publications and conferences.

1. Introduction

Reproducibility is a cornerstone of the scientific method. The ability and effort required from other researchers to replicate experiments and explore variations depends heavily on the information provided when the original work was published. Reproducibility is challenging for many sciences, for example when the variability of physical samples and reagents can significantly affect the outcome (Begley and Ellis 2012; Lithgow et al. 2017). In computer science, a large portion of the experiments are fully conducted on computers, making the experiments more straightforward to document (Braun and Ong 2014). Most AI and machine learning research also fall under this category of computational experimentation. However, reproducibility in AI is not easily accomplished (Hunold and Träff 2013; Fokkens et al. 2013; Hunold 2015). This may be because AI research has its own unique reproducibility challenges. Ioannidis (2005) suggests that the use of analytical methods which are still a focus of active investigation is one reason it is comparatively difficult to ensure that computational research is reproducible. For example, Henderson et al. (2017) show that problems due to non-determinism in standard benchmark environments and variance intrinsic to AI methods require proper experimental techniques and reporting procedures. Acknowledging these difficulties, computational research should be documented properly so that the experiments and results are clearly described.

The AI research community should strive to facilitate reproducible research, following sound scientific methods and proper documentation in publications. Concomitant with reproducibility is open science. This involves sharing data, software, and other science resources in public repositories using permissive licenses. Open science is increasingly associated with FAIR principles to ensure that science resources have the necessary metadata to make them findable, accessible,

interoperable, and reusable (Wilkinson et al. 2016). Modern digital scholarship promotes proper credit to scientists who document and share their research products through citations of datasets, software, and innovative contributions to the scientific enterprise.

The focus in this article is on best practices for documentation and dissemination of AI research to facilitate reproducibility, support open science, and embrace digital scholarship. We begin with an analysis of recent AI publications that highlights the limitations of their documentation in support of reproducibility.

2. State of the Art: How AI Research is Currently Documented

Gundersen and Kjensmo (2018) analyzed how well empirical AI research is documented to facilitate reproducibility. Empirical AI research involves evaluating how well computational AI methods solve a problem. An *AI method* refers to an abstract method for solving such problems. Examples include agent systems that perform collaborative tasks and neural network architectures trained using backpropagation.

The analysis by Gundersen and Kjensmo (2018) is based on a literature review and a framework for reproducibility. Their framework divides documentation into three factors: (1) *Method*, which specifies the AI method under investigation and the problem to be solved; (2) *Data*, which describes the data used for conducting the empirical research; and (3) *Experiment*, which documents how the experiment was conducted. How well these three factors are documented is indicated by 16 yes/no variables (see Table 1) that are directly relevant for facilitating reproducibility.

A publication that documents an empirical research study can be scored using these variables. Three reproducibility metrics are proposed. The three metrics are: (1) R1D, which calculates the average of all variables for all three factors (Method, Data, and Experiment), (2) R2D, which computes the average of the variables for the Method and Data factors, and (3) R3D, which calculates the average of all variables for the Method factor. These three metrics provide an indication of how well the scored papers document the research for three different degrees of reproducibility (we provide more detail on these degrees of reproducibility in Section 3).

Table 1: Description of all variables and their factors.

Factor	Variable	Description
Method	*Problem*	Is there an explicit mention of the problem the research seeks to solve?
	Objective	Is the research objective explicitly mentioned?
	Research method	Is there an explicit mention of the research method used (empirical, theoretical)?
	Research questions	Is there an explicit mention of the research question(s) addressed?
	Pseudocode	Is the AI method described using pseudocode?

Data	*Training data*	Is the training set shared?
	Validation data	Is the validation set shared?
	Test data	Is the test set shared?
	Results	Are the relevant intermediate and final results output by the AI program shared?
Experiment	*Hypothesis*	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Method source code	Is the AI system code available open source?
	Hardware	*Is the hardware* used for conducting the experiment specified?
	Software dependencies	Are software dependencies specified?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
	Experiment source code	Is the experiment code available open source?

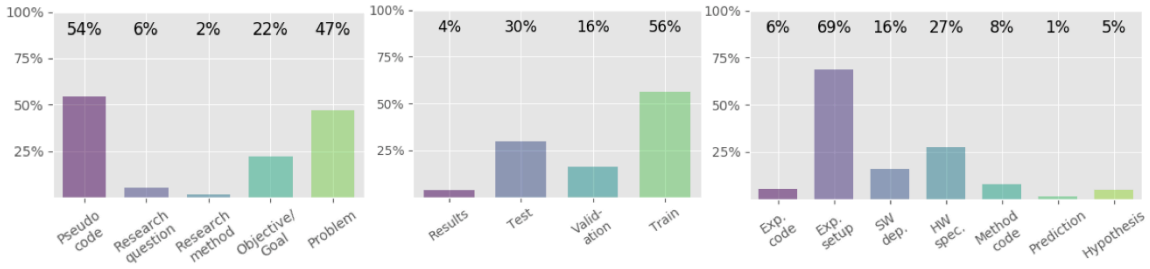

Figure 1: Percentage of papers documenting each variable for the three factors: Method (left), Data (middle), and Experiment (right) (Gundersen and Kjensmo 2018).

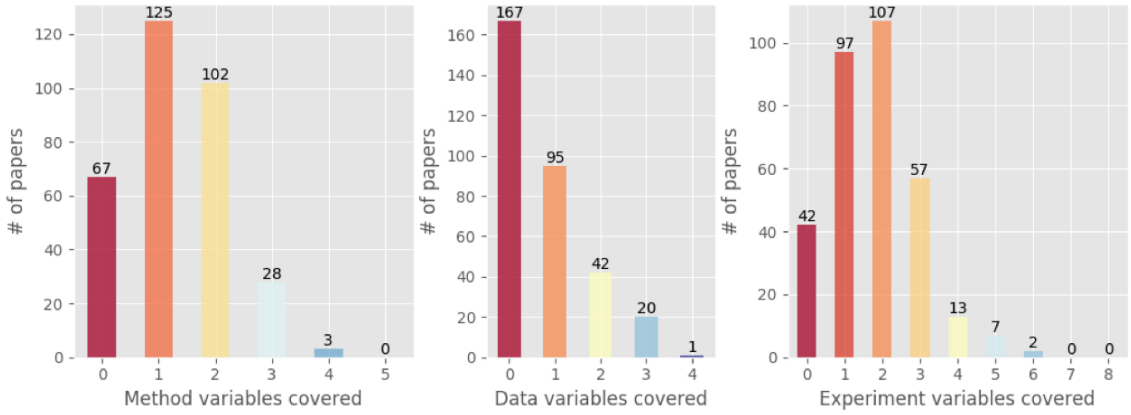

Figure 2: The number of variables for the three factors as documented for all the papers describing empirical research in the study by Gundersen and Kjensmo (2018): Method (left), Data (middle), and Experiment (right).

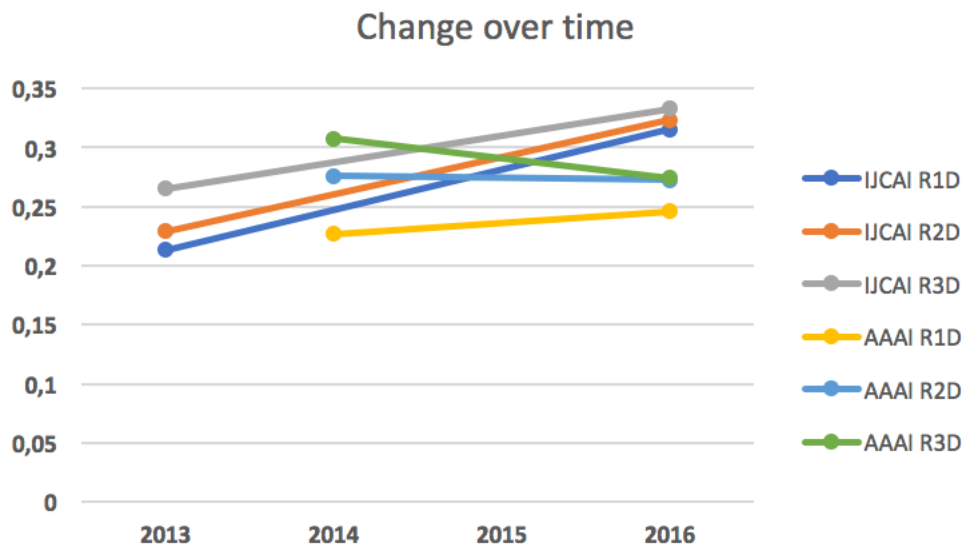

Figure 3: Change over time of the three reproducibility metrics for the two conferences AAAI and IJCAI (Gundersen and Kjensmo 2018).

In total, Gundersen and Kjensmo sampled 400 papers from the AAAI 2014, AAAI 2016, IJCAI 2013 and IJCAI 2016 conferences. Among these, 325 papers describe empirical studies, while the remaining 75 papers do not. Figure 1 displays the percentage of the surveyed papers that documented the different variables while Figure 2 summarizes how many of the variables were documented for each factor per paper.

We make a few observations. As seen in Figure 1, few of the papers explicitly mention the research method that is used, and only around half explicitly mention which problem is being solved. Only about a third of the papers share the test data set and only 4% share the result produced by the AI program. Only 8% of the papers share the source code of the AI method that is being investigated while only 5% explicitly specify the hypothesis and 1% specify their prediction. Figure 2 shows that: 67 papers do not explicitly document any of the variables for the factor Method; only one paper documents and shares training, validation and test sets as well as the results; and approximately 90% of the papers document no more than three of the seven variables of the factor Experiment.

As seen in Figure 3, the trends are unclear. Statistical analysis showed that only two of the metrics, R1D and R2D, for IJCAI had a statistically significant increase over time. While R2D and R3D for AAAI decrease over time, the decrease is not statistically significant.

The study by Gundersen and Kjensmo (2018) has some limitations. For example, the study required that for the variable *problem* to be set to *yes (true)*, the paper must explicitly state the problem that is being solved. Another shortcoming is that all the AI methods that are documented in the research papers are not necessarily

described better with pseudocode than without, but this fact was not given any consideration. If a paper described an AI method and pseudo code was not provided, the pseudocode variable was set to *false* for that paper. Finally, some of the variables might be redundant (e.g., problem, goal, or research questions). Still, despite these shortcomings, the findings indicate that computational AI research is not documented systematically and with enough information to support reproducibility.

3. Degrees of Reproducibility

Gundersen and Kjensmo (2018) distinguish between three degrees of reproducibility, which are defined as follows:

R1: Experiment Reproducible The results of an experiment are *experiment reproducible* when the execution of the same implementation of an AI method produces the same results when executed on the same data. This is often called *repeatability*.

R2: Data Reproducible The results of an experiment are *data reproducible* when an experiment is conducted that executes an alternative implementation of the AI method that produces the same results when executed on the same data. This is often called *replicability*.

R3: Method Reproducible The results of an experiment are *method reproducible* when the execution of an alternative implementation of the AI method produces consistent results when executed on different data. This is often called *reproducibility*.

Empirical research that is R1 (Experiment Reproducible) 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 experiment setup, while R2 (Data Reproducible) research must only document the AI method and the data. R3 (Method Reproducible) research must only document the AI method. Figure 4 illustrates the different factors that must be documented for the three reproducibility degrees.

If an independent team reproduces research and gets results that are consistent with the original results, the generality of the results depends on the level of documentation that was provided to the independent team. If the original research was R1 (Experiment Reproducible), the independent team has confirmed that the specific implementation of the AI method provided by the original research team achieved the reported results on the specific data that also was provided by the original research team. Hence, the generality of the results is limited to that specific implementation and that specific data. However, if the independent team reproduces the results of some research that was R3 (Method Reproducible) and gets consistent results, the results are more general, as they apply to a re-

implementation and other data. This leads to different incentives for the researchers who conducted the initial empirical study and the independent researchers who attempt to reproduce the results.

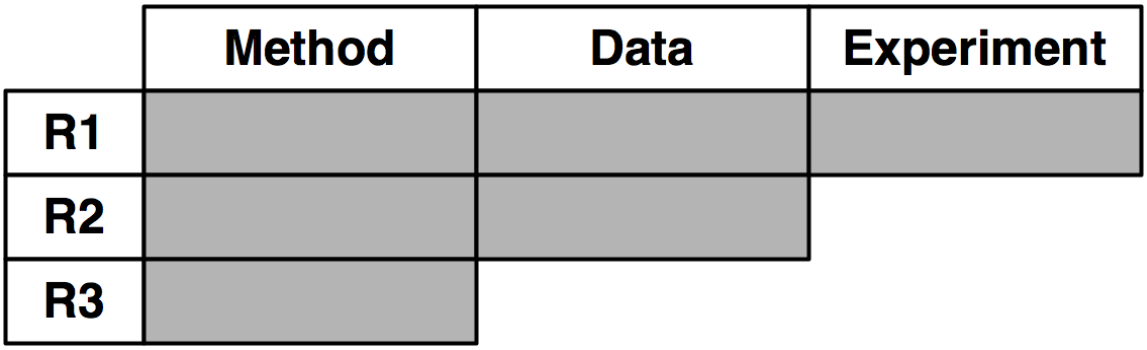

Figure 4: The three degrees of reproducibility are defined by which documentation is used to reproduce the results. The three degrees of reproducibility each require a different set of factors to be documented.

Independent researchers trust an empirical study's results increasingly with the amount of documentation that is shared with them, while the effort to reproduce the results increases when the amount of documentation is reduced. This situation is illustrated in Figure 5. Hence, independent researchers would prefer R1 (Experiment Reproducible) research.

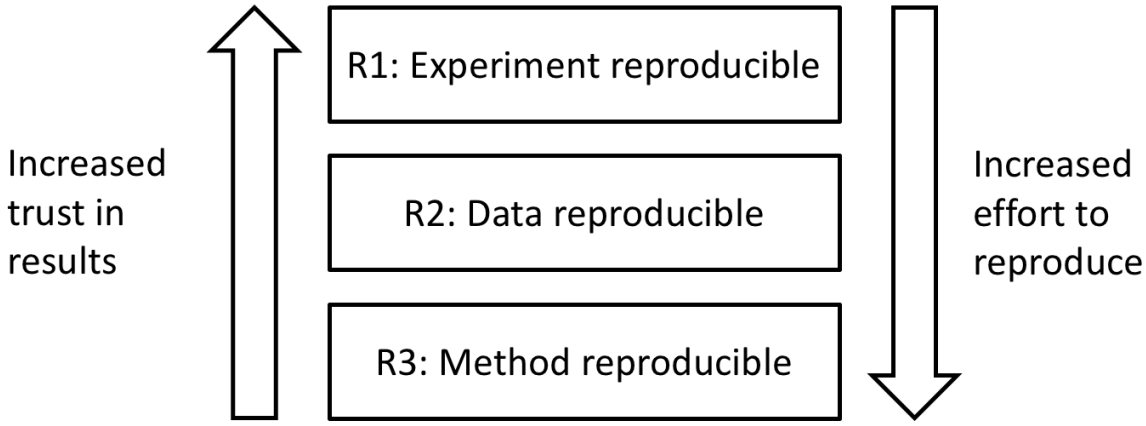

Figure 5: Effects of documentation as seen from the perspective of independent researchers.

On the other hand, the effort to document the research increases for the original researchers with the amount of documentation that needs to be shared, while the generality of the method is increased if independent researchers reproduce the results given less documentation. Hence, the original researchers may prefer to document their research to be R3 (Method Reproducible).

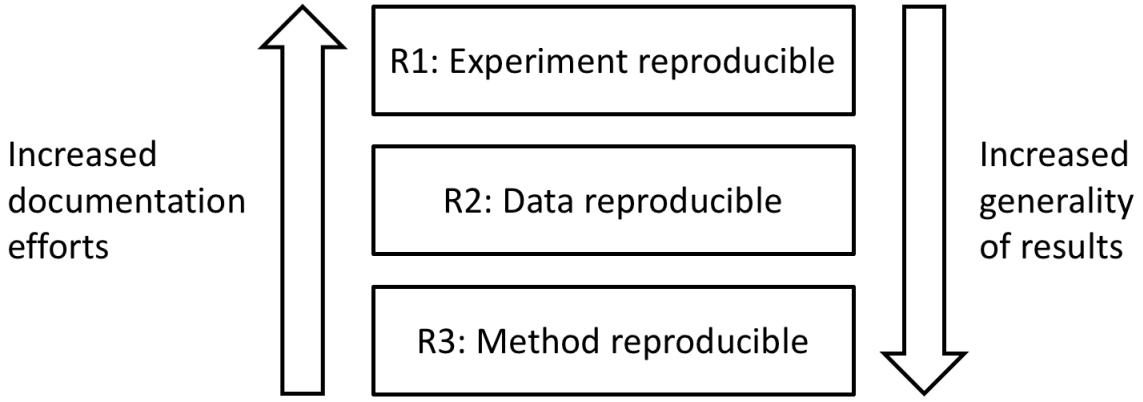

Figure 6: Effects of documentation as seen from the perspective of the original researchers.

Combine this conflict of incentives for the original and independent researchers with the pressure to publish and it is easy to see how this can lead to research being documented less vigorously. However, by following the recommendations given here the trustworthiness and reproducibility of research results can be increased with little effort required from authors. Still, changes cannot be expected solely from individual researchers alone. The research community, funding sponsors, employers of researchers, and publishers should , in their respective roles, with incentivize and reward reproducible research.

4. Best Practices and Recommendations

The recommendations we introduce in Sections 4.1-4.4 are based on best practices put forward by scientific organizations (RDA 2015; CODATA 2013; DataCite 2015; FORCE11 2014; ESIP 2012), scholars (Wilkinson et al. 2016; Stodden et al. 2016; Gil et al. 2016; Nosek et al. 2015; Starr et al. 2015; Uhlir et al. 2012; Downs et al. 2015; Ball and Duke 2012; Mooney and Newton 2012; Goodman et al. 2014; Garijo et al. 2013; Altman and King 2007), and publishers (Hanson et al. 2015; COPDESS 2015).

Strong momentum is building in support of FAIR practices, i.e., to make data Findable, Accessible, Interoperable, and Reusable Wilkinson et al. 2016). Our recommendations support FAIR principles and extend them to promote reproducible research, open science, and digital scholarship.

Implementing these recommendations requires extra space in publications. We suggest including this additional content in appendices that technical reviewers will not be required to assess but can quickly check. For electronic publications, there should not be any space limitations imposed for such appendices.

When these recommendations cannot be met, a brief explanation should be included about the reasons. Possible reasons may be restricted access (e.g., proprietary or

sensitive data), ownership by close collaborators who do not wish to disclose certain details, inadequate resources (e.g., to house large datasets), or an unreasonable burden on authors.

We begin with recommendations for data (Section 4.1) and source code (Section 4.2) as the basic ingredients of a computational experiment. Then we describe recommendations to document AI methods (Section 4.3) and the experiments themselves (Section 4.4). If all recommendations for AI methods (Table 4) are implemented, then the publication should in theory be R3 (Method Reproducible), while if all recommendations for data (Table 2) are also implemented, then the research should be R2 (Data Reproducible). Finally, all four sets of recommendations (Tables 2-5) must be implemented for the research to be fully R1 (Experiment Reproducible).

We will refer to the complete set of 20 recommendations as an *author checklis*t, we provide examples to demonstrate that they are synergistic, and we argue that they can be easily implemented.

4.1 Recommendations for Data

Table 2 summarizes our recommendations for documenting data, which concern: (1) repository use, (2) metadata, (3) licenses, (4) persistent unique identifiers, and (5) citations. They can be easily implemented if researchers use community data repositories that support recommended best practices.

Table 2. Author checklist (Part I): Recommendations for data in publications.

RECOMMENDATIONS 1-5: Data mentioned in a publication should:

1. *Be available in a shared community repository*, so anyone can access it
2. *Include basic metadata,* so others can search and understand its contents
3. *Have a license*, so anyone can understand the conditions for reuse of the data
4. *Have an associated digital object identifier (DOI) or persistent URL (PURL)* so that the data is available permanently
5. *Be cited properly in the prose and listed accurately among the references*, so readers can identify the datasets unequivocally and data creators can receive credit for their work

Data Repositories

Data repositories exist for many domains, and as such they are available to the AI community. Examples of these general repositories are Zenodo (Zenodo 2018), figshare (figshare 2018), and Dataverse (Dataverse 2018). These repositories will automatically assign a DOI to any uploaded data and will also accept software, figures, movies, and slide presentations. They will also inquire about choosing a license, and with specifying a descriptive name and authors for a submitted dataset. AAAI could, as a service, provide a list of recommended data repositories. This could

be modelled on a service provided by COPDESS, which is a large coalition for publishing data in the earth and space sciences (COPDESS 2015). Universities also offer general repositories, whether developed in house or as installations of general infrastructure such as Dataverse. University repositories are typically maintained by library departments, and always offer DOIs, licenses, and citations.

We encourage maintainers of data repositories that serve the AI community to adopt mechanisms for assigning DOIs or persistent URLs (PURLs) to datasets that they provide. The management of PURLs or DOIs can be complex. We suggest consulting with organizations such as FORCE11 and the Research Data Alliance, which have working groups with extensive and detailed recommendations on this topic.

Metadata
Basic metadata includes a descriptive title, the dataset's authors, and creation date. Additional metadata is always valuable to others in terms of understanding and reusing the dataset.

Licenses for Data
Recommended licenses for data are Creative Commons licenses (Creative 2018), preferably CC-BY (unlimited reuse as long as there is attribution) or CC0 (unlimited reuse without conditions).

Permanent Unique Identifiers for Data
Many authors make data available by providing a URL to their personal or lab pages. These references may not last long due to changes in sites and in author affiliations (Klein et al. 2014). Instead, we encourage authors to use persistent unique identifiers so that their data is always available. DOIs are managed by data repositories and given to individual datasets or to collections (DeRisi et al. 2013). Most data repositories provide DOIs, and for this they forge an agreement with a DOI authority. Another option that anyone can use is PURLS. PURLs can be assigned by anyone to any web resource using a trusted service such as the W3C's w3id (w3id 2018). Data repositories also have the option of using PURLs.

Data Citation
A data citation can be directly provided by a data repository, or it can be constructed by hand. A citation for a dataset consists of a descriptive name (or title) for the dataset, its creators, the name of the repository where it can be accessed, and the permanent URL. For example, a citation for a dataset in (Gil et al. 2017) is:

> *Adusumilli, Ravali. (2016). Sample datasets used in (Gil et al. 2017) for AAAI 2017 (Data set). Zenodo. http://doi.org/10.5281/zenodo.180716.*

Note that by simply uploading the dataset to the Zenodo repository we obtained the DOI and the citation. Specifying the authors, the name, and the license take negligible effort. The author checklist for data required little time to implement.

4.2 Recommendations for Source Code

We refer to *source code* as the human readable computer instructions written in plain text and *software* as computer programs that are executable by a computer. Typically, source code is compiled to software for a computer to run it. Our recommendations for source code are summarized in Table 3.

Table 3. Author checklist (Part II): Recommendations for source code implementing AI methods and experiments in publications.

RECOMMENDATIONS 6-10: Source code used for implementing an AI method and executing an experiment should:

6. *Be available in a shared community repository*, so anyone can access it
7. *Include basic metadata,* so others can search and understand its contents
8. *Include a license*, so anyone can understand the conditions for use and extension of the software
9. *Have an associated digital object identifier (DOI) or persistent URL (PURL)* for the version used in the associated publication so that the source code is permanently available
10. *Be cited and referenced properly in the publication* so that readers can identify the version unequivocally and its creators can receive credit for their work

---

Source Code Repositories
Source code repositories can be used by any scientists to share code, and as such they are available to the AI community. These include general repositories such as GitHub and BitBucket, and language-specific repositories such as CRAN for R code or File Exchange in MATLAB Central. General data repositories such as those mentioned above accept source code as an entry, and as with any dataset they always offer DOIs, licenses, and citations.

For a specific publication, the version of the source code that is being used should be clearly specified, and the source code repository should support the identification and future access of specific versions.

Source Code Metadata
Basic metadata includes a descriptive title, the source code's authors, and the creation date. Additional metadata is always valuable to others in terms of understanding and reusing the source code.

Licenses for Source Code
Recommended licenses for source code are the standard licenses from the Open Source Initiative. Licenses such as Apache v2 or MIT are recommended because

they provide unlimited reuse (as long as there is attribution). Other more restrictive licenses are available to limit commercial use or impose licensing conditions on extensions of the original source code.

Permanent Unique Identifiers for Source Code
A separate DOI should be assigned to meaningful versions of the source code, such as a version used for a publication. GitHub offers an option to obtain a DOI for a source code version, which is done by storing that version permanently in the Zenodo data repository. Any source code can be uploaded manually to community data repositories such as Zenodo, figshare, and Dataverse. PURLS can be assigned by anyone to any source code version that has a URL on the Web, using a trusted service such as w3id.org.

Source Code Citation
Source code citation can be directly provided by a source code repository, or it can be constructed by hand. A citation for a source code version consists of a descriptive name (or title) for the source code, its creators, the name of the repository where it can be accessed, the version, and the permanent URL. For example, a citation for GitHub code in (Gil et al. 2017) is:

Ratnakar, Varun. "DISK software" (v1.0.0). Zenodo. 2016. http://doi.org/10.5281/zenodo.168079

By uploading the source code into the GitHub code repository, we obtained a persistent identifier for the version used in the publication as well as the citation. Specifying the authors, the name, and the license take negligible effort. Implementing the author checklist for source code required little time.

4.3 Recommendations for AI Methods

Our recommendations for AI methods are listed in Table 4.

Table 4. Author checklist (Part III): Recommendations for AI methods in publications.

RECOMMENDATIONS (11-13): AI methods used in a publication should be:

11. Presented in the context of a *problem description* that clearly identifies what problem they are intended to solve
12. *Outlined conceptually* so that anyone can understand their foundational concepts
13. *Described in pseudocode* so that others can understand the details of how they work

Problem Description

The problem that a conceptual AI method solves should be explicitly described in the publication. In (De Weerdt et al. 2013) the following example can be found: ”To address this problem, we propose a novel navigation system ...”. The authors explicitly describe the problem that they address. Another good example of this practice can be found in (He et al. 2016). Here the authors state the problem explicitly: ”In this paper, we address the degradation problem by introducing a deep residual learning framework.” The degradation problem is also properly described in their publication.

Conceptual Method
A high-level, textual description of the AI method should be provided to readers to allow them to gain an understanding of it. This description should include a broad overview of how the AI method works and specify input variables and the resulting output. In general, the AI research community excels at providing this information in publications.

Pseudocode
Pseudocode for the AI method should also be provided. In cases where detailed pseudocode cannot be provided due to the complexity of the proposed algorithm or system, a more abstract pseudocode summary can be provided that illustrates the AI method’s flow.

Both a high-level description and pseudocode help independent researchers to decide whether their own implementation of the method is correct. If these are not presented carefully, then the empirical study cannot always be easily reproduced.

4.4 Recommendations for Experiments

Authors should, to the degree possible, detail how their experiments are designed, and indicate the rationale for their design. Our recommendations for documenting experiments are summarized in Table 5.

Table 5. Author checklist (Part IV): Recommendations for experiments described in publications.

RECOMMENDATIONS (14-23): Descriptions of experiments in a publication should:

14. *Explicitly present the hypotheses to be assessed, before other details concerning the empirical study are presented*
15. *Present the predicted outcome of the experiment, based on beliefs about the AI method and its application*
16. Include the *experimental setup* (parameters and the conditions to be tested) and its *motivation,* such as why a specific number of tests or data points are used based on the desired statistical significance of results and the availability of data
17. *Present the results (i.e., measures and metrics) and the analysis*

18. *An explicit indication of whether the analyses support the hypotheses*
19. *Justify* why the datasets used are appropriate for the experiment, why the chosen empirical design is appropriate for assessing the hypothesis, and why the metrics and measures are appropriate for assessing the results
20. Be described as a *workflow* that summarizes how the experiment is executed and configured
21. Include *documentation on workflow executions or execution traces* that provide parameter settings and initial, intermediate, and final data
22. *Specify the hardware used to run the experiments*
23. Be *cited and published separately when complex*, so that others can unequivocally refer to the individual portions of the method that they reuse or extend

---

Hypotheses and predictions
Hypotheses and predictions should be stated explicitly before describing the other components of an empirical study to ensure that the results analysis is meaningful (Baker 2016).

Empirical design
A textual description and justification of the experiment's design should be provided, to include a description of each test condition. This should also describe, for example, why a specific number of tests or data points are used based on the desired statistical significance of the results and the availability of data.

Data sets
Researchers should justify the use of their selected data sets.

Evaluation protocol
A justification should be provided for the chosen protocol when documenting an experiment. To avoid hypothesis myopia, this should not be designed to collect only evidence that is guaranteed to support the stated hypotheses. Instead, to encourage an insightful study, this should include conditions the could lead to the rejection of these hypotheses. The measures and metrics to be used to evaluate the research must be described, and so should the analysis procedure(s) (e.g., for assessing statistical significance) be as well.

Results and analysis
The results and the analysis should be presented in detail.

Workflow
This workflow should describe, in a machine-readable way, how software and data are used to implement the evaluation protocol. A workflow step is an invocation of the software. Each step has input data and parameters as well as output data. Input data and the output of any step can be used as input to subsequent steps. The

simplest workflow languages capture methods that are directed acyclic graphs, while other languages can represent iterations and conditionals. A publication that simply mentions what software was used usually leaves out critical information about how the software was configured or invoked.

Scripts or electronic notebooks can be an effective way to document workflows, although the organization of source code is more modular in a workflow structure.

Executions
A general workflow can be run many times with different datasets or parameter settings and generate different results. Execution traces of executed workflows provide a complete provenance trail of how each result was generated.

Hardware specification
The hardware that is used should be specified if this is important for the experiment. This may include specification of the processor type, the number of cores and processors, RAM and hard disk requirements. Also, the provider of the cloud solution that is used, if any, should be specified. The machine architecture and operating system may need to be specified, so that any discrepancies in results can be properly diagnosed. Library dependencies should also be described. Virtualization technologies, such as docker and Kubernetes, facilitate these specifications through artifacts called containers. Containers can be provided as appropriate to share the experiment hardware setup.

Workflow citation
Citing a publication does not make explicit whether the citation is to its AI method, source code, data, empirical design, workflows, execution traces, results, or a general body of work or contributions. If it is important that others are explicit about what aspects of the work are being reused, then separate citations should be given to each, as appropriate. Although workflow repositories are not as common as data and software repositories, many general data repositories accept any research product and can be used for this purpose.

For example, a citation for a bundle containing workflows and execution details for (Gil et al. 2017) is:

> *Adusumilli, Ravali, Ratnakar, Varun, Garijo, Daniel, Gil, Yolanda, and Mallick, Parag. (2016). Additional materials used in the paper "Towards Continuous Scientific Data Analysis and Hypothesis Evolution" on the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) (Data set). Zenodo. http://doi.org/10.5281/zenodo.190374*

By organizing the workflows and executions described into this publication and bundling them to upload to a general data repository, these authors obtained a persistent identifier as well as a citation. The author checklist for experiments was implemented quickly.

5. Benefits to Authors

Recognizing that our recommendations will require effort from authors, we want to highlight the following benefits:

1. *Practice open science and reproducible research*. This ensures the kinds of checks and balances that lead to better science.
2. *Receive credit for all your research products* (i.e., through citations for software, datasets, and other products).
3. *Increase the number of citations to your publications*. Studies have shown that well-documented articles receive more citations (Piwowar et al. 2007).
4. *Improve your chances of being funded* (i.e., by writing coherent and well-motivated empirical study and data management plans).
5. *Extend your CV*. Include data and software sections with citations. Maintaining data sets and writing code are important contributions to the field of AI.
6. *Improve the management of your research assets* (e.g., so your new students, and others, can more easily locate materials generated by your earlier students).
7. *Allow for the reproduction of your work* (e.g., so you and others can leverage it in new studies, even if it was conducted many years ago).
8. *Address new sponsor and journal requirements*. They are steadfastly driving research towards increased reproducibility and open science.
9. *Attract transformative students*. They strive for a rigorous research methodology.
10. *Demonstrate leadership*. Step into the future.

By explicitly citing datasets and source code, and by providing workflows that are machine readable, we create the structure needed that can allow for the development of AI systems that can analyze and reason about our literature (Gil 2017). These AI systems would have access to a vast amount of structured scientific knowledge with comprehensive details about experimental design and results. This could revolutionize how we approach the scientific research process.

6. Discussion

It is reasonable to expect a limited release of data and source code until the creator has completed the research for which the data was collected, or for which the source code was written, or until their draft is published. Many journals impose this, such as *Science* and *Nature*. See (Joly et al. 2012) for a review of data retention policies.

The creation and documentation of additional information we recommend should be done by researchers who publish their studies. By documenting and sharing code and data in such a way that they can be easily used and cited by others gives researchers credit for a larger portion of their research effort. For academic researchers, we advocate that tenure committees give weight to the publication of data and source code when evaluating candidates for tenure. Thus, the publication velocity should not be reduced, but include research products other than publications.

The recommendations we suggest should be a part of daily research practices. According to Irakli Loladze, despite increasing work load by 30%, "Reproducibility is like brushing your teeth. It is good for you, but it takes time and effort. Once you learn it, it becomes a habit" (Baker 2016).

Another recommendation for improving the readability and comparability of research papers is to require structured abstracts, which are commonly used in medical journals. Structured abstracts can be used to efficiently communicate a research objective, the motivation for and process by which an empirical study was conducted, and what results were achieved. Structured abstracts also require researchers to structure their own thoughts about their research. We suggest a five-part structured abstract containing (1) the research motivation, (2) the research objective, (3) the method used to conduct any empirical studies, (4) the results of the research, and (5) the conclusion. This structure enforces a coherent research narrative, which is not always the case for unstructured abstracts. The abstract for this article is an example of the proposed structure, while (Gundersen and Kjensmo 2018) provides an abstract for empirical research that follows these recommendations and includes an explicit description of the hypothesis and an interpretation of the results.

7. Call to arms

As a community, we should ensure that the research that we conduct is properly documented. To make AI research reproducible and more trustworthy, we proposed best practices that should be adopted by editors and program chairs and incorporated into the review forms of AAAI publication venues.

Publishers should provide extra space to document and cite data, source code, and empirical study designs. AAAI leadership should encourage AI researchers to increase the reproducibility of their published work. This could include providing structured templates to organize appendices and extra space in publications to accommodate the needed documentation.

For AI research to become open and more reproducible, the research community and publishers have to establish suitable practices. Authors need to

adopt these practices, disseminate them to colleagues and students, and help develop mechanisms and technology to make it easier for others to adopt them.

Our objective with this article is to highlight the benefits of reproducible science, and propose initial, modest changes that can increase the reproducibility of AI research results. There are many additional actions that could and should be taken, and we look forward to further dialogue with the AI research community on how to increase the reproducibility and scientific value of AI publications.

ACKNOWLEDGEMENTS

This research was funded in part by the National Science Foundation under grant ICER-1440323. This work has in part been carried out at the Telenor-NTNU AI Lab, Norwegian University of Science and Technology, Trondheim, Norway.
The recommendations proposed are based on the Geoscience Paper of the Future and the Scientific Paper of the Future best practices developed under that award. Thanks to Sigbjørn Kjensmo for all the effort put into surveying the state of the art of reproducibility of AI.

REFERENCES

(Altman and King 2007) “A proposed standard for the scholarly citation of quantitative data.” Altman, M., and King, G. D-Lib Magazine, 13(3/4). doi:10.1045/march2007-altman

(Baker, 2016) "Is there a reproducibility crisis?." Monya Baker. Nature, 533. May 2016. DOI: doi:10.1038/533452a

(Ball and Duke 2012) “How to Cite Datasets and Link to Publications”. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides - See more at: http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sthash.MJQjNn3i.dpuf

(Begley and Ellis 2012) "Drug development: Raise standards for preclinical cancer research." Begley, C. G., and Ellis, L. M. Nature 531. March 2012. DOI: 10.1038/483531a

(Braun and Ong 2014) Braun, M. L. and Ong, C.S. Open science in machine learning. In Implementing Reproducible Research, page 343. CRC Press. 2014.

(CODATA 2013) “Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data.” CODATA-ICSTI Task Group on Data

Citation Standards and PractOut of Cite, Out of Mind: The Current Sices . Data Science Journal, 2013. DOI: 10.2481/dsj.OSOM13-043

(COPDESS 2015) “Statement of Commitment from Earth and Space Science Publishers and Data Facilities.” Coalition on Publishing Data in the Earth and Space Sciences (COPDESS). January 14, 2015. http://www.copdess.org/statement-of-commitment/

(Creative 2018) Creative Commons. Available from https://creativecommons.org. Last accessed 18 May 2018.

(DataCite 2015) DataCite. Available from https://www.datacite.org/. Last accessed 3 August 2015.

(Dataverse 2018) The Dataverse project. Available from https://dataverse.org. Last accessed 18 May 2018.

(De Weerdt et al. 2013) "Intention-aware routing to minimise delays at electric vehicle charging stations." De Weerdt, M. M., Gerding, E. H., Stein, S., Robu, V., and Jennings, N. R. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 83–89. AAAI Press, 2013

(DeRisi et al 2013) “The What and Whys of DOIs.” Susanne DeRisi, Rebecca Kennison, Nick Twyman. PLoS Biology 1(2): e57, 2013.

(Downs et al. 2015) “Data Stewardship in the Earth Sciences.” Robert R. Downs, Ruth Duerr, Denise J. Hills, and H. K. Ramapriyan. D-Lib Magazine, 21(7/8). doi:10.1045/july2015-downs

(ESIP 2012) “Interagency Data Stewardship/Citations/provider guidelines.” Federation of Earth Science Information Partners (ESIP), 2 January 2012. Available from http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelin

(figshare 2018) figshare. Available from https://figshare.com. Last accessed 18 May 2018.

(Fokkens et al. 2013) "Offspring from repro- duction problems: What replication failure teaches us". Fokkens, A., Erp M. V., Postma, M., Pedersen, M., Vossen, P., and Freire, N. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1691–1701. Association for Computational Linguistics (ACL), 2013.

(FORCE11 2014) Joint Declaration of Data Citation Principles. Martone M. (ed.) and the Data Citation Synthesis Group, San Diego CA: FORCE11 2014. Available from https://www.force11.org/datacitation.

(Garijo et al. 2013) "Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome" Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. PLOS ONE, 27 November 2013.

(Gil 2017) "Thoughtful Artificial Intelligence: Forging A New Partnership for Data Science and Scientific Discovery." Yolanda Gil. Data Science (1):1-2, 2017. DOI:10.3233/DS-170011

(Gil et al. 2016) "Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance." Gil, Y.; David, C. H.; Demir, I.; Essawy, B. T.; Fulweiler, R. W.; Goodall, J. L.; Karlstrom, L.; Lee, H.; Mills, H. J.; Oh, J.; Pierce, S. A; Pope, A.; Tzeng, M. W.; Villamizar, S. R.; and Yu, X. Earth and Space Science, 3. 2016.

(Gil et al. 2017) "Towards Continuous Scientific Data Analysis and Hypothesis Evolution." Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, Arunima Srivastava, and Parag Mallick. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, 2017.

(Goodman et al. 2014) Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Stefano, R. D., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., and A. Slavkovic (2014), Ten simple rules for the care and feeding of scientific data, PLOS Computational Biology, 10, April 24, 2014, doi: 10.1371/journal.pcbi.1003542.

(Gundersen and Kjensmo 2018) "State of the Art: Reproducibility in Artificial Intelligence." Odd Erik Gundersen and Sigbjørn Kjensmo. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, 2018.

(Hanson et al. 2015) "Committing to Publishing Data in the Earth and Space Sciences." Brooks Hanson, Kerstin Lehnert, and Joel Cutcher-Gershenfeld. EOS 95, 15 January 2015. doi:10.1029/2015EO022207. https://eos.org/agu-news/committing-publishing-data-earth-space-sciences

(He et al. 2016) "Deep residual learning for image recognition." He, K., Zhang, X., Ren, S., and Sun, J. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

(Henderson et al. 2017) Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2017). Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*.

(Hunold 2015) "A survey on reproducibility in parallel computing". Hunold, S. *CoRR, abs/1511.04217*, 2015.

(Hunold and Träff 2013) "On the state and importance of reproducible experimental research in parallel computing". Hunold, S. and Träff, J. S. *CoRR, abs/1308.3648*, 2013.

(Ioannidis 2005) "Why most published research findings are false." Ioannidis, J. P. PLoS Medicine. August 2005. DOI: 10.1371/journal.pmed.0020124

(Joly et al. 2012) "Open science and community norms: Data retention and publication moratoria policies in genomics project." Yann Joly, Edward S. Dove, Karen L. Kennedy, Martin Bobrow, B.F. Francis Ouellette, Stephanie O.M. Dyke, Kazuto Kato, and Bartha M. Knoppers. Medical Law International. Vol 12, Issue 2, pp. 92 - 120. October 9, 2012 DOI: 10.1177/0968533212458431

(Klein et al 2014) Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. (2014) "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot." PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253

(Lithgow et al. 2017). "A long journey to reproducible results." Lithgow, G. J., Driscoll, M., and Phillips, P. Nature News. August 2017. DOI: 10.1038/548387a

(Mooney and Newton 2012) Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. Journal of Librarianship and Scholarly Communication 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035

(Nosek et al. 2015) "Promoting an open research culture." B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni. Science 348, 1422-1425, 26 June 2015. DOI: 10.1126/science.aab2374

(Piwowar et al 2007) “Sharing Detailed Research Data Is Associated with Increased Citation Rate.” Heather A. Piwowar, Roger S. Day, Douglas B. Fridsma. PLoS ONE, March 21, 2007. DOI: 10.1371/journal.pone.0000308

(RDA 2015) Outcomes of the Research Data Alliance (RDA). Available from https://rd-alliance.org/outcomes. Last accessed July 30, 2015.

(Starr et al. 2015) “Achieving human and machine accessibility of cited data in scholarly publications.” Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. PeerJ Computer Science 1:e1, 2015. DOI: 10.7717/peerj-cs.1

(Stodden et al. 2016) “Enhancing reproducibility for computational methods.” Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer. Science 354, 1240 (2016) DOI:10.1126/science.aah6168

(Uhlir et al. 2012) “For Attribution: Developing Data Attribution and Citation Practices and Standards.” Paul F. Uhlir, Rapporteur; Board on Research Data and Information; Policy and Global Affairs; National Research Council. Report of CODATA Data Citation Workshop. National Academies Press, 2012. Available from http://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards.

(Wilkinson et al. 2016) “The FAIR Guiding Principles for scientific data management and stewardship.” Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. Nature Scientific Data 3, 2016. doi:10.1038/sdata.2016.18

(w3id 2018) ”Permanent Identifiers for the Web.” World Wide Web Consortium (W3C), Available from http://www.w3id.org

(Zenodo 2018) Zenodo. Available from https://zenodo.org. Last accessed 18 May 2018.

Odd Erik Gundersen (PhD, NTNU) is an adjunct associate professor at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, where he teaches courses and supervises master students in AI. Gundersen has applied AI in the industry, mostly for startups, since 2006. Currently, he investigates how AI can be applied in the renewable energy sector and for driver training.

Yolanda Gil (PhD, CMU) is Director of Knowledge Technologies and Associate Division Director at the Information Sciences Institute of the University of Southern California, and Research Professor in Computer Science and in Spatial Sciences. Her research is on intelligent interfaces for knowledge capture and discovery, semantic workflows and metadata capture, social knowledge collection, computer-mediated collaboration, and automated discovery. Dr. Gil has served in the Advisory Committee of the Computer Science and Engineering Directorate of the National Science Foundation. She initiated and chaired the W3C Provenance Group that led to a community standard in this area. Dr. Gil is also Fellow of the AAAI and was elected as its 24th President in 2016.

David W. Aha (PhD, UC Irvine, 1990) leads a section within NRL's Navy Center for Applied Research in AI, in Washington, DC. In addition to encouraging reproducible research, his interests include mixed-initiative intelligent agents, deliberative autonomy, explainable AI, case-based reasoning, and machine learning. He has co-organized many events on these topics, launched the UCI Repository for ML Databases, served as an AAAI Councilor, co-created AAAI's AI Video Competition, and currently leads the evaluation team for DARPA's Explainable AI program.