Standing on the Feet of Giants — Reproducibility in AI

Odd Erik Gundersen

■ *A recent study implies that research presented at top artificial intelligence conferences is not documented well enough for the research to be reproduced. My objective was to investigate whether the quality of the documentation is the same for industry and academic research or if differences actually exist. My hypothesis is that industry and academic research presented at top artificial intelligence conferences is equally well documented. A total of 325 International Joint Conferences on Artificial Intelligence and Association for the Advancement of Artificial Intelligence research papers reporting empirical studies have been surveyed. Of these, 268 were conducted by academia, 47 were collaborations, and 10 were conducted by the industry. A set of 16 variables, which specifies how well the research is documented, was reviewed for each paper and each variable was analyzed individually. Three reproducibility metrics were used for assessing the documentation quality of each paper. The findings indicate that academic research does score higher than industry and collaborations on all three reproducibility metrics. Academic research also scores highest on 15 out of the 16 surveyed variables. The result is statistically significant for 3 out of the 16 variables, but none of the reproducibility metrics. The conclusion is that the results are not statistically significant, but still indicate that my hypothesis probably should be refuted. This is surprising, as the conferences use double-blind peer review and all research is judged according to the same standards.*

Traditionally, artificial intelligence (AI) research has been conducted by academia, but lately there has been a shift toward the technology industry. One indication of this is the fact that leading academics, such as Geoffrey Hinton, Yann LeCun, and Zoubin Ghahramani, double as academics and industry experts. A second indication is the number of industry sponsors that the large AI conferences manage to secure. Large companies such as Google, Intel, Tencent, Facebook, Baidu, Microsoft, Disney, Sony, JP Morgan, Amazon, IBM, and many more line up to sponsor conferences such as those of the Association for the Advancement of Artificial Intelligence (AAAI), the International Joint Conferences on Artificial Intelligence (IJCAI), and the International Conference on Machine Learning. Just compare IJCAI 2018 sponsors to those from 2011, or AAAI 2018 sponsors to those from 2012. There were more sponsors in 2018, and they were, to a larger degree, global, rather than local, companies. A third indication is how much harder it has become to hire and keep qualified people skilled in machine learning and AI, as "demand for software engineers with AI expertise continues to increase, while supply flattens."[1] Finally, some of the recent results in AI research that have a big impact on society — so that even mass media reports on it — are the result of industry research.

In theory, one could expect that this movement of the center of gravity of AI toward industry would lead to more secrecy and closed-down AI research, and that the industry would see the methods that they develop and use, as competitive advantages. In practice, however, this is not the case. The AI and machine learning software that is most commonly used by the community is developed by tech giants such as Google, Facebook, and Microsoft. Examples include PyTorch and Caffe, which were developed by Facebook; TensorFlow, which was developed by Google; and Cognitive Toolkit, which was developed by Microsoft. This software is free of charge, and even open-source. The tech giants not only share the software they develop; they also publish the wide variety of research they conduct at top conferences and in journals. Topics range from deep reinforcement learning (Silver et al. 2017), machine translation (Lample et al. 2018; Ott et al. 2018), and vision-to-language for people who are blind (Salisbury, Kamar, and Morris 2018), to machines that learn and think for themselves (Botvinick et al. 2017).

One of the primary reasons that industry is interested in AI is because of digitization and the huge growth in data generated by internet usage and Internet of Things sensors as well as the introduction of methods, mainly deep neural networks, that are capable of utilizing all the data that is owned by these companies. There is a saying that *data is the new oil*,[2] and hence a valuable asset. This could indicate that the industry is a bit less eager to share data than software, as machine learning software to a large degree is only as good as the data it is trained on. By sharing the software that is developed and used internally in a company, the companies not only become thought leaders, they also prepare potential employees to become efficient workers even before they apply for a job. Allowing employees to publish research not only keeps employees happy, it is also a marketing tool. The companies that publish research at top conferences are looked at as innovative and interesting companies to work for. Apple even changed their policy on not publishing research papers to make itself competitive when it comes to hiring AI and machine learning talents, according to a paper by Steven Levy in *Wired*.[3] Hence, allowing employees to publish their research is a means for attracting top talent. Importantly, high-quality research also builds the reputation of a company, which again could be used to increase sales. So, it is clear that sharing software and publishing research is advantageous, while sharing data comes at a higher risk with regard to attaining and keeping competitive advantage.

Industry research is submitted to the same tracks as academic research, and it is judged according to the same standards. This indicates that the quality of the documentation of the research should be the same for industry and academe.

However, to maintain the edge over the competition, a strategy might be to keep some important details of the research from the research papers that are put into the public domain. By doing this, competitors might spend time and resources on pondering important details when trying to reproduce the results. Hence, it could be expected that the quality of the industry research documentation is lower than quality of academic research documentation, although academics also could have incentives for keeping some parts to themselves. This begs the question of whether the empirical research presented by academic and industry researchers at the same conferences have the same quality of documentation. Is the quality of the documentation of AI methods presented by academia and industry the same, or are there actual differences? Do industry researchers share less data? Does the industry specify the experiments and hyperparameter settings as thoroughly as is done by academia? Does industry share the code for the experiments, or only the code implementing the AI methods?

My objective is to investigate whether the quality of the documentation is the same for industry and academic research. Are there any differences between the experiment documentation made by industry and academia, and if so, what are these differences? I investigate the hypothesis that empirical research presented at top AI conferences is equally well documented whether the research is conducted by industry or academe. Given the analysis above, my prediction is that the documentation of academic research is better than industry research. My contribution is a comparison of the documentation quality of AI research presented at four installments of the top two AI conferences, IJCAI and AAAI, followed by a discussion of the results.

Reproducibility

According to an paper by Gundersen and Kjensmo (2018), reproducibility in empirical AI research is the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team. The key is that an independent research team should produce the same results as the original team based only on the documentation created by the original team. Hence, the documentation is the enabler for the independent team to ensure that they actually conduct the exact same experiment as the original team. In AI research, the documentation has three components: the documentation of the AI method that the original research team has developed and wants to test; the experiment description, which is written both as text and as code; and the data that are used for evaluating the AI method.

The grouping of the documentation allows Gundersen and Kjensmo (2018) to define three degrees to which the original results can be reproduced: *R1: Experiment reproducible* means that 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. *R2: Data reproducible* means that 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. *R3: Method reproducible* means that the results of an experiment are method-reproducible when the execution of an alternative implementation of the AI method produces the same results when executed on different data.

Figure 1 illustrates how the three degrees relate and which degree requires which documentation. When an independent research team conducts research based on a description of the AI method, the experiment implementation, and the data provided by the original team, the results are less generalizable than if the independent team only get the description of the AI method from the original team and have to implement the method themselves and conduct the experiment on different data. There is a conflict between the incentives for the original and independent research teams, as an independent team has higher trust in research documented at a lower reproducibility degree while the original team would like independent researchers to reproduce the results with less documentation to prove generalizability. This conflict of interest is discussed in more detail in Gundersen, Gil, and Aha (2018).

Several definitions of reproducibility exist in the literature. Stodden (2011) distinguishes between replication and reproduction; replication is seen as rerunning the experiment with code and data provided by the author, while reproduction is a broader term "implying both replication and the regeneration of findings with at least some independence from the [original] code and/or data" (p. 22). Drummond (2009) states that replication, as the weakest form of reproducibility, can only achieve checks for fraud. Due to the inconsistencies in the use of the terms replicability and reproducibility, Goodman, Fanelli, and Ioannidis (2016) propose to extend reproducibility into methods reproducibility, results reproducibility, and inferential reproducibility: *methods reproducibility* is the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results; *results reproducibility* is the production of corroborating results in a new study, having used the same experimental methods; and *inferential reproducibility* is the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.

Replication, as used by Drummond (2009) and Stodden (2011), is in line with methods reproducibility as proposed by Goodman, Fanelli, and Ioannidis (2016), while reproducibility seems to entail both results reproducibility and inferential reproducibility. Peng (2011), on the other hand, suggests that reproducibility is on a spectrum from publication to full replication. This view neglects that results produced by AI methods can be reproduced using different data or different implementations. Results generated by using other implementations or other data can lead to new interpretations, which broadens the beliefs about the AI method, so that generalizations can be made. Despite the disagreements in terminology, there is a clear agreement on the fact that the reproducibility of research results is not just one *thing*, but that empirical research can be assigned to some sort of spectrum, scale, or ranking that is decided based on the level of documentation.

The degrees proposed by Gundersen and Kjensmo (2018) differ from the degrees suggested by Stodden (2011), Goodman, Fanelli, and Ioannidis (2016), and Peng (2011) in that the degrees are based on the different types of documentation that document a computer science experiment. In this way, one can specify the information that is required of the different types of documentation to enable reproducibility. This can even be tested empirically. It also allows the research community to discuss what needs to be documented and in the end — maybe — agree on a specification of what needs to be documented for an experiment to be reproducible.

Research Method

I have conducted an observational experiment in the form of a survey of research papers to generate quantitative data about the state of documentation quality of AI research. The research papers have been reviewed, and a set of 16 variables have been manually registered. To compare results between papers and groups of papers, I use three reproducibility metrics — R1D, R2D, and R3D — to score the documentation quality. I use the same research method and data (with some small revisions) that were used by Gundersen and Kjensmo (2018). The revised data set and the code for analyzing the data are shared online.[4]

Survey

To evaluate the hypothesis, I have surveyed a total of 400 papers where 100 papers have been selected from each of the 2013 and 2016 installments of the conference IJCAI and from the 2014 and 2016 installments of the conference series AAAI. With an exception of 50 papers from IJCAI 2013, all the papers have been selected randomly to avoid any selection biases. Table 1 shows the number of accepted papers (the population size), the number of surveyed papers (sample size), and the margin of errors for a confidence level of 95 percent for the four conferences. I have computed the margin of error as half the width of the confidence interval; for this study, the margin of error is 4.29 percent.

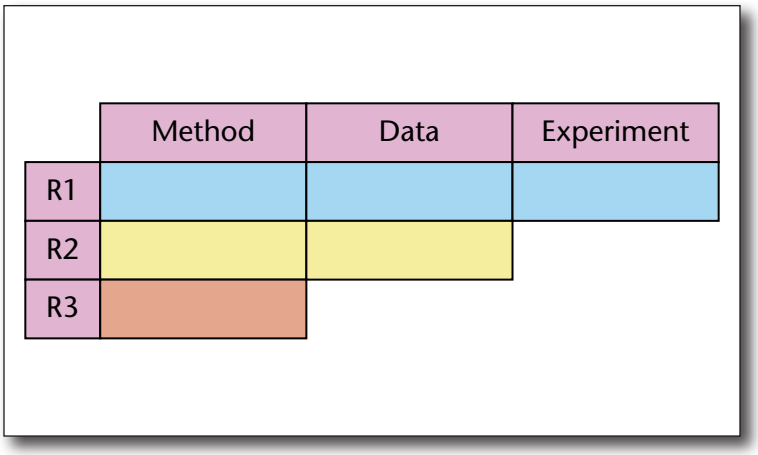

Figure 1. The Three Degrees of Reproducibility Are Defined by Which Documentation Is Used to Reproduce the Results.

Conference	Population Size	Sample Size	Margin of Error
IJCAI 2013	413	100 (71)	8.54%
AAAI 2014	213	100 (85)	7.15%
IJCAI 2016	551	100 (84)	8.87%
AAAI 2016	549	100 (85)	8.87%
Total	1726	400 (325)	4.30%

Table 1. Population Size, Sample Size (with Number of Empirical Studies), and Margin of Error.

Confidence level is 95 percent for the four conferences and total population.

Factors and Variables

The three types of documentation, Method, Data, and Experiment, are treated as factors that are specified by 16 different variables. The factors and variables that are used in the analysis are presented in figure 2. For each surveyed paper, I have registered the listed variables. All variables were registered as true (1) or false (0). When surveying the papers, I looked for explicit mentions of some of the variables: Problem, Objective, Research method, Research questions, Hypothesis, and Prediction. For example, when reviewing the variable Problem, I have looked for an explicit mention of the problem being solved, such as "To address this problem, we propose a novel navigation system …" (de Weerdt et al. 2013, p. 83). The reasons for this choice are discussed by Gundersen and Kjensmo (2018).

It should be noted that although both the variables and the factors are the same as in the paper by Gundersen and Kjensmo (2018), I have moved three variables (hypothesis, prediction, and experiment setup) from the factor Experiment to the factor Method. The reason for this change is that reproducing results based only on the factor Method requires the experiment to be described in the textual documentation. This change affects the calculation of the reproducibility metrics.

Quantifying Reproducibility

I have defined three metrics to quantify whether an experiment e is R1-, R2-, or R3-reproducible, and to what degree. The metrics $R1D(e)$, $R2D(e)$, and $R3D(e)$ measure how well the three factors, Method, Data, and Experiment, are documented for experiment e:

$$R1D(e) = \left(\delta_1 Method(e) + \delta_2 Data(e) + \delta_3 Exp(e)\right) / \left(\delta_1 + \delta_2 + \delta_3\right), \quad (1)$$

$$R2D(e) = \left(\delta_1 Method(e) + \delta_2 Data(e)\right) / \left(\delta_1 + \delta_2\right), \quad (2)$$

$$R3D(e) = Method(e), \quad (3)$$

where $Method(e)$, $Data(e)$, and $Exp(e)$ are the weighted sums of the truth values of the variables listed under the three factors Method, Data, and Experiment. The weights of the factors are δ_1, δ_2, and δ_3, respectively. This means that the value for $Data(e)$ for experiment e is the summation of the truth values for whether the training, validation, and test data sets as well as the results are shared for e. It is of course also possible to give different weights to each variable of a factor. I use a uniform weight for all variables and factors for this survey, $\delta_1 = 1$. For an experiment e_1 that has published the training data and test data, but not the validation set and the results, $Data(e_1) = 0.5$. Note that some papers have no value for the training and validation sets if the experiment does not require them. For these papers, the δ_i weight is set to 0.

Results

I have investigated how academic research compares to industry and collaborations between academia and industry. A total of 325 papers documenting empirical research were surveyed. Out of these, 268 documented research conducted by authors with academic affiliations, 10 were done by authors from industry alone, and 47 were collaborations where some authors were from academia and some from industry (see figure 3). As only 10 of the 325 papers were from industry, the errors in this analysis are high and the results are highly uncertain.

To reduce the uncertainty in the results, I grouped industry and collaborations in a group I called $C + I$. Here, I interpret this group to represent the research in which industry has partaken. This group include all collaborations between academia and nonacademic entities, of which private research institutions (such as the Allen Institute for AI), government institutions (such as the New York State Department of

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?
	Hypothesis	Is there an explicit mention of the hypotheses being investigated?
	Prediction	Is there an explicit mention of predictions related to the hypotheses?
	Experiment setup	Are the variable settings shared, such as hyperparameters?
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	Method source code	Is the AI system code available open source?
	Experiment source code	Is the experiment code available open source?
	Software dependencies	Are software dependencies specified?
	Hardware	Is the hardware used for conducting the experiment specified?

Figure 2. Method, Data, and Experiment, and the Variables That Specify Them.

Health), and industry (such as IBM and Microsoft) are examples. Only eight of the papers in the collaboration group are from collaborations between private research and government institutions. In this study, I present the results from collaboration studies and industry studies as well, despite small sample sizes.

Variables

Table 2 presents the mean values for the eight variables comprising the factor Method for each group of papers. Industry scores highest on the variables Problem description, Goal, and Experiment setup, while the combination (C + I) of collaborations and industry have the same score as academic for Problem description. Academic research scores higher than industry, collaboration, and the combination for Research method, Research question, Pseudo code, and Prediction. None of these results are statistically significant. Academic research also scores highest on Hypothesis, and this is statistically significant.

Table 3 presents the mean values for the four variables comprising the factor Data for each of the groups of papers. Academic research has the highest score for Training data. The result for this variable is statistically significant when compared with industry and the combination. Academia also has the highest score for Validation data and Test data as well, but these results are not statistically significant. Industry has the highest score for Results, and C + I has a lower score than academia. None of these findings are statistically significant.

Table 4 presents the mean values for the four variables comprising the factor Experiment for each of the groups of papers. Academic research scores highest on Hardware specification, and this result is statistically significant when compared with C + I. Industry has the best score on Method code, Experiment code, and Software dependencies. However, the confidence is low as the error is very high. The scores for C + I are lower for all these variables when comparing to academic research.

Factors

Figure 4 shows three spider plots of the mean for the variables of each of the three factors for all the

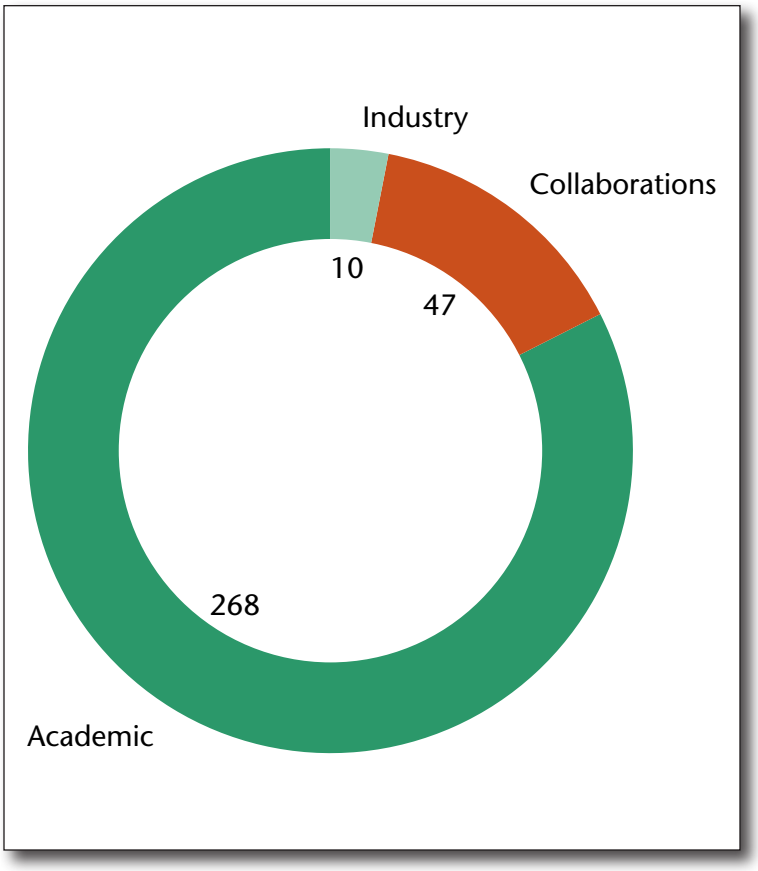

Figure 3. How Academic Research Compares to Industry and Academic-Industry Collaboration.

Of the 325 empirical papers that were surveyed, 265 of them were written by researchers from academe only, 47 were collaborations by academia and industry, and 10 had authors from industry alone.

surveyed empirical research, while figure 5 shows the same for the combination (C + I) and academic research. When comparing the outline of the spider plots for academia and all, one can see that they have very similar forms. This is no surprise, as academic research comprises 81.5 percent of all papers. Figure 5 shows that academic research has higher or equal scores on all variables for the factors Method, Data, and Experiment as the plots fully envelop the plots for the C + I research.

An observation is that most of the scores are quite low. The only variables scoring higher than 50 percent are Pseudo code, Experiment setup, and Training data. Pseudo code is very good for conveying an AI method in a concise way, so this is very positive. The fact that 56 percent of the research papers share the training data is also very positive. Experiment setup is the highest-scoring variable with a score of almost 70 percent. However, I have not checked whether the experiment can be reproduced based on the description of the experiment setup, so the descriptions of the experiments might not be complete.

Table 5 shows mean and median for the three factors grouped on research affiliations. The mean values indicate that the factor Experiment is documented at the same level as Data, and that Method is documented significantly better for all the surveyed studies. However, the median values of the factors differ widely with Experiment and Data on one side and Method on the other, as the median value for Method is 0.25 while it is 0.00 for the other two. Hence, the distribution is positively skewed for Experiment and Data and almost symmetric for Method. It should be noted that the median values, surprisingly, are the same for all groups. The factor Method is, on average, the one best documented. This observation is supported by both mean and median values. According to the mean values, academic research is documented better than industry, collaborations, and the combined group of collaborations and industry research. For the factor Experiment, the result when comparing academic and the combination between industry and collaborations is statistically significant.

Figure 6 shows one bar chart for each of the three factors. The y-axis of the bar charts is the frequency and the x-axis represents the mean value of the variables for each of the factors. The bar chart is not stacked so the frequency count starts at 0 for all of them. Let me explain how to interpret the bar charts by looking at the bar chart for the factor Data.

The x-axis of the bar charts ranges from 0 to 1, and this range has been divided into five equally sized partitions, that is, one partition for each variable that the factor comprises and one partition for those papers that have documented none of the variables. As part of the survey, every paper has been scored on each of the four variables that comprise Data. This means that a paper that has only documented one of the four data variables will have a mean for the factor Experiment of 0.25. Hence, it will be put into the group [0.20, 0.40), and thus increase the frequency of this group with 1. If a paper has documented all of the variables, the mean for the factor will be 1 and the paper will be put into the partition [0.8, 1.0]. The bar charts allow us to understand the distribution of the mean of the factors for all the papers that have been surveyed. As can be seen, the distributions are similar for all, academic and C + I papers. A total of 203 papers have not documented any of the variables for Experiment while 167 have not documented any of the variables of Data. Only 18 papers have not documented any of the variables of Method.

Reproducibility Metrics

Table 6 presents the mean and median scores for each of the three reproducibility metrics, R1D, R2D, and R3D. Academic research has the highest scores for all the three reproducibility metrics. Compared with C + I and collaborations, industry scores higher on R1D and R3D, but the confidence in the industry

Type	Problem Description	Goal	Research Method	Research Question	Pseudo Code	Hypothesis	Prediction	Experiment Setup
All	0.47 ± 0.05	0.22 ± 0.05	0.02 ± 0.01	0.06 ± 0.02	0.54 ± 0.05	0.05 ± 0.02	0.01 ± 0.01	0.69 ± 0.05
Academic	0.47 ± 0.06	0.22 ± 0.05	0.02 ± 0.02	0.06 ± 0.03	0.57 ± 0.06	0.06 ± 0.03	0.01 ± 0.01	0.69 ± 0.06
Collaborations	0.45 ± 0.14	0.19 ± 0.11	0.00 ± 0.00	0.04 ± 0.06	0.46 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.62 ± 0.14
Industry	0.60 ± 0.32	0.30 ± 0.30	0.00 ± 0.00	0.00 ± 0.00	0.20 ± 0.26	0.00 ± 0.00	0.00 ± 0.00	0.80 ± 0.26
C+I	0.47 ± 0.13	0.21 ± 0.11	0.00 ± 0.00	0.04 ± 0.05	0.42 ± 0.13	0.00 ± 0.00	0.00 ± 0.00	0.65 ± 0.12

Table 2. The 95-Percent Confidence Interval for the Mean of all Variables of the Factor Method for the Different Types of Papers.
$\varepsilon = 1.96\sigma_{\bar{x}}$ and $\sigma_{\bar{x}} = \hat{\sigma}/\sqrt{N}$.

Type of Paper	Training	Validation	Test	Results
All	0.56 ± 0.05	0.16 ± 0.04	0.30 ± 0.05	0.04 ± 0.02
Academic	0.61 ± 0.06	0.18 ± 0.05	0.31 ± 0.06	0.04 ± 0.02
Collaboration	0.44 ± 0.14	0.12 ± 0.09	0.28 ± 0.13	0.00 ± 0.00
Industry	0.22 ± 0.27	0.00 ± 0.00	0.20 ± 0.26	0.10 ± 0.20
C+I	0.40 ± 0.13	0.10 ± 0.08	0.26 ± 0.12	0.02 ± 0.03

Table 3. The 95-Percent Confidence Interval for the Mean of All Variables of the Factor Data for the Different Types of Papers.
$\varepsilon = 1.96\sigma_{\bar{x}}$ and $\sigma_{\bar{x}} = \hat{\sigma}/\sqrt{N}$.

scores is low here. None of these results are statistically significant. The median of R1D for C + I and industry are lower than for academic research while the median for R2D and R3D are the same for C + I and academic research.

In figure 7, the frequency of papers is plotted against reproducibility metric scores for each group of papers. The reproducibility metric scores are divided into five equally sized partitions of 0.2. The bar chart is not stacked. When it comes to the three metrics, the distribution is very similar for all, academic and C + I. For both R1D and R2D metrics, both academic and C + I have the most papers in the lowest range and then fewer and fewer for the following partitions. Only academic research is represented in the highest partitions. The R3D distribution differs with most papers in the [2, 4) range. There are no C + I papers in the range [0.6, 1.0] while there are a few academic papers in the [0.6, 0.8) range and none in the [0.8, 1.0] range.

Figure 8 shows three scatter plots. Academic papers are plotted to the left, C + I papers are plotted in the middle, and both groups are plotted in the same chart to the right. For each paper, a dot is plotted with its R1D score on the x-axis and the R2D score on the y-axis. The size of each dot is scaled according to the R3D score for that paper. Academic papers are plotted in red and the C + I papers are blue. The dots are transparent, so that the color becomes less transparent for each dot that is drawn on top of another. This plot allows us to see the distribution of individual papers and see how the three reproducibility metrics relate. For R3D, R2D, and R1D, generally, papers with a high R1D score will have a high R2D score and R3D score and papers with a high R2D score will have a high R3D score. High R3D score does not correlate with high scores on R1D and R2D, as high-scoring R3D papers are spread all over the area covered by R1D and R2D. The spread of the C + I papers is smaller than for academic papers, meaning that the variability of academic papers is higher. All the highest-scoring papers at the top-right corner are academic papers. Although both groups have the highest concentration at the lower scores, there are more dark-colored dots at higher scores for academic papers. It should be noted that 18 of the papers have a 0.0 score on the R3D metric, which means that they vanish from the plot as they have no area.

Type of Paper	Method Code	Experiment Code	Hardware Specification	Software Dependencies
All	0.08 ± 0.03	0.06 ± 0.02	0.27 ± 0.05	0.16 ± 0.04
Academic	0.09 ± 0.03	0.06 ± 0.03	0.30 ± 0.06	0.18 ± 0.05
Collaboration	0.04 ± 0.06	0.04 ± 0.06	0.13 ± 0.10	0.04 ± 0.07
Industry	0.10 ± 0.20	0.10 ± 0.20	0.20 ± 0.26	0.20 ± 0.26
C+I	0.05 ± 0.06	0.05 ± 0.06	0.14 ± 0.09	0.07 ± 0.07

Table 4. The 95-Percent Confidence Interval for the Mean of All Variables of the Factor Experiment for the Different Types of Papers.

$\varepsilon = 1.96\sigma_{\bar{x}}$ and $\sigma_{\bar{x}} = \hat{\sigma} / \sqrt{N}$.

Shown separately and together, here the axes and sizes of the dots are individual papers' scores on R1D-, R2D-, and R3D-reproducibility metrics.

Discussion

The results, although not statistically significant, paint a clear picture: The quality of the documentation shared by industry is lower than the documentation shared by academia. Given the assumption that it would be harder to reproduce research results that are poorly documented than results that are well documented, it would be easier to reproduce results from academia than from the C + I group. Out of the 16 variables that the survey covered, the academic papers have higher scores on 15 variables when compared with the C + I. The variable Problem description has the same score for academic and C + I. This means that academia scores better on 94 percent of the variables. Also, academia scores better on all three factors as well as the mean of the reproducibility metrics. The median is the same for academia and C + I on all the reproducibility metrics. To be fair, there is still much to desire when it comes to documentation quality of AI research accepted at the top conferences — whether the research is presented by academic researchers, collaborations, or industry researchers.

Does academia share more of the data than industry? The answer is yes, academia scores higher than the C + I group for all the four variables describing the Data factor. The results, however, are not statistically significant, except for the variable training data. However, the scores for data sharing are relatively high. Academia shares the training data in over 60 percent of the papers, while this is true for only 40 percent of the papers in the C + I group.

Academia shared more code than industry as well, both method code (9 percent versus 5 percent) and experiment code (6 percent versus 5 percent) Industry shares the same amount of code whether it is for setting up the experiment or for implementing the AI method; academia shares more AI method code than the code used for setting up the experiment.

One of the questions I asked in the introduction was whether one could expect industry to more easily share code than data. The premise is that data hold the most value, as data are used to generate machine learning models; however, without the data, the value of the model is low, so the foregoing premise is refuted. Interestingly, the difference between data sharing and code sharing for industry is large (40 percent versus 5 percent). How can this be so? Does this indicate that industry values the code used for running the experiments more highly than the data? Is the code used when conducting the experiments, the same code that will be used in production? This does not sound right.

Typically, experiment code is used for prototyping. Different code that has been through proper quality assurance is typically deployed, especially for large companies. Startups might not follow this practice for obvious reasons. Is there something else that lies behind? Could it be that industry is less willing to spend time on maintaining the code or answer questions related to it than academia is? Does industry have higher expectations for code quality than academia has, and does not want to share the code because of this? Or could it be that the code specifies the hyperparameters and other experiment settings, and hence renders the complete experiment transparent?

Why are industry researchers eight times more willing to share data than code? Is the data shared not that valuable for industry? Does industry share data that are relevant for proving their methods, but which have little value to competitors? Does industry use open data shared by others to prove their methods and in this way share nothing — not the code and not their own data? I have not investigated these questions in my study.

Hyperparameters could be documented both as part of the experiment code and in the experiment description where the setup is explained. Although the experiment code is not shared to a large degree (only 5 percent for C + I), the experiment setup is

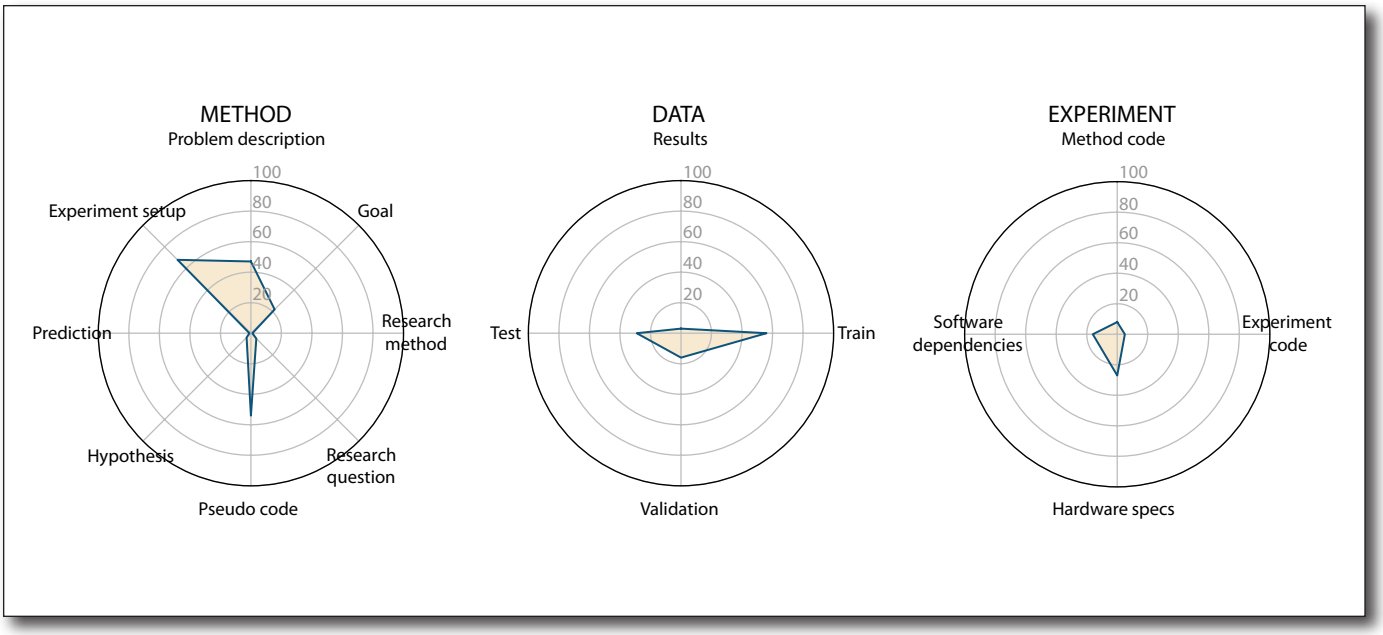

Figure 4. Spider Plots Showing the Variables of Method, Data, and Experiment for all Empirical Papers.

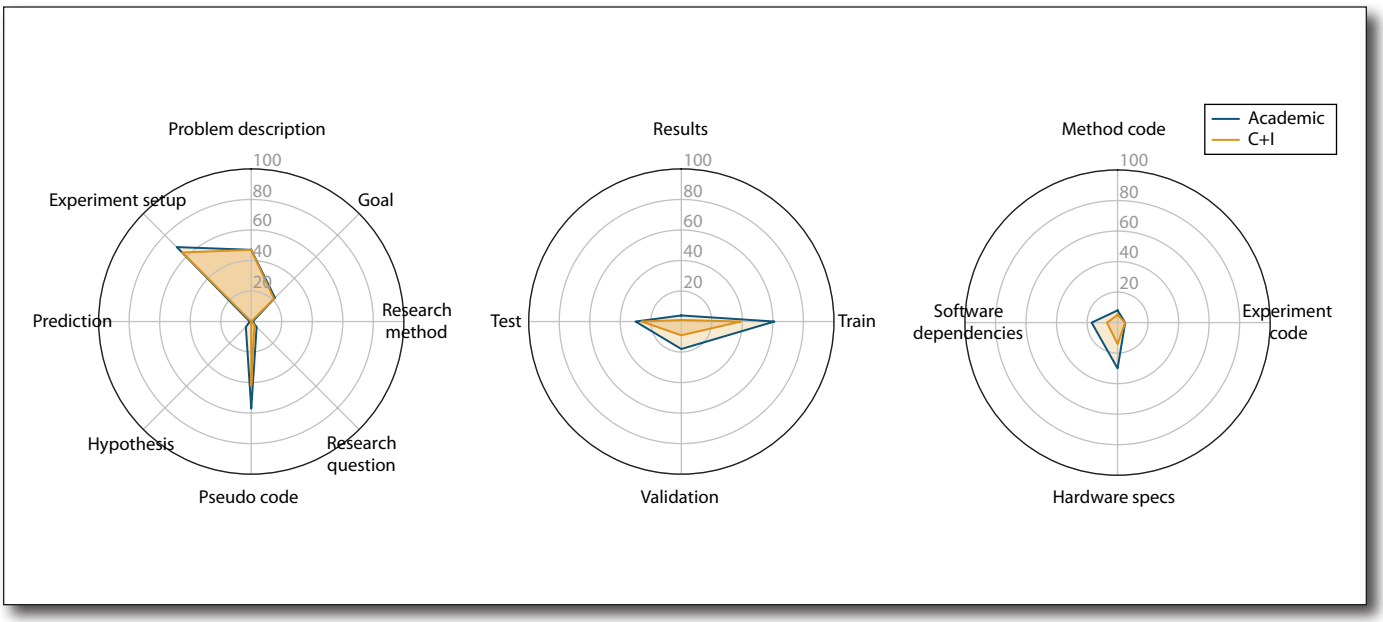

Figure 5. Spider Plots Showing the Variables of Method, Data, and Experiment for the Academic and Combined Collaboration and Industry Papers.

described for 63 percent of the papers in the C + I group. The result for experiment setup is higher for academia at 70 percent, but compared with the other variables, this is a very good result. I have not checked in detail whether all settings actually have been shared. Hence, one could imagine that some variables are described in detail, but not all — so that researchers would appear to be sharing, but are really not, as the experiment setup code is not shared.

All research presented at the top AI conferences is judged according to the same standards. There is a double-blind peer-review process where reviewers do not know who the authors are, or their affiliations. Hence, one should expect that there generally would be no differences in the documentation quality when comparing academic research and research that industry is involved in. The fact that there seems to be a pattern of research conducted by academia being

Metric	All	Academic	Collaboration	Industry	C + I
Mean experiment	0.14 ± 0.02	0.16 ± 0.03	0.06 ± 0.04	0.15 ± 0.13	0.08 ± 0.04
Mean data	0.19 ± 0.03	0.19 ± 0.03	0.17 ± 0.06	0.12 ± 0.15	0.16 ± 0.06
Mean method	0.26 ± 0.01	0.26 ± 0.02	0.22 ± 0.04	0.24 ± 0.08	0.22 ± 0.03
Median experiment	0.00	0.00	0.00	0.00	0.00
Median data	0.00	0.00	0.00	0.00	0.00
Median method	0.25	0.25	0.25	0.25	0.25

Table 5. Mean and Median Values for the Factors Experiment, Data, and Method Grouped for the Different Groups of Affiliations.

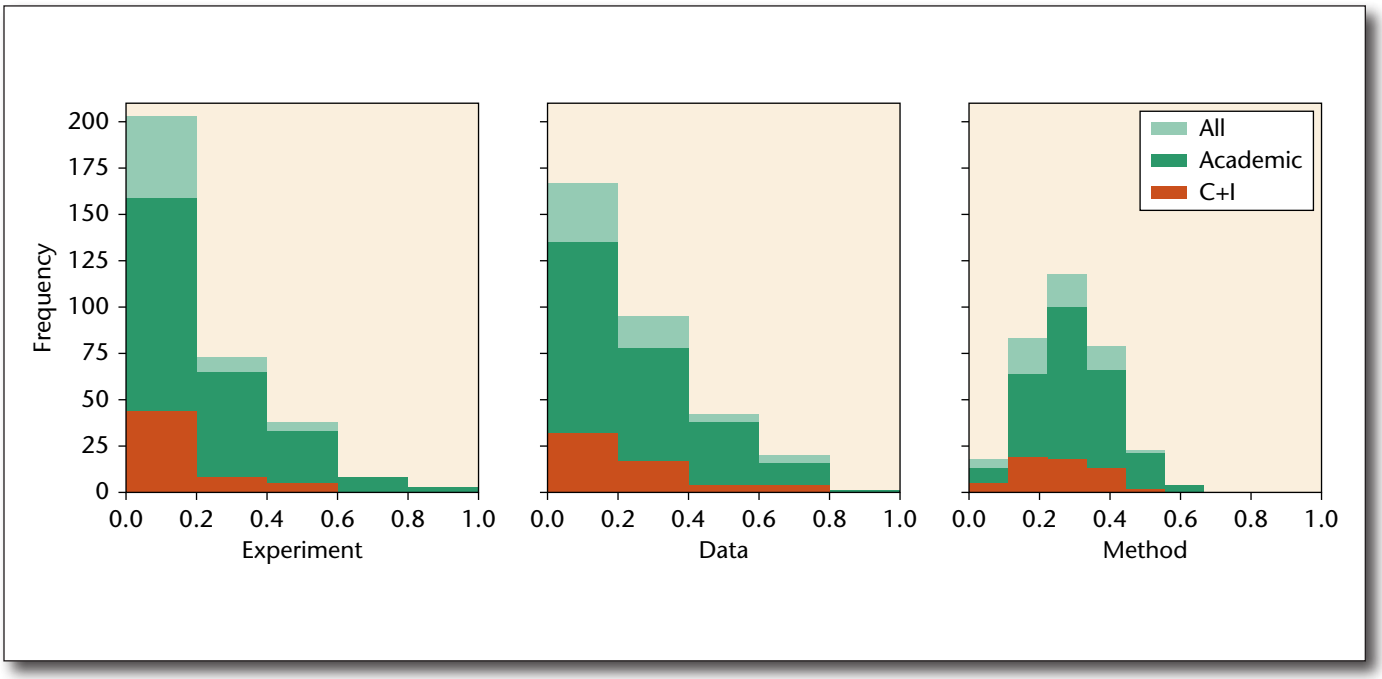

Figure 6. The Three Bar Charts Show the Frequency Distribution for All Papers Plotted Against the Mean Value for Experiment, Data, and Method.

documented better than industry research is intriguing. Why is the AI research community unable to hold industry research to the same standard as academic research in a double-blind peer-review process?

Out of the 57 surveyed papers in the C + I group, 32 involve the tech giants Microsoft, IBM, Didi, Baidu, and Facebook (see figure 9). This means that these five companies are in part responsible for 56 percent of the surveyed papers that involve industry, and that Microsoft and IBM alone stand for 49 percent. One could interpret the tech giants or the researchers that publish at the top AI conferences as the giants. No matter what, we — the AI research community — are not standing on their shoulders. Given the documentation quality of the surveyed papers, it is more like we are standing on each other's feet. The key is to improve the documentation, of course. What are the barriers that impede us?

Barriers to Reproducibility

Most results in AI and machine learning research could be made reproducible, as they are conducted on computers. Still, as follows from this study, most results seem not to be. Why is this so? I have identified some barriers for individual researchers:

Research Is Time-Consuming

Conducting research that is reproducible is time-consuming. It takes time to document research properly,

Metric	All	Academic	Collaboration	Industry	C + I
Mean R1D	0.20 ± 0.01	0.20 ± 0.02	0.15 ± 0.03	0.17 ± 0.07	0.15 ± 0.03
Mean R2D	0.22 ± 0.01	0.23 ± 0.02	0.20 ± 0.03	0.18 ± 0.09	0.19 ± 0.03
Mean R3D	0.26 ± 0.01	0.26 ± 0.02	0.22 ± 0.04	0.24 ± 0.08	0.22 ± 0.03
Median R1D	0.17	0.17	0.17	0.13	0.13
Median R2D	0.19	0.19	0.19	0.16	0.19
Median R3D	0.25	0.25	0.25	0.25	0.25

Table 6. Metrics for the 325 Papers Reporting Empirical Research Grouped by Affiliation.

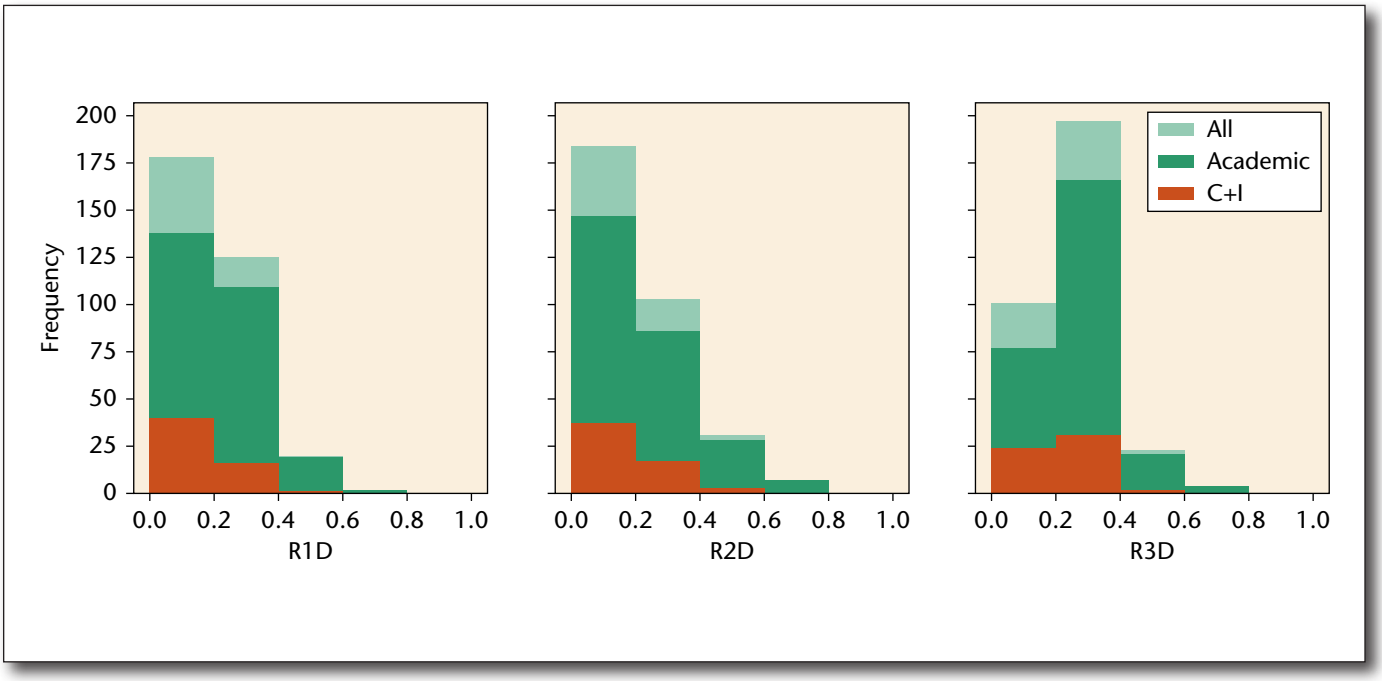

Figure 7. The Three Bar Charts Show the Frequency of the Reproducibility Metric Scores.

R1D, R2D, and R3D shown respectively for all papers, academic papers, and papers that are either collaborations or industry, C + I.

make code and data ready for sharing, and share them. If the research is successful, other researchers want to actually try to use the data and code. They might ask questions regarding the research, code, and data that take time to answer. Hence, it is not enough to share code and data. Typically, some type of maintenance (if errors are found) and support are required. The time and effort of conducting research is increased, but not only before presenting it. Time and effort is required even after the results are presented.

There Are No Incentives

Currently, there are few if any incentives for researchers to make their research reproducible. Publishers do not require that the research they publish is reproducible, and neither do grant makers. Also, whether research is reproducible is most often not a part of evaluating candidates for research positions, such as professorships. So, why bother when it requires extra effort and is time-consuming?

Future Work Might Be Put at Risk

Sharing of data, code, and detailed experiment procedures will enable independent researchers to quickly build on the published research. This might risk future research of the original researchers, and hence jeopardize possible new publications.

Given that most researchers are evaluated based on the number of research papers published in journals and presented at conferences, reducing the time

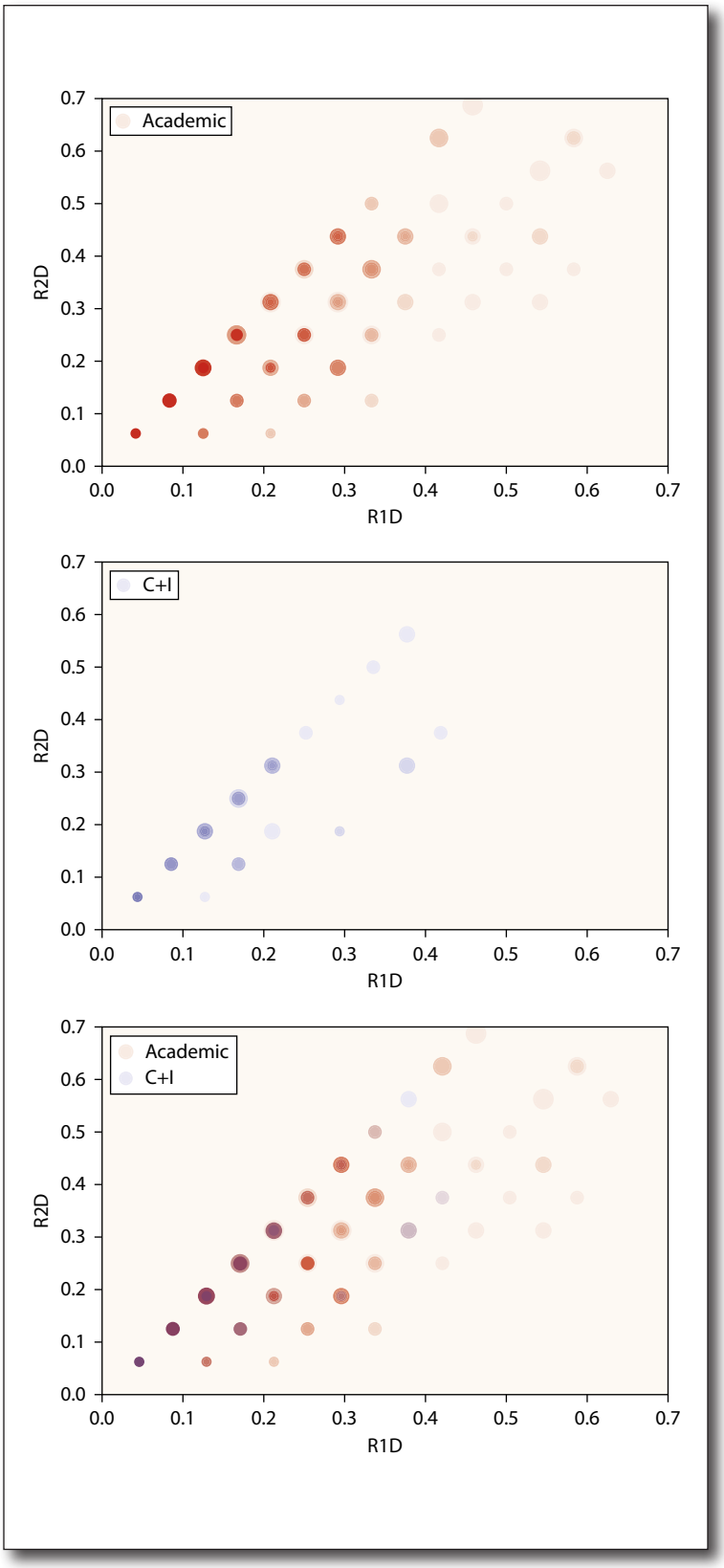

Figure 8. Individual Academic Papers and C + I Papers Are Plotted as Dots in Scatter Plots.

Academic papers are red; C + I papers are blue.

it takes to publish papers is important. Therefore, cutting corners and avoiding giving away advantages are rational actions.

How to Overcome the Barriers

What can be done to mitigate the effects of these barriers?

We Can Build Infrastructure

The time required for extra work related to making research reproducible could be reduced, although probably not completely removed, by building public infrastructure for experiment descriptions, data, results, and code. A lot of work has already been done; see, for example, Gundersen, Gil, and Aha (2018). However, more work is required.

We Can Provide Infrastructure

Publishers should provide infrastructure for data and code in addition to the infrastructure that is provided for publishing and sharing papers. Universities and other research institutions could provide infrastructure for sharing data and code maintained by their own staff. Grant makers could provide the infrastructure for the research they fund. In the era of open science where publishers fear the competition of open journals, they could provide more than they used to, and in this way meet the competition.

We Can Ensure Eligibility Requirements

Public funding sources could demand that the research that is conducted by their funding is accessible to the public. Hence, only researchers that agree to produce reproducible results by sharing code and data could be made eligible for receiving grants and funding. There are of course many issues with such a requirement, as data cannot be shared in many cases because of privacy issues and issues related to disclosing intellectual property. A possibility is to reserve parts of the available funding to applicants that agree to share everything. Another possibility is to adjust funding according to how much is shared.

We Can Share Rewards

When evaluating researchers for professorships or other research positions, the criteria could be expanded to include data sets and research software that have been published, as well as the quantity of research papers and quality of the journals in which they have been published. This is easier if the data sets and code are citable.

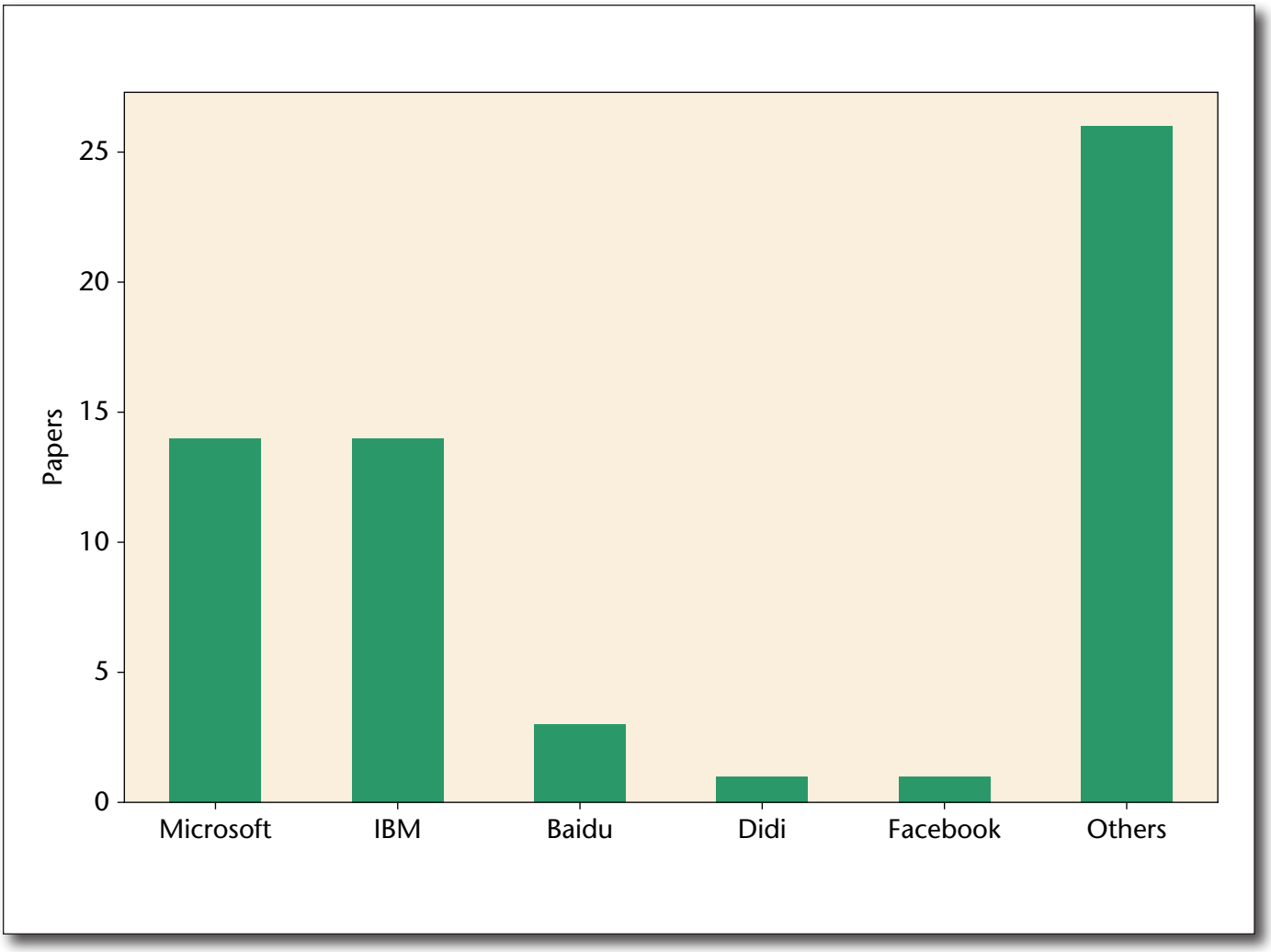

Figure 9. The Tech Giants Microsoft, IBM, Baidu, Didi, and Facebook Published 32 of 57 Papers in the Group C + I.

The total number of companies does not add to 57, as some papers have authors from more than one company.

We Can Reward Reproducibility

As reproducibility of research is a cornerstone of science, reproducibility should be rewarded in the review process and when assessing for scientific positions.

When it comes to reproducibility, academia could actually learn from industry — not necessarily from industry research practices, but from the software engineering practices that the industry follows. Software engineers focus on building quality software and continuously evaluating its performance. Software development methodologies including *Agile* (such as Scrum and Kanban), test-driven development, and code reviews have been developed to help increase the quality of the software. The reason is that the performance of the software is directly related to how well the companies themselves perform (and return financial investment), so reproducibility is a key concern together with proper performance evaluation. For companies that develop AI and machine learning software, this diligence in evaluating software extends to the AI and machine learning software. Versioning of code and data are required to ensure the capability of monitoring performance over time.

In science, reproducibility is key for ensuring that our beliefs regarding a concept, such as an AI program, are correct. It is through building and organizing the set of these beliefs that we expand our knowledge. As scientists, we should optimize for advancing knowledge. Therefore, we should ensure that our results are correct, which means that we must be able to reproduce our own results while enabling independent researchers to do the same. As discussed above, the incentives for individual scientists are not necessarily aligned for this right now, and we need an open discussion on what can be changed to get there.

For companies, maintaining a competitive advantage is important and sharing could enable competitors to close the gap. Hence, all openness can be considered a net win for the AI research community. The fact that companies share methods, code, and data should be applauded. However, given that there

is a divide in documentation quality between industry and academia, how could we reduce or remove this gap? Based on what we know about reproducibility, should we make more detailed checklists for peer-review that have check boxes for whether the problem is described well enough, a hypothesis is stated, or the code and data are shared?

If so, it will become clear what is expected from an IJCAI or AAAI paper, and that reproducibility is important for getting one accepted. Extending the acceptance criteria to include items related to reproducibility and making them explicit might help reduce the gap between industry and academia. However, if industry is required to share code or data, they might stop presenting their results at the conferences and journals that introduce such criteria. This is not a desired situation, so we should avoid it. Could we have authors register their research as R1-, R2-, or R3-reproducible research, so that it is clear what information the papers contain? This would require researchers to become aware of the documentation quality of their research — if they are not already. Also, one could imagine that a percentage of all accepted research is set for how much of the research could be R3- or R2-reproducible. Then, industry or any other researchers that would or could not share everything, could publish as much as they are able to. This would arguably make it harder to get the research accepted, so the incentives are to share.

To increase reproducibility of AI research, the culture must change. The high-impact conferences and journals have the power to make this change together with the grant makers that fund research. Although low-impact conferences and journals could see the need for reproducibility as an opportunity to get higher impact, they are afraid to scare researchers away from them.

Increased Interest in Reproducibility

In this survey, I have analyzed papers presented at IJCAI and AAAI between 2012 and 2016. However, over the last few years, the AI and machine learning communities have shown increased interest in reproducible research. A few workshops were organized before 2016, such as the Workshop on Replicability and Reusability in Natural Language Processing: From Data to Software Sharing[5] at IJCAI in 2015, which had a focus or partial focus on reproducibility. In 2017, the workshop Reproducibility in Machine Learning Research[6] was organized at that year's International Conference on Machine Learning, and the workshop Enabling Reproducibility in Machine Learning MLTrain@RML[7] was held at the 2018 International Conference on Machine Learning. The Reproducibility Challenge was organized at the 2018 International Conference on Learning Representation.[8] I organized the AAAI Workshop on Reproducibility in 2019 where the participants discussed how to improve the reproducibility of papers published by AAAI. At AAAI 2017, the tutorial *Learn to Write a Scientific Paper of the Future: Reproducible Research, Open Science, and Digital Scholarship*, was given.

This increased interest has resulted in several very interesting and relevant papers, of which a few are mentioned here. Sculley et al. (2018) discuss empirical rigor and stresses its importance for work that presents "methods that yield impressive empirical results, but are difficult to analyze theoretically" (p. 1). Mannarswamy and Roy (2018) suggest that we need to build AI software that can perform the verification task given a research paper that presents a technique and details on where to find the code and the data used in the paper. This could help mitigate the workload of reproducing research results. Exactly such a tool is presented by Sethi et al. (2018), who has made software that autogenerates code from deep learning papers with a 93-percent accuracy. Henderson et al. (2018) show that "both intrinsic (for example, random seeds, environment properties) and extrinsic sources (for example, hyperparameters, codebases) of nondeterminism can contribute to difficulties in reproducing baseline algorithms" (p. 3213).

Conclusion

We are not standing on each other's shoulders. It is more like we are standing on each other's feet. The quality of documentation of empirical AI research must clearly improve.

My findings indicate that the hypothesis that industry and academic research presented at top AI conferences is equally well documented is not supported.

Academic research score higher on the three reproducibility metrics than research to which industry has contributed. Academia also scores higher on all three factors, but these results are not statistically significant. Furthermore, academic research scoring higher than the industry research is involved in 15 out of the 16 surveyed variables while the two groups score the same on the last variable. The result is statistically significant for only three of the variables investigated. The difference in documentation quality between industry and academia is surprising, as the conferences use double-blind peer review and all research is judged according to the same standards.

I discussed three barriers for individual researchers to make research reproducible: It is time-consuming, there are no incentives, and future work is put at risk. Some suggestions for how to overcome these barriers were made: Infrastructure reducing the time and effort of making research should be built, and provided to researchers; funding sources could start demanding researchers to make the funded research conducted reproducible; and sharing of code and data should be rewarded, as should making the research reproducible. Some ideas for why there is a discrepancy between academia and industry in documentation quality were also discussed. Industry has many incentives to not share data or code, as both

can be used by competitors to reduce a company's competitive advantages.

This study suggests that industry researchers are eight times more willing to share data than code. Why this is the case is not clear. One reason could be that the data shared is already open data. Investigating this is potential future work, as well as finding out how to ensure that industry and academic research, accepted at the same conference, will have the same quality of documentation.

Acknowledgments

The authors thank Sigbjørn Kjensmo, who collected most of the data while working on his MS thesis at the Norwegian University of Science and Technology, and Rune Havnung Bakken, who read and commented on a draft of the paper. This work has been carried out at the Norwegian Open AI Lab at the Norwegian University of Science and Technology, Trondheim, Norway.

Notes

1. spectrum.ieee.org/view-from-the-valley/at-work/tech-careers/feeding-frenzy-for-ai-engineers-gets-more-i.

2. www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

3. www.wired.com/2016/08/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple.

4. github.com/kireddo/Standing on the Feet of Giants.

5. nl.ijs.si/rrnlp2015.

6. sites.google.com/view/icml-reproducibility-workshop/home.

7. mltrain.cc/events/enabling-reproducibility-in-machine-learning-mltrainrml-icml-2018.

8. www.cs.mcgill.ca/jpineau/ICLR2018-ReproducibilityChallenge.html.

9. w3id.org/rai.

References

Botvinick, M.; Barrett, D. G.; Battaglia, P.; De Freitas, N.; Kumaran, D.; Leibo, J. Z.; Lillicrap, T.; Modayil, J.; Mohamed, S.; Rabinowitz, N. C., Rezende, J., Santoro, A., Schaul, T., Summerfield, C., Wayne, G., Weber, T., Wierstra, D., Legg, S., and Hassabis, D. 2017. Building Machines That Learn and Think for Themselves: Commentary on Lake et al., Behavioral and Brain Sciences, 2017. *arXiv.org* arXiv: 1711.08378.

de Weerdt, M. M.; Gerding, E. H.; Stein, S.; Robu, V.; and Jennings, N. R. 2013. Intention-Aware Routing to Minimise Delays at Electric Vehicle Charging Stations. In *Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, 57. New York: Association for Computing Machinery. doi.org/10.1145/2516911.2516923.

Drummond, C. 2009. Replicability Is Not Reproducibility: Nor Is It Good Science. Paper presented at the Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning (ICML). www.researchgate.net/publication/228709155_Replicability_Is_Not_Reproducibility_Nor_Is_It_Good_Science.

Goodman, S. N.; Fanelli, D.; and Ioannidis, J. P. A. 2016. What Does Research Reproducibility Mean? *Science Translational Medicine* 8(341): 341ps12. doi.org/10.1126/scitranslmed.aaf5027.

Gundersen, O. E.; Gil, Y.; and Aha, D. 2018. On Reproducible AI — Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine* 39(3): 56–68. doi.org/10.1609/aimag.v39i3.2816.

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*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning That Matters. *arXiv.org* arXiv:1709.06560.

Lample, G., Ott, M., Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv.org* arXiv:1804.07755.

Mannarswamy, S., and Roy, S. 2018. Evolving AI From Research to Real Life — Some Challenges and Suggestions. Paper presented at the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI), Evolution of the Contours of AI. 5172–79. doi.org/10.24963/ijcai.2018/717.

Ott, M.; Auli, M.; Granger, D.; and Ranzato, M. 2018. Analyzing Uncertainty in Neural Machine Translation. Paper presented at the 35th International Conference on Machine Learning (ICML) 2018. *arXiv.org* arXiv:1803.00047.

Peng, R. D. 2011. Reproducible Research in Computational Science. *Science* 334(6060): 1226–7.

Salisbury, E.; Kamar, E.; and Morris, M. R. 2018. Evaluating and Complementing Vision-to-Language Technology for People Who Are Blind With Conversational Crowdsourcing. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI), 5349–53. doi.org/10.24963/ijcai.2018/751.

Sculley, D.; Snoek, J.; Wiltschko, A.; and Rahimi, A. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. Paper presented at the Sixth International Conference on Learning Representations (ICLR) 2018 Workshop, Vancouver, BC Canada, April 30–May 3.

Sethi, A.; Sankaran, A.; Panwar, N.; Khare, S.; and Mani, S. 2018. Dlpaper2code: Auto-Generation of Code From Deep Learning Research Papers. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 7339–46. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A., Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the Game of Go Without Human Knowledge. *Nature* 550(7676): 354–9. doi.org/10.1038/nature24270.

Stodden, V. C. 2011. Trust Your Science? Open Your Data and Code. *Amstat News* 2011(July): 21–2. web.stanford.edu/~vcs/papers/TrustYourScience-STODDEN.pdf.

Odd Erik Gundersen (PhD, Norwegian University of Science and Technology) is the Chief AI Officer at the renewable energy company TrønderEnergi AS and an Adjunct Associate Professor at the Department of Computer Science at the Norwegian University of Science and Technology. Gundersen has applied AI in the industry, mostly for startups, since 2006. Currently, he is investigating how AI can be applied in the renewable energy sector and for driver training, and how AI can be made reproducible.