The Diversity Crisis of Software Engineering for AI

Bram Adams
MCIS, Polytechnique Montreal
http://mcis.polymtl.ca/
bram.adams@polymtl.ca

Foutse Khomh
SWAT, Polytechnique Montreal
http://swat.polymtl.ca/
foutse.khomh@polymtl.ca

1. Intro

AI is experiencing a "diversity crisis" [1]. Several reports [1-3] have shown how the breakthrough of modern AI has not yet been able to improve on existing diversity challenges regarding (amongst others) gender, race and geography, neither for (1) the end users of those products, nor for (2) the companies and organizations building AI products. Plenty of examples have surfaced in which biased data engineering practices or existing data sets led to incorrect, painful or sometimes even harmful consequences for unassuming end users [4]. The problem is that ruling out such biases is not straightforward due to the sheer number of different bias types [5]. In order to have a chance to eliminate as many biases as possible, most of the experts agree that the teams and organizations building AI products should be made more diverse [1-3]. In essence, this harkens back to Linus' Law [6] for open source development ("given enough eyeballs, all bugs are shallow"), but applied to the development process of AI products.

Unfortunately, current AI organizations and companies are not diverse. For example, AI NOW Institute's West et al. [1] report on how work floor discrimination during hiring/promotion persists, or on how AI companies still reason in terms of binary gender. Worse, current diversity initiatives tend to introduce new biases, for example towards "white women". Element AI's Gagné et al. [7] report how AI experts (self-identified via their LinkedIn profiles) primarily reside in North America, the UK, France and Germany. Even in academia, which could be considered to be more progressive, they found less than 1 out of 5 authors at major AI conferences to be women (which coincidentally corresponds to the percentage of AI faculty estimated to be women [8]), which was confirmed by Stathoulopoulos et al. for arXiv papers [9] and by Simonite for publications (and the percentage of women employees) of R&D labs at major tech companies like Google, Microsoft and IBM [2].

In this column, which is based on our MSR 2018 publication [10], we argue that the "diversity crisis" of AI goes even further, and that it has other consequences that might hamper further evolution. This claim is based on an empirical study of the GitHub data of the 20 most popular company- and community-driven frameworks for machine learning (ML), and the LinkedIn and Google Scholar profiles of the top contributors of these frameworks [10]. Basically:

1. the top ML software frameworks, while being open source, are primarily driven by companies and these frameworks are optimized for specific cloud offerings;
2. ML software projects converge towards highly specialized skills and roles.

Both observations show additional sources of bias in modern AI software projects that, if ignored, might quickly widen the gap between the companies at the top of the tech chain, and others. We also discuss a number of promising directions to proceed.

2. Methodology

To understand diversity in ML/AI software projects, we analyzed 598 core contributors of 20 top open source ML frameworks, out of a set of 104 ML-related open source frameworks.

The frameworks were selected by querying GitHub's search API using technical terms related to ML, such as "machine learning", "deep learning", "statistical learning", "neural network", "supervised learning", "unsupervised learning" and "reinforcement learning", and known keywords such as "toolkit", "tool", "framework" and "library". The resulting search results were filtered manually based on the license file, README, list of contributors (and affiliations) and GitHub's built-in wiki, eventually arriving at a list of 104 projects.

Next, to approximate the number of developers who adopted these frameworks, we used GitHub's built-in popularity measure, i.e., the number of people who star-ed the framework's repository[1]. By selecting the set of frameworks whose cumulative number of stars (adopters) represents 80% of the total number of stars of the 104 ML frameworks, we obtained 29 popular frameworks. We then qualitatively analyzed these projects' descriptions on GitHub, their corresponding organizations, and their official Websites to determine whether they are mostly supported by a company or by the open source community. From this qualitative analysis, we selected the top-10 company-driven and the top-10 community-driven projects. These 20 projects represent 70% of the total number of stars attributed to the 104 ML frameworks on GitHub.

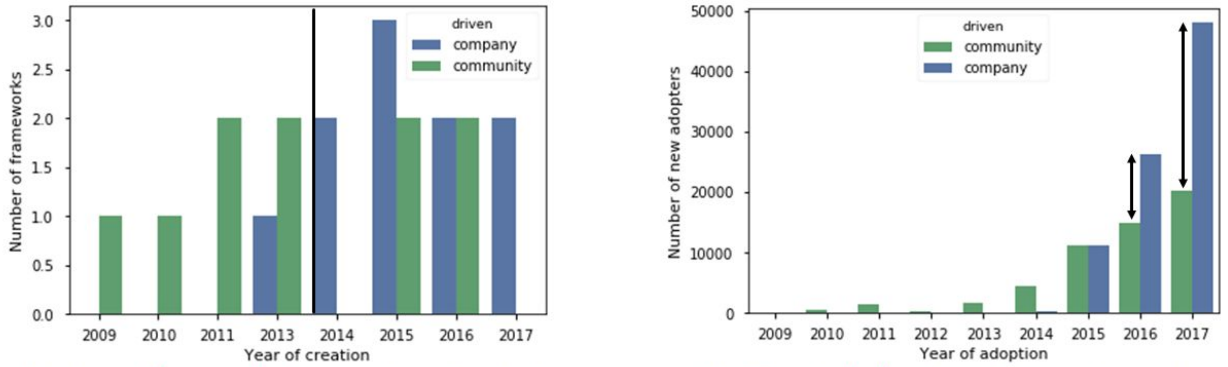
To identify the core contributors of these frameworks, we computed the total contribution of each contributor and retain the set of contributors whose total number of contributions in the source code repository accounted for 90% or more of the total contributions. We then extracted GitHub, LinkedIn, and Google Scholar profiles of each contributor in this set, whenever available.

[1] Our study also considered a second measure of adoption, i.e., 4,099 projects that used the ML frameworks. We refer to our paper for those results [10].

3. Results

a. Company- vs. community-driven

Our study shows that, although the development of open source ML frameworks was initially driven by the open source community, since 2013 the number of ML frameworks backed by companies has surpassed the number of community-driven ML frameworks. Furthermore, the number of adopters of ML frameworks supported by companies largely surpasses the number of adopters of community-driven ML frameworks, as shown on Figure 1.

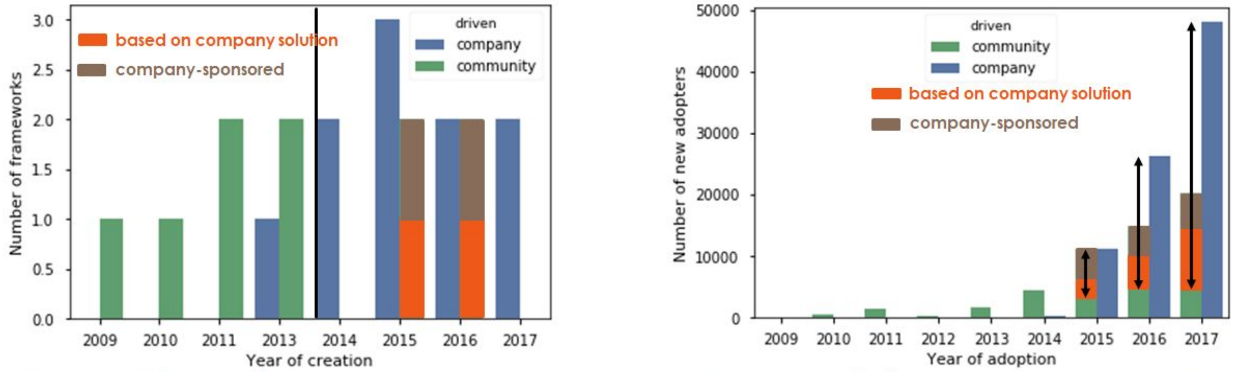


Number of frameworks for each type of project per year of creation

Number of adopters for each type of project per year of adoption

Figure 1: Company-driven vs. community-driven ML frameworks.

In reality, the dominance of companies is much larger than suggested in Figure 1 because, apart from releasing their own ML frameworks, companies also actively contribute to community-driven ML frameworks by providing and-or supporting skilled professionals. Moreover, a deep analysis of the source code and dependencies of community-driven frameworks revealed that all community-driven frameworks that appeared from 2014 on are in fact built on company-driven solutions, as shown on Figure 2. This increases even more the discrepancy between the number of adopters of “purely” community-driven frameworks and frameworks with companies involved. In other words, the last 6 years companies have been dominating the development of ML technologies.



Number of frameworks for each type of project per year of creation

Number of adopters for each type of project per year of adoption

Figure 2: Community-driven ML frameworks based on company-driven solutions.

We also observed that company-driven ML frameworks are often deployed and optimized for commercial cloud offerings, which raises the risks that these open source frameworks serve as baits to trap ML users and community developers into commercial platforms. It is utterly important to ensure that open-source ML technologies remain accessible from a diverse type of infrastructures, to prevent them from being biased towards the goals of some specific organizations. Currently, the richest datasets reside within the servers of large companies, hence more effort should be directed towards the development of open datasets, to empower a more diverse user base.

b. Skill specialization

Our analysis of the profile of contributors of the 20 top ML-frameworks learnt that, except for 3 frameworks, all ML development teams are hybrid. In company-driven ML projects, professional researchers and engineers often contribute equally. The contribution of academic researchers is often limited to model design and they rarely contribute to the development or production of code. In contrast, in community-driven ML projects, the picture is significantly different, with professional and academic researchers writing almost the totality of the code.

Taking a closer look at the profile of professional researchers from either company-driven or community-driven frameworks, we observed that they typically hold a Ph.D. degree and have experience working in either a R&D lab or an innovation product team, which signals that ML development is currently mostly driven by few highly skilled professionals. Such a high entry bar is likely to have an adverse effect on innovation; it is important to develop tools and policies that enable the adoption of ML by a wider and diverse user base, allowing them to express their creativity and achieve their use cases.

4. Discussion & Way Forward

The two diversity challenges that we studied in open-source ML frameworks, i.e., the dominance of companies and reliance on highly specialized personnel, are not necessarily new [3] [7] [11] [12], yet the scale at which these ML software frameworks and projects are developing, as well as the (often safety-critical) domains in which they are applied, are cause for concern. Most intuitively, these challenges could lead to a shortage of qualified personnel, on top of the current shortage of open source contributors [13]. At a higher level, the need to add specialized roles like industry researchers increases the need for communication, while at the same time the dominance of companies makes the development process less transparent. A clear example of the latter is the "tensorflow-gardener" bot that re-commits TensorFlow changes made internally by Google employees on the public GitHub repository, typically to hide employee names or to merge multiple commits at once. For technologies that increasingly are controlling humans' lives, transparency of the development process is essential. Finally, going back to Linus' Law [6], a too biased, less diverse contributor community also risks to correlate with lower code quality, or at least higher costs to achieve a certain code quality.

So, what can be done about this? First of all, for the company dominance challenge, it is important to stress that we are not finger pointing at the companies, since they have pushed many of the recent breakthroughs in ML/AI and also enabled democratization of those breakthroughs by open-sourcing advanced frameworks like TensorFlow or PyTorch. In fact, they as well are fighting for the (relatively) scarce AI talent [7], and would welcome any chance at attracting both competent researchers and engineers. Instead, the core issues seem to be (1) lack of AI training of potential contributors and (2) the proliferation of ML frameworks backed by different companies, each with their own terminology and agenda. While issue (1) could be addressed (or at least to some extent [1]) by including AI courses in undergraduate curricula and organizing workshops and seminars on advanced AI technologies, which most universities and AI research centres have been doing the past couple of years, issue (2) would require a standardization effort [14], led by a neutral consortium.

Both of these issues require long-term thinking. In the meantime, the company and/or high-tech bias in open source ML projects might be controlled by embracing more open development processes (e.g., MLOps [15]) that stress communication between all stakeholders, including data scientists and industry researchers. Openness will likely spark more collaboration between different AI stakeholders and increase trust in AI technologies.

In the end, we believe that improving diversity, whether in terms of gender, race, geography, equity or expertise, of (AI) open-source software projects is essential to improve not only the quality [6] of these projects, but also the cohesion of their contributor communities [16] and (ultimately) their sustainability.

5. Bibliography

[1] Sarah Myers West, Meredith Whittaker and Kate Crawford (2019). "Discriminating Systems: Gender, Race and Power in AI", AI Now Institute (https://ainowinstitute.org/discriminatingsystems.html).

[2] Tom Simonite (2018). "AI Is the Future - But Where Are the Women?", WIRED (https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/).

[3] Carlos M. Melendez (2019). "In AI, Diversity Is A Business Imperative", Forbes Technology Council Post (https://www.forbes.com/sites/forbestechcouncil/2019/11/14/in-ai-diversity-is-a-business-imperative/#51c5aac8103f).

[4] Anna Lauren Hoffmann (2018). "Data Violence and How Bad Engineering Choices can Damage Society", in Medium (https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4).

[5] https://www.research.ibm.com/5-in-5/ai-and-bias/

[6] Eric S. Raymond and Bob Young (2001). "The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary", O'Reilly & Associates, Inc.

[7] Jean-François Gagné, Grace Kiser and Yoan Mantha (2019). "Global AI Talent Report 2019", Element AI (https://jfgagne.ai/talent-2019/).

[8] Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz and Zoe Bauer (2018). "The AI Index 2018 Annual Report", AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA (http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf).

[9] Kostas Stathoulopoulos and Juan Mateos-Garcia (2019). "Gender Diversity in AI Research", Nesta (https://www.nesta.org.uk/report/gender-diversity-ai/).

[10] Houssem Ben Braiek, Foutse Khomh and Bram Adams (2018). "The Open-Closed Principle of Modern Machine Learning Frameworks", in Proceedings of the 15th International Conference on Mining Software Repositories (MSR), pp. 353–363, Gothenburg, Sweden.

[11] Eric Horvitz (2017). "AI, people, and society", in Science, Vol. 357, Issue 6346, pp. 7.

[12] Kate Crawford (2016). "Artificial Intelligence’s White Guy Problem", in New York Times, 25/06/2026 (https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html).

[13] The Linux Foundation (2018). "The 2018 Open Source Jobs Report" (https://www.linuxfoundation.org/publications/2018/06/open-source-jobs-report-2018/).

[14] Carlos E. Perez (2018). "Why Deep Learning Needs Standards for Industrialization" (https://medium.com/intuitionmachine/challenges-for-ai-standardization-eab1de4fab0b).

[15] Nisha Talagala (2018). "Why MLOps (and not just ML) is your Business’ New Competitive Frontier" (https://www.aitrends.com/machine-learning/mlops-not-just-ml-business-new-competitive-frontier).

[16] Gemma Catolino, Fabio Palomba, Damian A. Tamburri, Alexander Serebrenik and Filomena Ferrucci (2019). "Gender diversity and women in software teams: how do they affect community smells?", In Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), pp. 11-20, Montreal, Canada.