The Business of Open-Sourcing Artificial Intelligence: Commercial Interests and Collaboration Dynamics in Machine Learning Developer Communities

Document submitted for the degree of DPhil in Social Data Science

at the Oxford Internet Institute, University of Oxford.

Word Count: 63,985

Cailean Osborne

Wolfson College, University of Oxford

Advisors: Prof. Mark Graham and Dr. Xiaowen Dong

Hilary Term

31 March 2025

Abstract

Open source is widely celebrated as a means to democratise the development and governance of artificial intelligence (AI) technologies. In the last decade alone, open source AI (OSAI) has evolved from a scientific endeavour to a cornerstone of the AI industry, with over a million open source software (OSS) and open model (OM) repositories on GitHub and Hugging Face Hub respectively. The growth of the OSAI ecosystem is partly credited to commercial activity, from the open-sourcing of AI software and models to collaborations thereon. The pervasiveness of commercial activity raises the question of why and how companies participate in OSAI development, and the implications thereof on development practices, governance norms, and the trajectory of the OSAI ecosystem. This thesis addresses this question through four research papers (RP), which make both theoretical and empirical contributions to the nascent research agenda on the political economy of OSAI. RP1 examines the patterns and practices of "open source co-opetition"—that is, open source collaboration among companies, including market competitors—in three major OSAI projects: PyTorch, TensorFlow, and Transformers. RP2 investigates and taxonomises commercial incentives for "democratising AI" through the lens of 43 AI software donations to the Linux Foundation. RP3 presents a large-scale quantitative analysis of development activity on Hugging Face Hub, shedding light on the uneven patterns of OM development and adoption. RP4 examines the roles of private and public funding in sustaining non-commercial, community-led OSAI projects through a case study on scikit-learn. The thesis concludes with a discussion of its key contributions and future research directions.

Declarations

This thesis was funded by the Economic and Social Research Council's Grand Union Doctoral Training Partnership Digital Social Science Pathway (Grant Number: ES/P000649/1).

Ethical clearance was obtained from University of Oxford's Research Ethics Committee prior to beginning data collection for this thesis (CUREC 1A, OII C1A 22 089).

Contents

List of Terms

Acronyms

AI Artificial intelligence

API Application programming interface

AWS Amazon Web Services

CLA Contributor license agreement

CSCW Computer-Supported Cooperative Work

CV Computer vision

DL Deep learning

EU European Union

FLOSS Free, Libre, Open Source Software

FM Foundation model

GPUs Graphics processing units

HF Hugging Face

Inria French Institute for Research in Computer Science and Automation

LF Linux Foundation

LLM Large language model

LOC Lines of code

ML Machine learning

MRQ Meta research question

NLP Natural language processing

OECD Organisation for Economic Co-operation and Development

OM Open model

ONNX Open neural network exchange

OSAI Open source artificial intelligence

OSI Open Source Initiative

OSS Open source software

PACMHCI CSCW Proceedings of the ACM on Human-Computer Interaction track on Computer-Supported Cooperative Work

PR Pull request

QDA Qualitative data analysis

RP Research paper

RQ Research question

R&D Research and development

SaaS Software as a Service

SDS Social data science

SEA-LION South East Asian Languages in One Network

SNA Social network analysis

USA United States of America

Definitions

Algorithm: An algorithm is a finite sequence of pre-defined instructions for solving a specific problem or performing a computation (Cormen, Leiserson, Rivest, & Stein, 2022).

Artificial intelligence (AI): AI concerns the development of "machine-based systems that, for explicit or implicit objectives, infer, from the input [they] receive, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (OECD, 2024). The field of AI encompasses sub-fields, such as machine learning, deep learning, natural language processing, computer vision, and robotics, among others (Russell, 2021).

Deep learning (DL): DL is a sub-field of machine learning that uses artificial neural networks with multiple layers to progressively learn representations of data (LeCun, Bengio, & Hinton, 2015).

Digital commons: The digital commons are digital resources or corresponding resource systems that are developed and governed by an online community (Dulong de Rosnay & Stalder, 2020).

Foundation model (FM): FMs are large-scale AI models with up to trillions of parameters, which are trained on diverse datasets using neural network architectures (e.g. transformers) and are characterised by their adaptability to perform a variety of downstream tasks with minimal task-specific training, such as the generation of text, code, images, and audio (Bommasani et al., 2022).

Large language model (LLM): LLMs are a type of FM that contain between billions and trillions of parameters, which are trained on massive text corpora and are capable of processing and generating natural language (Mökander, Schuett, Kirk, & Floridi, 2023).

Machine learning (ML): ML is a sub-field of AI that concerns statistical learning, where "a computer program is said to learn from experience (`E`) with respect to some class of tasks (`T`) and performance measure (`P`), if its performance at tasks in `T`, as measured by `P`, improves with experience `E`" (Mitchell, 1997). ML encompasses various approaches, including supervised, unsupervised, and reinforcement learning, as well as various tasks, including classification, regression, and clustering, among others.

Open models (OMs): OMs are AI models whose architecture and parameters (i.e. pre-trained weights and biases) are publicly available under open licenses that permit their use, study, modification, and redistribution (White et al., 2024).

Open source artificial intelligence (OSAI): OSAI is used as an umbrella term to encompass various digital resources that are used in AI research and development, including OSS, OMs, and open datasets, which are made available under licenses that allow anyone to freely study, use, copy, modify, and redistribute them. Alternatively, "open AI" is used by Widder, Whittaker, and West (2024).

Open source software (OSS): OSS is software whose source code is made available under a license that allows anyone to freely study, use, copy, modify, and redistribute it (OSI, 2007). Note that, per the Open Source Initiative, "open source" is not hyphenated as a lexicalised compound noun, while "to open-source" is hyphenated as a compound verb, which means to release or make software freely available for anyone to use, study, modify, and distribute (OSI, 2024b).

Political economy: Political economy has a myriad of definitions across disciplines and traditions. This thesis adopts the definition in Mosco's (2009, p.24) *Political Economy of Communication*, who defines political economy as the study of "the social relations, particularly power relations, that mutually constitute the production, distribution, and consumption of media resources." Hence, the political economy of OSAI is defined as the study of the social relations, particularly power relations, that mutually constitute the production, distribution, and consumption of OSAI technologies.

Software ecosystems: In software engineering research, software ecosystems are understood as "a set of businesses functioning as a unit and interacting with a shared market for software and services, together with the relationships among them" (Jansen, Brinkkemper, & Finkelstein, 2009, p.9).

Supply chain capitalism of AI: Supply chain capitalism is a theoretical framework for understanding the inter-continental scale and constitutive diversity of global capitalism (Tsing, 2009). Drawing on this framework, Valdivia (2024, pp.5-6) defines "the supply chain capitalism of AI" as "the orchestration of commodity chains that extract, ship, and manufacture the natural resources needed to develop AI from an infrastructural perspective, such as mines, data centres, and e-waste dumps together with their human resources (miners, drivers, directors of data centres operations, e-waste dismantlers, etc.). Within this chain, digital elements such as datasets and algorithms together with human labour (data annotators, data scientists, data engineers, etc.) are also key elements."

List of Figures

List of Tables

Acknowledgements

As I wrap up this thesis, I cannot help but think back to my first doctoral seminar in the first week of October 2021. The convener explained to us—that is, myself and my fellow doctoral peers who had, for better or worse, also decided to embark on a PhD—that writing a thesis is like dropping a grain of sand on a beach. Behind her, a projected slide glistened on the wall, showing a beach. This beach represented our respective research fields. We should not be disheartened by the seeming insignificance of our contribution to this beach, she told us. On the contrary, we should take pride in making a contribution, no matter how humble, to a collective effort by a community of scholars which we would be joining. One grain of sand at a time, one step at a time, we advance our knowledge as a community, she explained. At that time, I had a faint idea of what my grain of sand would be—I had decided on my topic and I had my fair share of research ideas; some good, some less good—but the idea of actually finishing the thesis and dropping my grain of sand seemed so distant.

Well, here I am, writing my acknowledgements in the final moments before submitting my thesis. Running with the metaphor, this work represents my contribution to the nascent but rapidly evolving research agenda on the critical political economy of open source artificial intelligence (OSAI). This topic has gone from a rather niche topic when I began working on this thesis in October 2021, let alone when I applied for the doctorate in November 2020, to a mainstream talking point among AI researchers and developers. It is remarkable to think that in the summer of 2021, prior to beginning my PhD, when I worked in AI policy at the UK government, the topic of open source did not even make it into the UK's national AI strategy. Nowadays it is hard to keep up with the never-ending flow of announcements of breakthroughs and innovations that push the frontiers of AI, including the many open AI models that seem to be more and more powerful by the day.

The rapid pace of change in fields like AI presents concrete challenges for (doctoral) researchers like myself, who, with limited resources, task themselves with investigating the social dynamics of such developments. This thesis represents my best effort to do just this. The four research papers included in this thesis tackle specific research problems and gaps concerning commercial participation in OSAI development, shedding light on commercial incentives for different types of involvement, from open-sourcing AI software or models to when companies collaborate with market rivals to

collectively develop software, and how community-led projects sustain themselves in light of the prevalence and intensity of commercial activity. But much work remains to be done to advance this research agenda and, beyond the research world, to ensure that OSAI serves broader societal interests beyond solely commercial agendas. My research was made possible by standing on the shoulders of the many scholars who wrote on the critical political economy of the digital commons and AI before me, and I hope that my contributions can provide a useful foundation or even as a source of inspiration for the minds and souls who come after me.

The truth is that I could not have done all of this work without the love and support from so many people over the years, to whom I owe many more thanks than I could possibly express in words.

First, for enabling me to do this PhD in the first place, I would like to thank the professors at the Oxford Internet Institute who decided to give me the chance to pursue this PhD. Thank you Dr. Balazs Vedres and Dr. Joss Wright, who interviewed me, and those behind the scenes who agreed to take a chance on me. Then, thank you to the incredible Economic and Social Research Council Grand Union Doctoral Training Programme, in particular Alan McKechnie, for awarding me a scholarship that not only allowed me to pursue my research with financial stability but which also sent me around the world: to Paris for my fieldwork with the scikit-learn team; to the Linux Foundation for a summer internship that evolved into an eye- and door-opening part-time job; to Taipei for a two-month summer school in Mandarin; to Beijing for a six month research visit at Peking University; to New York City for a conference on open source at the United Nations; and to San José, Costa Rica to present my research at my first ACM conference. It truly was the gift that kept on giving. I pinch myself because I cannot believe how lucky I have been, how many places I have been able to visit, and how many people I have been able to meet thanks to this scholarship. Thank you!

I would also like to thank my ever so thoughtful and available supervisors, Prof. Mark Graham and Dr. Xiaowen Dong. Thank you Mark for pushing me to think big, for inspiring me by the real-world impact and public accessibility of your work, and for not letting me forget that there's no such thing as a free lunch—in open source and elsewhere. Thank you Xiaowen for your patience and kindness, for teaching me how to break down problems, and for our laughs in 中文. Thank you also to the wider Oxford Internet Institute community, in particular my dear friends (more about you below), professors, and the administrative team that got me to the finishing line. In particular, I would like to thank Prof. Vili Lehdonvirta, Dr. Fabian Stephany, and Dr. Bernie Hogan for taking the time to assess me at the Transfer of Status and Confirmation of Status milestones. Your feedback and advice was extremely valuable and helped me improve my research. I also thank Prof. Renaud Lambiotte for his feedback and advice on network science in the early stages. Of course, a massive thank you goes out to the administrative team, in particular Chrissy Bunyan, Laura Maynard, and

David Pepper, for ensuring such a seamless PhD journey. Thank you, dear OII staff and community.

Thank you to Prof. Minghui Zhou for hosting me for 6 months at her Open Source Software Data Analytics Lab at Peking University. Those six months were life-changing. While the Chinese work culture is not for me, I left feeling inspired by the brilliance, work ethic, and the humility of my colleagues. I am also pleased that I chose the hard but rewarding immersion route by living in a Beijing hutong. Being the only foreigner in a downtown neighbourhood, where only one other person spoke English, challenged me every day. Some days were hard indeed, but I am so grateful for the sweetness and generosity of my neighbours, in particular 李涵 who took me under his wing and opened my eyes to how a society, more foreign to me than expected, lived and loved. It opened my eyes to myself, too, challenging things I took for granted or valued in life. 谢谢北京!

I also want to thank everyone who has participated in my research in some capacity: those who participated in research interviews, those who helped me decipher error messages, those who provided feedback on drafts of my research design and research papers, and the anonymous reviewers—even my demanding reviewer 2s—who pushed me to get my work to the next level. Thanks for all your help in getting my research ideas from my head to paper and eventually to publication.

Thank you also to all my colleagues and collaborators at the Linux Foundation for their kindness, support, and education in the world of open source. This includes but is certainly not limited to Hilary Carter, Mirko Böhm, Ana Jiménez Santamaría, Adrienn Lawson, Ibrahim Haddad, Gabriele Columbro, Matt White, Paul Sharratt,, Jan Krewer, Nick Gates, Anni Lai, Arnaud LeHors, and others.

Last but not least, I want to thank my family and friends for all their love, support, company, and inspiration. I should start by thanking my loving parents, Solveigh and Iain, for bringing me into this world and encouraging me, in their own ways, to always strive and struggle for better, to be kind to others, and to share love with the world. Life has not been easy or kind for our family but they showed my brothers, Max and Ossian, and I to keep up the fight and to never stop smiling and laughing along the way. Thank you Omi for your stubborness to stay alive and the jokes we have enjoyed together. Then, Luisa, mi amor, thank you for your laughs, your smiles, your curiosity, and passion for life and justice. Every day, you make me want to be a better person (and a better dancer at Cumbia, Bachata, Salsa...the list goes on jaja). Te amo hasta la luna y vuelta, y otra vez.

And thank my friends and loved ones who have made me the person I am today. First of all, Maud Barret Bertolini, thank you for our many years together, for your love, for your inspirations and insights, for growing up with me. I think of our years fondly. In Oxford, thank you Marta Ziosi, Hannah Rose Kirk, Calvin Yixiang Cheng, Damian Maher, Julia Slupska, Liam Bekirsky, Jo and Paul Rivera-Carlisle, Giuliano Formisano, Andreas Tsamados, Prathm Juneja, Manu Tonneau, Felix Simon, Jakob Mökander, Yung Au, Sruj Katta, Alec Schellinx, Noo and Harj Narulla, Hanna Wetzel and Foo

Coovadia, Cat Fan, Patrick Gildersleve and Nayana Prakash; as well as from my previous time at Oxford: Matt Pierri, Jack LaViolette, Publio Adrianza, Ollie Ballinger, Andrew Strait, Bethan Charnley, Scarlet Dawson Duckworth, Phoebe Bright, David Watson, Siân Brooke, Carl Öhman, among many other champs. I also want to thank friends from along these paths of my life. In Taipei, thanks to Erdem for being the best flatmate and to Saul for being such a serious classmate and friend later on in China. In Beijing, thanks to 李涵 , Sophie, Alex, Kenza, Emma, Rox, Lexi, Raz, Ben, Marlon, Fabs, Runzhi, Hengzhi, Yuxia, and many others for an incredible 6 months in China. Thanks to my friends from previous moments in my life, too: Gregorio, Hans, Tom, Bruno, Kat, Rob, Billy, Ollie, Greg, Will, Alasdair, Susanna, Alice, Sandra, Josh, Adrian, and many others for their love and friendship.

The world often feels like a dark place these days. But with empathy, solidarity, and courage (and maybe a wee bit of open source!), I hope that we can nurture a more peaceful, green, and open future.

Cailean Osborne

Bogotá, Colombia

30 December 2024

1. Introduction

1.1 Background and Meta Research Question

In October 2007, Sonnenburg et al., a coalition of sixteen eminent researchers in the field of artificial intelligence[1] (AI), published an urgent call to action in "The Need for Open Source Software[2] (OSS) in Machine Learning[3] (ML)." They painted a desperate picture of their field: despite major theoretical advances in the design of algorithms, few researchers were publishing source code alongside their publications, which created obstacles for reproducibility and scientific progress, as researchers were spending months reimplementing algorithms in code before they could even build upon them. They noted that this scientific culture contrasted sharply with fields like bioinformatics, where the sharing and development of OSS had become the foundation of scientific advances. They sought to solve this problem by launching a new publication track in the *Journal of Machine Learning Research* that was specifically dedicated to OSS, with the hope that it would incentivise researchers across the world to share and collaborate on OSS for ML and, more broadly, AI (Sonnenburg et al., 2007).

Fast forward to 2024 and the role of open source in AI research and development (R&D) is radically different than it was in 2007. Today, AI researchers and developers can choose from and contribute to a rapidly growing ecosystem of open source AI (OSAI) technologies, counting over a million AI OSS repositories on GitHub and over a million open model[4] (OM) repositories on Hugging Face (HF) Hub. The OSAI ecosystem is not only large in quantity, but also diverse in scope. It includes OSS, OMs, and open datasets that span AI sub-fields, such as ML, natural language processing, and computer vision. It includes deep learning (DL) frameworks like PyTorch and TensorFlow

[1] AI is a research field that concerns the development of "machine-based systems that, for explicit or implicit objectives, infer, from the input [they] receive, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (OECD, 2024).

[2] OSS is software whose source code is made available under a license that allows anyone to freely study, use, copy, modify, and redistribute it (OSI, 2007).

[3] ML is a sub-field of AI that concerns statistical learning, where "a computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997).

[4] OMs are AI models whose learned parameters (i.e. pre-trained weights and biases) and architecture are released under licenses that permit their use, study, modification, and redistribution (White et al., 2024).

and open foundation models[5] (FMs) like Meta's Llama models and Mistral AI's Mixtral models. It includes grassroots OSAI projects that are led and governed by communities of developers,[6] as well as corporate OSAI projects that are hosted and governed by a single company. In short, the concerns of the 16 scientists in 2007 now seem like distant history, as open source has become a cornerstone of AI R&D and the wider AI industry (Langenkamp & Yue, 2022; Law & Krier, 2023).

The meteoric growth of open source is not unique to AI—the last twenty years have witnessed the emergence and proliferation of open source technologies across the software industry, turning OSS from a fringe activity of hobbyists and hackers to a quotidian mode of software development in the global digital economy (Broca, 2021). The last decade, in particular, has seen the acceleration of commercial engagement with open source (Germonprez, Link, Lumbard, & Goggins, 2018), leading to what scholars have called "the incorporation of the digital commons" (Birkinbine, 2020) and the emergence of a "commons of capital" (Calimaq, 2018). The growth of the OSAI ecosystem is part of this wider trend, with companies contributing both code and cash, from open-sourcing AI software and large language models to funding and collaborating in OSAI projects. However, what is unique about AI is the eye-watering sums that companies, in particular technology giants, have invested in AI R&D, which make public funding look trivial by comparison. For example, while the European Union (EU) provided $1.2 billion in funding for AI R&D in 2021, Alphabet spent $1.5 billion on its subsidiary Google DeepMind alone in 2019 (Ahmed, Wahed, & Thompson, 2023).

This OSAI boom has received mixed reactions. On the one hand, under the banner of "AI democratisation," proponents celebrate open access to state-of-the-art AI technologies as a significant barrier-remover that enables AI R&D beyond a handful of industry leaders that have an unparalleled command of the necessary finance, talent, data, and compute resources (Kapoor et al., 2024; Seger, Ovadya, Garfinkel, Siddarth, & Dafoe, 2023). Others cite the benefits for reproducibility and scientific progress, as researchers can build on the collective shoulders of the community (Law & Krier, 2023; Sonnenburg et al., 2007). OSAI technologies are viewed as a potential antidote to an industry that is dominated by a few technology giants and well-funded scale-ups (Acquisti et al., 2024). World leaders have pledged to invest millions in OSAI development (Chatterjee & Volpicelli, 2023) and venture capitalists have bullishly invested in OSAI start-ups (Wiggers, 2023; Abboud, Levingston, & Hammond, 2024). Non-profit foundations have also thrown their hat in the ring; for example, the Mozilla Foundation has invested $30 million in its *mozilla.ai* initiative that seeks to build a trustwor-

[5] FMs are large AI models (up to trillions of parameters), which are based on neural network architectures and characterised by their ability to perform a variety of tasks, such as generating text, code, and images (Bommasani et al., 2022) Open FMs are FMs whose parameters are released under licenses that permit their use, study, modification, and redistribution (White et al., 2024).

[6] Such OSS projects can be considered digital commons; that is, digital resources that are developed and governed by an online community (Dulong de Rosnay & Stalder, 2020).

thy and independent OSAI ecosystem "outside of Big Tech and academia" (Mozilla, 2023).

On the other hand, critical scholars temper the optimism surrounding OSAI, highlighting how, behind the guise of ethical capitalism, it presents an avenue through which industry leaders actively seek to extend their influence in AI R&D (Widder, West, & Whittaker, 2023). For example, Srnicek (2022) argues that apparent acts of altruism like Meta and Google's respective open-sourcing of their DL frameworks PyTorch and TensorFlow obfuscate fierce market competition between the giants, who seek to shape open standards and crowdsource innovation. This argument was evident in a leaked Google memo, which warned that "open source [AI] solutions will out-compete companies like Google or OpenAI" and cited the lack of their "moat" against OSAI development as a reason to "own the ecosystem" and "to let open source work for us" (Patel & Ahmad, 2023). Another reason to temper the optimism is that the development and deployment of AI systems, open source or not, are dependent on material dependencies, from AI accelerators (Sastry et al., 2024) to data centres (Lehdonvirta, Wú, & Hawkins, 2024). In short: open source may be disruptive at the software or model layer of the AI stack, but it does not alter the material dependencies in this wider supply chain.

The debate about the promise and perils of OSAI animates the nascent research agenda on the political economy of OSAI, which has been spearheaded by Widder et al. (2023) in their seminar work, "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI," as well as, "Why 'open' AI systems are actually closed, and why this matters" (2024).[7] In these texts, the authors apply a political economy lens to scrutinise the interests, rhetoric, and actions of dominant corporations in the OSAI ecosystem, and the implications thereof for alternative AI futures. By drawing on historical examples of how companies have sought to assert market "dominance through—not in spite of—open-source," they (2024, p.828) draw our attention to how corporations are pursuing similar strategies to OSAI, provoking the question of how the development of OSAI may serve the interests of already powerful corporations, rather than leading to the democratising or equalizing effects that are celebrated. Furthermore, they discuss how the rhetoric of "openness" distracts practitioners and policymakers alike from the material dependencies of AI development—in particular, data, talent, and compute—that remain concentrated among a handful of technology giants. This critical analysis underscores that, when viewed within the wider supply chain, OSAI may offer limited "democratising" or equalising effects to the structure of the AI industry.

This thesis contributes to this research agenda through four empirical RPs that are collectively guided by the following meta research question (MRQ): ***"Why and how do companies participate in the development of OSAI technologies, and what are the implications thereof for development practices, governance norms, and the potential trajectories of the OSAI ecosystem?"*** This inquiry

[7] N.B. I deviate from the authors' use of "the political economy open AI" and opt for "the political economy of OSAI" to avoid confusions with OpenAI, the proprietary AI company that is famous for its chatbot ChatGPT.

is timely because commercial participation in OSAI development shapes not only which technologies are developed but also how they are governed and who benefits from them. While commercial activity has undoubtedly contributed to the growth of the OSAI ecosystem and advances in AI R&D, we face the risk that commercial agendas shape the practices, norms, and trajectory of the OSAI ecosystem, whilst crowding out digital commons or public interest alternatives. Ultimately, understanding where and why companies invest their resources in OSAI——as well as where and why they do not——is crucial for identifying pathways to promote public interest OSAI initiatives.

1.2 Theoretical Lens

In this section, I explain the theoretical lens that informs my empirical approach to the aforementioned MRQ and the four RPs that are presented in this thesis.

The primary objective of this thesis is to contribute to the nascent research agenda on the political economy of OSAI through empirical research. For this reason, I approach this MRQ through a political economy lens, building on the aforementioned work of Widder et al (2023, 2024). However, their work, as foundational texts of the political economy of OSAI literature, is marked by, at least, two theoretical limitations. First, they do not explicitly define the terms "political economy" and "power". These are terms that require defining, as their definitions inform what one can "see" or question. Second, they do not adequately draw on complementary theoretical frameworks from prior scholarship on the critical political economy of the digital commons and AI, which can inform this research agenda. In this thesis, I seek to address these limitations by defining these terms and bridging their work with these streams of scholarship. While I discuss these terms and prior work in more detail in Chapter 2, I briefly define them and explain how they shape my theoretical lens here.

To define "political economy" and "power", I draw on Mosco's (2009, p.24) political economy of communications, which provides a congruent framework for Widder et al's analysis of the interests, rhetoric, and actions of dominant corporations in the AI industry and the implications thereof for alternative AI futures. In particular, Mosco examines the material operation of the media sector, with a focus on "the social relations, particularly the power relations, that mutually constitute the production, distribution, and consumption of media resources." This framework has previously been applied by Birkinbine (2020) to study commercial interests and involvement in OSS development. As Birkinbine (2020, p.29) notes, research in this tradition of political economy is typically "directed at large corporations that hold extensive market power and the ability to influence the production, distribution, exhibition of, or access to communication resources." Here, power is understood both as a preventative and potential force; that is, "power manifests itself not just as a resource to achieve goals,

but also as a form of control that is embedded within a broader set of social relations" (Birkinbine, 2020, p.34). In the context of OSAI, power as a preventative force may be understood as the ability of dominant corporations to shape the conditions of participation in OSAI ecosystems in ways that discourage competition; for example, by making it costly for developers to switch away from their OSS or OMs. Meanwhile power as a potential force may be understood as the capacity of these corporations to mobilise external developers and resources in ways that reinforce their leadership in the AI industry, such as leveraging open source contributions to enhance their OSS or OMs or to entrench them as de facto industry standards. It is important to note that in this tradition it is also common to examine sites of struggle or resistance to corporate power. In the context of OSAI, this includes communities that develop and govern OSAI projects outside of the orbit of powerful corporations.

Now that I have defined these terms, I introduce two streams of scholarship that I draw on to expand Widder et al's foundation for the political economy of OSAI: the critical political economy of digital commons (Birkinbine, 2020) and the supply chain capitalism of AI (Valdivia, 2024).

In the former, Birkinbine (2020) applies a critical political economy lens to interrogate the social relations, particularly the power relations, that mutually constitute the production, distribution, and consumption of OSS. Through historical case studies, he examines how companies commodify OSS developed by OSS developer communities and the exploitative labour relations that underpin those processes. Unlike natural commons like rivers or pastures, he argues that companies do not seek to *enclose* digital commons like OSS projects; rather, they seek to *incorporate* their commons-based peer production in their own proprietary production, as the collective labour power of globally distributed OSS developer communities offers companies a fast and cheap way to develop and adopt software. Concretely, by participating in OSS development, Birkinbine contends that companies do not merely seek to capitalise on the *products* of OSS developer communities, but crucially on their *processes*. Following the critical political economy tradition, this framework also considers the conditions required to safeguard and sustain community-led OSS projects in the wake of exploitative tendencies by corporations. This framework is therefore complementary to Widder et al, offering theoretical tools for interrogating commercial interests and involvement in OSAI development, as well as identifying mechanisms that can support public interest OSAI initiatives and strategies.

In addition to the critical political economy of the digital commons, the supply chain capitalism of AI provides a complementary framework for contextualising OSAI development within the material dependencies that are required to develop and deploy AI systems, whether open source or not. The supply chain capitalism of AI concerns "the orchestration of commodity chains that extract, ship, and manufacture the natural resources needed to develop AI from an infrastructural perspective, such as mines, data centres, and e-waste dumps, together with their human resources [and labour]...Within

this chain, digital elements such as datasets and algorithms together with human labour (data annotators, data scientists, data engineers, etc.) are also key elements" (Valdivia, 2024, pp.5-6). I draw on this framework to expand Widder et al's (2024, p.827) analysis of "the materials—models, data, labour, frameworks, and computational power—frequently involved in creating and using large AI systems." In particular, locating OSAI development within the global AI supply chain equips us with a critical lens on the rhetoric surrounding the "democratising" promise of open-sourcing AI. In other words, this lens helps us see that the open-sourcing of AI software or models does little to change the fact that, for example, Nvidia commands a near-monopoly in the AI accelerator market (Leswing, 2024; Vipra & Myers West, 2023) or that three cloud hyperscalers—that is, Google Cloud, Microsoft Azure, and Amazon Web Services (AWS)—dominate the cloud compute market (Lehdonvirta, Wu, & Hawkins, 2023). Furthermore, in Chapter 2, I use this framework to organise prior work on the political economy of AI, in particular on industry concentrations at each stage of the AI supply chain.

By drawing on and bridging these previously disconnected streams of scholarship, I extend Widder et al's foundations for the nascent research agenda on the political economy of OSAI. In turn, I apply this holistic theoretical lens to empirically examine why and how companies participate in the development of OSAI technologies, and the implications thereof for development practices, governance norms, and the potential trajectories of the OSAI ecosystem. In the next section, I outline the structure of the thesis and discuss how the four empirical RPs collectively address the MRQ.

1.3 Outline of Thesis

This thesis is an integrated thesis that comprises eight chapters, including an overarching literature review, a methodology chapter, four RPs (see overview in Table 1.1), and a conclusion. In this section, I provide an overview of each chapter and explain how the four empirical RPs address the MRQ.

Following the Introduction, Chapter 2 presents an overarching literature review for the thesis, which extends the work of Widder et al (2023, 2024) by bridging it with scholarship on critical political economy of AI and the digital commons. It is structured in two parts. First, it situates OSAI in the global supply chain of AI and discusses work on industry concentrations across the supply chain. It also discusses prior work on the history of open source in AI R&D. Then, it reviews prior work on the political economy of OSS, with a particular focus on the critical political economy of the digital commons. This chapter also reviews prior work on the history, incentives, and effects of commercial participation in OSS development. Please note that, as the RPs were written independently, there is repetition between this literature review and the literature reviews in the RPs.

Chapter 3 discusses ex-ante considerations and ex-post reflections on the research design beyond

what will be discussed in the RPs. It includes thesis-wide considerations, such as my pragmatic epistemology and approach to case study research design, as well as RP-specific research methods and reflections, such as semi-structured interviews, data mining, and social network analysis (SNA). The reflections include a reflexive discussion of my positionality and how it influenced my research, as well as the challenges that I faced during the research and how I navigated them.

Chapter 4 (RP1) examines collaboration between companies in AI OSS projects that are hosted and governed by one company. With prior work identifying open governance and vendor-neutral hosting of OSS projects as structural enablers for inter-company collaborations, the effects of single-vendor governance and commercial hosting thereon remain unexplored. RP1 addresses this gap through a mixed-methods analysis of three critical AI OSS projects: Google's TensorFlow, Meta's PyTorch, and HF's Transformers. This RP shows how single-vendor governance creates a power imbalance that shapes collaboration in multiple ways, from the centralisation of development to varying community involvement approaches (from over-control to over-delegation). Furthermore, it characterises three types of inter-company collaborations—strategic, contractual, and non-strategic—-which encompass varying development incentives and practices. Among the strategic collaborations, it observes three types of strategic collaborations between AI companies, AI accelerator manufacturers, and cloud providers that reflect competitive dynamics among dominant companies across the AI supply chain. While the fruits of strategic collaborations are shared openly in repositories, they mostly take place privately between companies with the mutual aim of ensuring the competitiveness of their interdependent hardware, software, models, and cloud compute services.

Beyond maintaining control over their own OSS projects, companies have donated AI software to vendor-neutral foundations under the banner of AI democratisation. While donations are commonly marketed as altruistic acts, the commercial interests are typically hidden from public view. Chapter 5 (RP2) investigates the commercial incentives behind such donations through a mixed-methods analysis of 43 AI software donations to the Linux Foundation (LF). The RP identifies and categorises social, economic, and technological incentives at the level of the individual developer and company that motivate donations. The findings highlight how companies, from start-ups to multinational corporations, strategically adopt open governance and vendor-neutral hosting as means to attract new developers, reduce development costs, and shape industry standards, among others. The RP concludes with a discussion of the generalisability of the findings of commercial incentives to OSS donations more broadly as well as to other AI democratisation efforts, such as the release of OMs.

Chapter 6 (RP3) focuses on the development of OMs. While OMs are viewed as a potential threat to proprietary AI models (Patel & Ahmad, 2023), the ways in which OMs are developed or adopted remains poorly understood. RP3 addresses this gap with a quantitative analysis of development

activity in over 500,000 repositories on HF Hub, an increasingly popular platform for developing, sharing, and demonstrating OMs. This RP contributes three key findings. First, various types of activity, from likes and downloads to discussions and commits, in the repositories of models, datasets, and "spaces" (i.e. model demos) are characterised by extreme imbalances. For example, over 70% of OMs have had zero downloads whilst 1% account for 99% of all downloads. Second, the social network of collaboration in model repositories is characterised by a core-periphery structure, with 89% of developers working in isolation and a small core of prolific developers who collaborate intensively. Third, a small subset of models, developed by a few industry leaders, account for the majority of model adoption in spaces. The findings provide an empirical baseline for further study on OM development, whilst challenging prevalent narratives about the "democratising" effects of OSAI.

The commercial dominance in the OSAI ecosystem raises critical questions about the feasibility and sustainability of community-led OSAI projects; that is, AI OSS or OMs that are developed and governed by a community of developers, who are not collectively motivated by commercial gain. These concerns extend beyond OSAI to the digital commons more widely, where funding is increasingly understood as a vital yet undersupplied support mechanism. Chapter 7 (RP4) contributes to this debate through a case study on the role of funding in supporting the scikit-learn project, a widely used Python library for ML that is developed by researchers based at the French Institute for Research in Computer Science and Automation (Inria) and a community of over two thousand volunteers. Through 25 interviews with its maintainers and funders, this RP sheds light on the merits and drawbacks of public and private funding, and offers recommendations for how governments and companies can effectively co-operate to fund community-led OSAI projects.

Overall, the four RPs address different aspects of the MRQ. Chapter 4 (RP1) addresses the "how" and "why" of the MRQ by examining collaborative and competitive dynamics between companies, as well as the "implications" by examining how unilateral governance influences inter-company collaboration practices and possibilities. Chapter 5 (RP2) tackles the "why" of the MRQ by exploring the commercial incentives behind AI democratisation. Chapter 6 (RP3) addresses both the "how" and "implications" by analysing OM development and adoption patterns on HF Hub. Chapter 7 (RP4) examines the "implications" by investigating the role of funding as a lever for sustaining community-led OSS projects, as well as how funding influences governance choices and project priorities. Collectively, these RPs provide a multi-faceted investigation of commercial participation in OSAI development, contributing both theoretical and empirical findings to the political economy of OSAI literature.

Finally, Chapter 8 concludes with a discussion of the overarching contributions to and future directions for the nascent research agenda on the political economy of OSAI.

Table 1.1: Summary of Research Papers

MRQ: *Why and how do companies participate in the development of OSAI technologies, and what are the implications thereof for development practices, governance norms, and the potential trajectories of the OSAI ecosystem?*				
RP	**RP1**	**RP2**	**RP3**	**RP4**
Title	*Characterising Open Source Co-opetition in Company-hosted OSS Projects: The Cases of TensorFlow, PyTorch, and Transformers*	*Why Do Companies Democratise AI? The Case of Software Donations to Foundations*	*The AI Community Building the Future? A Quantitative Analysis of Development Activity on the HF Hub*	*OSS Developers' Views on Public and Private Funding: A Case Study on scikit-learn*
Research Questions	**RQ1:** What, if any, are typical structures and patterns of open source co-opetition in company-hosted OSS projects? **RQ2:** What types of collaborative relationships do companies pursue in company-hosted OSS projects, and why? **RQ3:** What similarities and differences characterise co-opetition practices in company-hosted versus foundation-hosted OSS projects?	**RQ1:** Why do companies democratise AI?	**RQ1**: What are typical patterns of development activity on HF Hub? **RQ2**: What is the social network structure of the HF developer community? **RQ3**: What is the distribution of model development and adoption activity on HF Hub, and which actors develop the most adopted models?	**RQ1:** What are the interests of the public and private funders of the scikit-learn project, and how have they aligned or conflicted with the interests of the maintainers? **RQ2:** How do the scikit-learn maintainers view the relative benefits and drawbacks of public and private funding?
Methods	Data mining SNA Interviews	Questionnaires Interviews Document analysis	Data mining SNA	Interviews Document analysis
Relevance to MRQ	"How" and "why": Collaborative and competitive dynamics between companies	"Why": Commercial incentives for OSS donations as method of AI democratisation	"How" and "implications": OM development activity on HF Hub	"Implications": Role of funding for sustainability of non-commercial projects
Peer Review Status	**Accepted/Published:** Proceedings of ACM HCI (CSCW)	**Accepted/Presented:** Creative Commons Global Summit 2023, Cuidad de México **Under Review:** ACM Transactions on Software Engineering & Methodology	**Accepted:** International Conference for Computational Social Science **Accepted/Published:** Journal of Computational Social Science	**Accepted/Presented:** OpenForum Academy Symposium 2023, TU Berlin, Berlin **Accepted/Published:** Proceedings of ACM HCI (CSCW)

2. Literature Review

This chapter presents the overarching literature review for the thesis. It is structured into two parts. First, it reviews prior work on the political economy of AI, with a focus on the supply chain capitalism of AI and prior work on industry concentrations across the AI supply chain. It also discusses prior work on the history and role of open source in AI R&D. Second, it presents prior work on the political economy of OSS, with a focus on how commercial activity has reshaped OSS development from a social movement of hackers and hobbyists to a quotidian practice in the global software industry. In particular, it discusses the critical political economy of the digital commons as a theoretical framework for investigating commercial interests and involvement in OSS development. The purpose of this review is to bridge these two streams of scholarship, which taken together provide a comprehensive foundation for ongoing and future scholarship on the political economy of OSAI development.

2.1 Building Moats and Ecosystems: The Political Economy of OSAI

2.1.1 Situating AI R&D within the AI Value Chain

2.1.1.1 Defining AI: A Moving Target

AI is a contested and confused term. There are many different definitions of AI, and terms such as ML and AI are often conflated (Krafft, Young, Katell, Huang, & Bugingo, 2020). For example, a survey found that AI researchers and policymakers have divergent understandings of AI: "While AI researchers favour definitions of AI that emphasise technical functionality, policymakers instead use definitions that compare systems to human thinking and behaviour" (Krafft et al., 2020, p.1). This survey found that the definition that won the most agreement among both AI researchers and policymakers was the definition developed and adopted by the OECD (2024):

> *"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

I adopt this definition of AI because it captures a wide spectrum of AI implementations without being tied to specific technical approaches, from various ML approaches used for prediction or classification to content-generating language models. In addition, it has practical value due to its support by both technical and non-technical stakeholders (Krafft et al., 2020) and its adoption by governments and international institutions, including the EU's AI Act (H. Roberts et al., 2023). At the same time, I acknowledge that "there is much at stake in how we define AI, what its boundaries are, and who determines them: it shapes what can be seen and contested... The description of AI as fundamentally abstract distances it from the energy, labour, and capital needed to produce it" (Crawford, 2021, p.217). Furthermore, I note that the definitional goal posts of AI keep on changing. Changes are in part due to advances in AI research. Indeed, the research field has mushroomed in a short period. For example, a study in 2019 found that 77% of AI research papers on *arXiv*, a repository of open access preprints, were published between 2014 and 2019 (Mateos-Garcia, Klinger, Stathoulopoulos, & Winch, 2019). What is more, there have been a number of scientific breakthroughs in sub-fields such as deep reinforcement learning, exemplified by Google DeepMind's development of the AlphaGo programme (Silver et al., 2017), to the invention of the transformers model architecture that is used to developlarge language models (LLMs) like OpenAI's GPT models (Vaswani et al., 2023).

However, critical scholars contend that changes in definitions of AI are not merely a result of research advances; rather, commercial or financial logics are at play when corporate or academic labs claim to achieve scientific breakthroughs or request the funds to do so (Widder et al., 2023). For example, in August 2021, Stanford University announced the establishment of its Center for Research on Foundation Models, which was "a new initiative [that] brings together more than 175 researchers across 10+ departments at Stanford University to understand and build a new type of technology that will power AI systems in the future"(Stanford, 2021). The centre focuses on making "fundamental advances in the study, development, and deployment of foundation models" (Stanford, 2021); that is, large AI models that have up to trillions of parameters, which are characterised by their ability to perform a variety of tasks, such as generating text, code, and images, thus providing a "foundation" that can be adapted through fine-tuning for downstream tasks (Bommasani et al., 2022). Whittaker (2021, p.51) has a more critical perspective on these breakthroughs, contending that, "The 'advances' in AI celebrated over the past decade were not due to fundamental scientific breakthroughs in AI techniques. They were and are primarily the product of significantly concentrated data and compute resources that reside in the hands of a few large tech corporations." These critiques underline that the definition of AI remains a subject of fierce debate characterised by vested interests.

The growth of the scientific field is in large part due to the eye-watering investments into AI R&D by both the public and private sectors. Governments have industriously published AI strategies to

unlock public funds for AI R&D and boost national competitiveness (H. Roberts et al., 2023), aiming to seize on the promises of AI, from medicine (Yeasmin, 2019; Rajpurkar, Chen, Banerjee, & Topol, 2022) to education (Chan & Hu, 2023; L. Chen, Chen, & Lin, 2020) to defence (Widder, Gururaja, & Suchman, 2024; Carlo, 2021). However, public funding remains minuscule compared to industry's investments (Evas et al., 2022). In 2021, government agencies except for the defence sector in the United States of America (USA) allocated $1.5 billion and the European Union (EU) planned to spend $1.2 billion on AI R&D, whilst industry globally spent over $340 billion on AI R&D that year and in 2019 Alphabet spent $1.5 billion on its subsidiary Google DeepMind alone (Ahmed et al., 2023).

Consequently, corporate labs have overtaken academic labs as the leaders of the AI research field, measured by the quantity of their scientific publications and model benchmarks (Ahmed et al., 2023). Furthermore, an analysis of over 170,000 papers from prestigious computer science conferences between 2000 and 2019 shows that technology giants and elite universities (i.e. those ranked in the top 1-50 in the QS World University Rankings) have increased their dominance in AI research since 2012 through increased firm-only publications and firm-elite university collaborations, crowding out mid- and lower-tier universities from these conferences (Ahmed & Wahed, 2020). Industry giants also wield considerable influence on AI R&D in academic labs through the selective financing of research areas that are relevant to their commercial agendas (Gulson & Webb, 2021; Klinger, Mateos-Garcia, & Stathoulopoulos, 2020; Whittaker, 2021). Furthermore, the open science ethos in AI R&D is marked by knowledge asymmetries between academia and industry, as academic labs share their preprints which can be capitalised on by commercial actors while corporate labs are more selective about what they share in the interest of maintaining their market competitiveness (Gulson & Webb, 2021).

The advances in AI research have been widely commercialised and consequently AI technologies are becoming increasingly commonplace by the day, from recommender systems on social media platforms to generative AI chatbots like ChatGPT that made AI a household name (C. David & Paul, 2023). While they are prized for their benefits for research and innovation, the development and deployment of AI technologies in a range of settings have caused a number of social, economic, and environmental harms. These include algorithmic discrimination against marginalised communities (Buolamwini & Gebru, 2018; Raji et al., 2020), the exploitation of click workers in the Global South (Muldoon, Graham, & Cant, 2024; Gray & Suri, 2019), and environmental damages from the AI supply chain (Valdivia, 2024), from mineral extraction (Crawford, 2021; Bobba, Carrara, Huisman, Mathieux, & Pavel, 2020) to wasteful energy and water consumption (Valdivia, 2024; Luccioni, Jernite, & Strubell, 2024) to e-waste (Wang, Zhang, Tzachor, & Chen, 2024; Gabrys, 2013). In response, governments have developed both hard and soft regulation to govern the harms of AI development, from the EU's AI Act to the USA's AI Bill of Rights (H. Roberts et al., 2023).

2.1.1.2 The Supply Chain of AI: Mapping its Structure and Industry Concentrations

With the ever-prevalent role and impact of AI technologies on civilian life, scholars have begun to examine the critical political economy of AI; that is, the social relations, particularly the power relations, that mutually constitute the development, distribution, and use of AI technologies (Mosco, 2009). In this section, I draw on frameworks from prior work to situate AI R&D within its wider supply chain and discuss the power configurations that characterise its supply chain.

To do so, I draw on the supply chain capitalism of AI, which provides a critical infrastructural framework that maps the global supply chain of AI development and its constituent parts (Valdivia, 2024). More specifically, drawing on the supply chain capitalism framework (Tsing, 2009), Valdivia (2024, pp.5-6) defines the supply chain capitalism of AI as:

> *"The orchestration of commodity chains that extract, ship, and manufacture the natural resources needed to develop AI from an infrastructural perspective, such as mines, data centres, and e-waste dumps, together with their human resources (miners, drivers, directors of data centres operations, e-waste dismantlers, etc.). Within this chain, digital elements such as datasets and algorithms together with human labour (data annotators, data scientists, data engineers, etc.) are also key elements."*

The value of this framework for the political economy of OSAI is that it provides a sobering reminder that the development of more and more efficient algorithms or larger and larger AI models does not simply take place in the online communities that meet in repositories, mailing lists, or Discord channels. Rather, the development of AI technologies depends on a number of material dependencies, from raw minerals and hardware to data centres and developer platforms. In addition, it highlights the many devastating material impacts of AI development on the environment and local communities throughout the supply chain. Crawford's (2021) "Atlas of AI" is complementary to this framework. Upon mapping the minerals, labour, and data that fuel AI development, Crawford characterises AI as a "technology of extraction," whose labour force extends far beyond high-tech researchers and developers in innovation hubs in the Global North like Silicon Valley. For example, it includes an"invisible" labour force of click workers, largely based in the Global South, who play an indispensable role in supporting "the illusion of automation" by cleaning and labelling data under arduous working conditions (Crawford, 2021; Gray & Suri, 2019; Muldoon et al., 2024).

In addition, I draw on the political economy of open AI, a novel framework that critically analyses "the materials—models, data, labour, frameworks, and computational power—frequently involved in creating and using large AI systems" (Widder, Whittaker, & West, 2024, p.827). In this work, Widder, Whittaker, and West (2024, pp.827-828) highlight the material dependencies in the development of

large “open source” AI models, and draw parallels to historical examples of commercial methods of “asserting market dominance through—not in spite of—OSS” in order to illustrate “how rhetoric around ‘open’ AI is frequently wielded in ways that exacerbate rather than reduce concentration of power in the AI sector.” While OSAI may offer benefits for AI R&D through enabling transparency, reusability, and extensibility, they argue that these affordances alone cannot un-do industry concentrations in the contemporary, compute-intensive nature of AI R&D (Widder, Whittaker, & West, 2024). Taken together, these frameworks form a useful organising framework for the growing body of scholarship on industry concentrations, including monopolies and oligopolies, that exist at each stage of the AI supply chain, which I discuss below and summarise in Table 2.1.

On top of these frameworks, I draw on the platform capitalism literature to view OSAI development within the context of the market power and influence wielded by the companies that own and operate the digital platforms that AI researchers and developers depend on to share, access, develop, and deploy OSAI technologies, from the cloud hyperscalers to developer platforms like GitHub and HF Hub (Gawer & Cusumano, 2002; Srnicek, 2017). Gawer and Cusumano (2002, p.417) define industry platforms as the “products, services, or technologies that act as a foundation upon which external innovators, organized as an innovative business ecosystem, can develop their own complementary products, technologies, or services.” Meanwhile Srnicek (2017, p.43) defines digital platforms as “the digital infrastructures that enable two or more groups to interact; they are intermediaries between different user categories and often they come with tools that enable users to build their own products, services, and marketplaces.” This literature illustrates that the ownership and operation of platforms confers platform owners two key benefits. First, they gain “control and governance over the rules of the ‘game’,” which range from policies that permit or disallow certain types of activity to what types of activity are technically possible through platform features. Second, the value of platforms grow through network effects as their user base expands. (Srnicek, 2017, p.45, p.95) contends that, “Platforms produce and are reliant on network effects, which have a natural tendency towards monopolisation” and “network effects tend to mean that early advantages become solidified as permanent positions of industry leadership.” Similarly, Gawer and Cusumano (2014, p.421) argue that, “platforms tend to facilitate and increase the degree of innovation on complementary products and services. The more innovation there is on complements, the more value it creates for the platform and its users via network effects, creating a cumulative advantage for existing platforms: As they grow in adoption, they become harder to dislodge by rivals or new entrants, with the growing number of complements acting like a barrier to entry.”

What is more, Gawer and Cusumano (2014, p.417) connect the concepts of industry platforms with ecosystem innovation. Software ecosystems are“a set of businesses functioning as a unit and

Phase	Dominant Firms	Control	Implications
Hardware	Nvidia: Near-monopoly for GPUs. Arm, AMD, Google, Intel, IBM (market rivals).	Dominate production of specialized hardware (e.g., GPUs, TPUs) required for training large-scale AI models.	Pricing, supply chain decisions, and technological advancements influence AI developers' capabilities.
Cloud Compute	AWS, Google Cloud, Microsoft Azure: Oligopoly of hyperscalers.	Control computational resources for training, hosting, and deploying AI models.	Set pricing, access rules, and availability, impacting feasibility of OSAI projects.
Developer Platforms	Microsoft (GitHub) and HF (HF Hub): Key platforms for OSS and OM sharing, hosting, and collaboration.	Centralise repositories, tools, and collaborative environments.	Policies (e.g., access fees, content moderation, API changes) shape direction and accessibility of open source activity.
Training Data	OpenAI, Google, Meta: Access to training datasets via their own platforms and/or partnerships with providers (e.g., publishers).	Control critical data resources, despite the existence of open data alternatives.	Proprietary datasets often outperform open data, influencing competitive dynamics in AI.

Table 2.1: Leading Companies in Four Stages of the AI Supply Chain

interacting with a shared market for software and services, together with the relationships among them" (Jansen et al., 2009, p.9). The strategic benefits of commercial ownership of software ecosystems are similar to the ownership of industry platforms, as other market actors are dependent on this software and build complementary products, technologies, or services on their technologies. This framework is instructive for interrogating the commercial logics of industry giants like Google and Meta which, as I discuss in more detail in section 2.1.2, seek to "own the [OSAI] ecosystem" and "to let open source work for us" (Patel & Ahmad, 2023).

While this thesis does not specifically respond to debates in the platform capitalism literature, it provides useful theoretical tools to contextualise OSAI development within the power structures of its wider supply chain. For example, it helps us to understand how AI democratisation efforts do little to challenge the market power of technology giants that dominate the different stages of the AI supply chain through their ownership of critical digital platforms. As Srnicek (2017, p.97) argues, "Open-sourcing all software or capital investments are not enough to overturn monopolies; access to data, network effects, and path dependency place even higher hurdles in the way of overcoming a monopoly like Google." In the rest of this section, I discuss industry concentrations across the AI supply chain: hardware, cloud compute, developer platforms, and data.

In the last decade, we have witnessed the emergence of a "bigger-is-better" paradigm that prizes the training of larger and larger AI models (Varoquaux, Luccioni, & Whittaker, 2024). This approach to the development and deployment of large AI models is extremely compute-intensive; and comput-

ing power required to train state-of-the-art AI models doubles approximately every six months (Heim et al., 2024). Consequently, compute power has become a critical resource in AI R&D (Lehdonvirta et al., 2024). The compute supply chain encompasses four main types of companies: (1) those that design and market accelerators like Nvidia's graphics processing units (GPUs), (2) those that fabricate and package chips, (3) those that deploy chips to provide compute services, and (4) those that consume compute to develop and deploy AI systems (OECD, 2023; Lehdonvirta et al., 2024).

The AI compute supply chain is characterised by monopolies and oligopolies at various stages: Nvidia has a near-monopoly over the accelerator (or "chip") market with 70-90% market share (Leswing, 2024); ASML has a monopoly over photolithography machine production (Miller, 2022); TSMC has a monopoly over chip fabrication (Miller, 2022); and three cloud hyperscalers—Google Cloud, Microsoft Azure, and AWS—dominate the cloud compute market (Lehdonvirta et al., 2023). Nvidia's dominance extends beyond its GPUs: with over four million developers dependent on CUDA, its partly proprietary framework that only supports training on GPUs, it has become the "de facto industry standard" for AI development (Economist, 2024). The expense of the compute-intensive, "bigger-is-better" approach to AI development has narrowed the field of companies that can afford to buy advanced accelerators for training frontier models (Varoquaux et al., 2024). For example, Nvidia's H100, their most powerful GPU at the time of writing, costs up to $40,000 per unit. Meta alone is estimated to have spent $18 billion on GPUs in 2024 (Hays, 2024). These exorbitant costs have created a divide between the "GPU rich" and "GPU poor" (Barr, 2023).

Beyond a few industry giants like Meta that can afford to pay for their own compute infrastructure, AI developers depend on cloud compute platforms that are powered by data centres—a market that is dominated by the three cloud hyperscalers: Google Cloud, Microsoft Azure, and AWS (Pilz & Heim, 2023). Their ownership of cloud compute infrastructure "is an immensely powerful and profitable position to be in... Cloud platforms are building up the basic infrastructure of the digital economy in a way that can be rented out profitably to others, while they collect data for their own uses" (Srnicek, 2017, pp.63-64). The commercial ownership of GPU-powered data centres is also of geopolitical significance, as these data centres are unevenly distributed geographically among "Compute North", "Compute South", and "Compute Desert" countries (Lehdonvirta et al., 2024). That is to say, countries that host data centres with advanced GPUs that are suited to frontier AI development, those with data centres with less powerful GPUs that are suited to AI deployment, and those no data centres who depend on foreign data centres (Lehdonvirta et al., 2024).

As a result, cloud hyperscalers have emerged as major power brokers in the AI supply chain, upon whose compute AI R&D depends (as shown in Figure 2.1a). Srnicek (2022) contends that the concentration of compute power, alongside skilled talent and data, among a handful of technology

giants and well-funded start-ups makes AI a centralising technology that compounds the concentration of power and capital in the IT sector. Furthermore, this concentration of compute power has led to a circular flow of capital between cloud hyperscalers and AI start-ups and scale-ups. The three cloud hyperscalers "contributed a full two-thirds of the $27 billion raised by fledgling AI companies in 2023" (Hammond, 2023), with the majority of capital raised by AI start-ups—"up to 80-90% in early rounds"—being paid back to the same cloud hyperscalers (Bornstein, Appenzeller, & Casado, 2023). This dependency is evident in the strategic partnerships formed between AI companies like Anthropic, OpenAI, and Mistral AI and the cloud hyperscalers (Sastry et al., 2024). The UK Competition and Markets Authority has documented the network of dependencies formed through strategic partnerships (see Figure 2.1b), noting that industry giants' control over crucial inputs for AI R&D provides them with significant leverage over AI start-ups and scale-ups (CMA, 2024).

In addition to cloud compute platforms, AI researchers and developers are dependent on two principal platforms for hosting (i.e. making available of), sharing (i.e. distribution of), and developing OSS and OMs: GitHub, which is owned by Microsoft, and HF Hub, which is owned by the start-up HF. As per the platform capitalism literature, these two companies benefit from their market position in two key ways: they are the rule-setters for what activity can take place and how, and their value increases though network effects (Srnicek, 2017). For example, they set development possibilities through the provision of certain platform features (Eghbal, 2020) and platform policies (Gorwa & Veale, 2024). Furthermore, GitHub benefits from network effects as its vast user base of developers makes it the default platform for hosting OSS projects, thereby attracting even more developers who want to collaborate on existing projects. Similarly, HF Hub benefits as its growing collection of models and datasets attracts more researchers and developers, further reinforcing its position as the go-to platform for sharing and accessing AI models and datasets, which in turn incentivises more researchers and developers to share their models or datasets on the platform to reach the platform's growing "AI community [that is] building the future."

Data is also a critical resource for AI R&D. In fact, according to James Betker, a researcher at OpenAI, it is the most important resource: "model behavior is not determined by architecture, hyperparameters, or optimizer choices. It's determined by your dataset, nothing else... When you refer to 'Lambda', 'ChatGPT', 'Bard', or 'Claude' then, it's not the model weights that you are referring to. It's the dataset" (Betker, 2023). Data access is not evenly distributed. Platforms—from social media platforms and e-commerce platforms to cloud compute platforms and developer platforms—have a data advantage, which stems from their ability to "record the activities that take place on their platforms as data and in turn extract value from this data" (Srnicek, 2017, p.44). The business model of surveillance capitalism has made a handful of industry giants that own the digital platforms in-

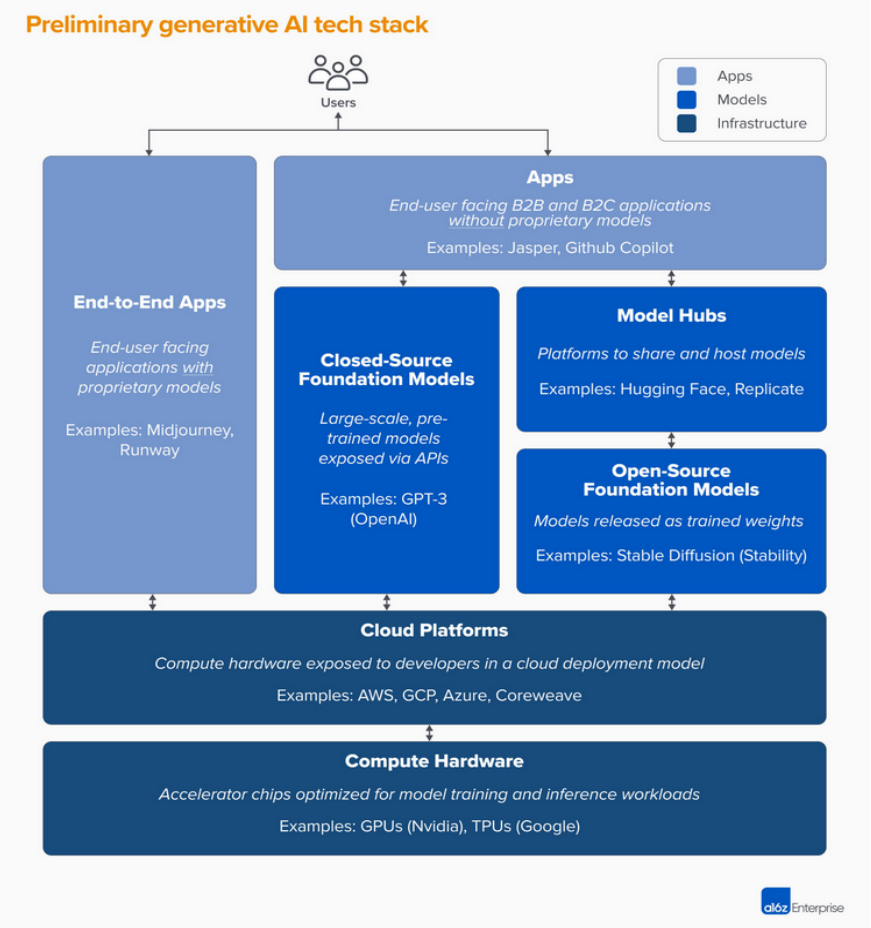
credibly data-rich (Zuboff, 2019). Beyond these platforms, AI companies have feverishly mined data from the Internet, including news websites, Wikipedia, and Stack Overflow. However, many websites have resisted data extraction by introducing blockers for crawlers; for example, "by the end 2023, 48% of the most widely used news websites across ten countries were blocking OpenAI's (ChatGPT) crawlers" (Fletcher, 2024). Furthermore, AI companies have been sued for using pirated copyrighted material to train their models, such as The Pile, a 886 GB text dataset created by EleutherAI in 2020 (Paul, 2023; Roth, 2024). AI companies like OpenAI and Antrophic have sought additional sources of data by forming strategic partnerships with media conglomerates. For example, OpenAI formed a partnership with News Corp, providing it with access to current and archived content from News Corp's news outlets (OpenAI, 2024). However, buying data is "costing more and more—making it out of reach for all but the wealthiest tech companies" (Wiggers, 2024).

These concentrations across the AI supply chain raise fundamental questions about the future trajectory of AI development. As industry giants consolidate their control over critical resources, there is growing concern about the sustainability and democratic implications of the current "bigger-is-better" paradigm, which I discuss in the following section.

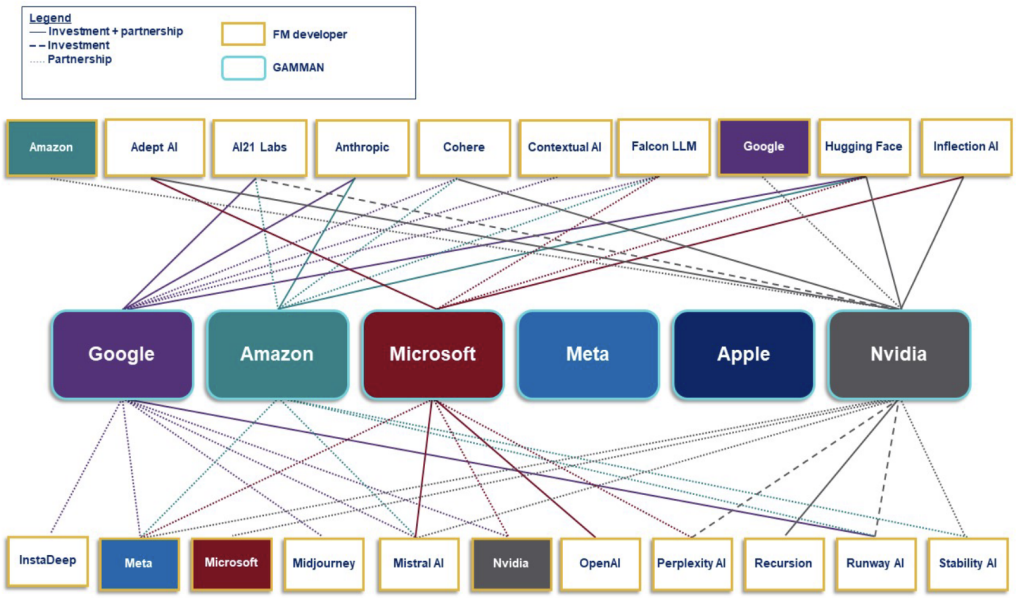
2.1.1.3 The Perils of the "Bigger-Is-Better" Paradigm and Alternative Public AI Pathways

The structural conditions in the AI supply chain have important implications for the potential of the OSAI commons to deliver the democratising effects that are celebrated or hoped for in AI R&D. As Widder, Whittaker, and West (2024, p.831) argue, "the pursuit of even the most open AI will not on its own lead to a more diverse, accountable or democratized ecosystem, even though it may have other benefits." Given the dominance of industry giants in contemporary AI R&D, there have been calls for alternative futures that are more conducive to the public interest.

The Mozilla Foundation has become a proactive advocate of public interest OSAI. In March 2023, it invested $30 million into its newly formed *mozilla.ai* to build a trustworthy and independent OSAI ecosystem "outside of Big Tech and academia" (Mozilla, 2023). In September 2024, Marda, Sun, and Surman (2024) published the "Public AI" manifesto, warning that we face the risk that "an AI ecosystem powered exclusively by the market will prioritize a narrow set of profitable applications," while socially-beneficial applications are under-resourced. They argue, "we can't rely on a few companies to build everything our society needs from AI, and we can't afford the risk that they won't," and propose "a vision for a robust ecosystem of initiatives that promote public goods, public orientation, and public use throughout every step of AI development and deployment" (Marda et al., 2024, p.4). They contend Public AI "will need serious financial, community, and political backing to enable it to become a meaningful alternative to the closed, commercial ecosystem" and diverse stakeholders



(a) The Technology Stack of Generative AI Systems
Source: Bornstein et al. (2023)



(b) Partnerships between Industry Giants and AI Scale-ups
Source: CMA (2024, p.18)

Figure 2.1: Dependencies and Partnerships in the AI Supply Chain

should have a role to play, from developers to policymakers to the public at large.

Similarly, in August 2024, the Public AI Network published a proposal for "Public AI: Infrastructure for the Common Good" (Public.AI, 2024). It outlines three essential features for AI development in the public interest: universal public access to key AI capabilities, public accountability in governance and development, and permanent public goods that resist corporate capture. It calls for public investments across three areas: horizontal platforms providing public options for core AI building blocks like compute and data, vertical platforms offering end-to-end public AI services, and new public goods targeting societal challenges that are underserved by companies. The proposal argues this approach would expand access to AI benefits while ensuring development aligned with public values. They note that initiatives demonstrating this vision are already emerging globally, such as the South East Asian Languages in One Network (SEA-LION) family of open LLMs by AI Singapore, a national R&D programme in AI, that "better understands Southeast Asia diverse contexts, languages, and cultures" (SEA-LION.AI, 2024). These examples demonstrate growing momentum for public alternatives to purely market-driven (OS)AI development.

An alternative proposal is to break away from the "bigger-is-better" paradigm altogether. Varoquaux et al. (2024) challenge two key assumptions underlying this paradigm: one, that improved performance necessarily requires increased scale; and two, that all significant AI problems require large-scale models. They argue that the focus on scale is both scientifically fragile and unsustainable, as compute demands grow faster than model performance, leading to unreasonable economic and environmental costs (Varoquaux et al., 2024). For example, GPUs require rare materials like silicon, copper, tantalum, aluminium, or tungsten (Bobba et al., 2020), and their short five-year lifespan generates significant e-waste, much of which ends up in landfills in the Global South (Valdivia, 2024). Furthermore, training LLMs can emit up to 550 tonnes of CO2, while model inference accounts for a substantial portion of AI-related energy use, ranging from one-third of Meta's AI carbon footprint (C.-J. Wu et al., 2022) to 60% of Google's AI energy consumption (Patterson et al., 2022). Varoquaux et al. (2024, p.10) make a call to action for an alternative trajectory of AI R&D that is more affordable, more environmentally friendly, and more conducive to democratic governance:

> *In recent years, AI research has acquired an unhealthy taste for scale. This comes with dire consequences—economic inequalities and environmental (un)sustainability, datasets that erode privacy and emphasize corrosive social elements, a narrowing of the field, and a structural exclusion of small actors such as most academic labs and many start-ups. This fixation on scale has emerged via norms that shape how the scientific community acts. We believe that scientific understanding and meaningful social benefits of AI will come from de-emphasizing scale as a blanket solution for all problems, instead focusing on models that can be run on*

widely-available hardware, at moderate costs. This will enable more actors to shape how AI systems are created and used, providing more immediate value in applications ranging from health to business, as well as enabling a more democratic practice of AI.

These proposals illustrate a growing momentum and political will to embark on a new path toward more democratic and sustainable AI development, in which open source—as a technology, as a mode of development, and as a mode of governance—plays a role, offering benefits like open access, open development, and open governance. However, as I discuss in the following section, open source is not a silver bullet to industry concentrations in AI R&D and is commonly used as an avenue by dominant industry actors to extend their influence and fortify their market position.

2.1.2 The Emergence of the OSAI Commons

2.1.2.1 A History of Open Source in AI R&D

The culture of openness in AI R&D has evolved significantly in the last 15 years. As recently as 2007, researchers lamented the dearth of OSS in 2007, arguing, "The field of ML has developed a large body of powerful learning algorithms for diverse applications... [However,] the true potential of these methods is not used, since existing implementations are not openly shared" (Sonnenburg et al., 2007, p.2444). However, nowadays AI R&D is hardly imaginable without open source thanks to a rapidly growing commons of hundreds of AI OSS libraries (Haddad, 2022), over a million AI OSS repositories on GitHub (GitHub, 2024a), and over a million OMs on HF Hub (HuggingFace, 2023b).

While AI OSS development began as an academic endeavour, commercial participation in AI OSS development became increasingly common in the mid-2010s and by many measures the OSAI ecosystem is dominated by industry (see Figure 2.2). In January 2022, 67.4% of the OSS libraries in the LF AI & Data database were hosted by companies (LFAI&Data, 2022). A handful of industry giants and non-profit foundation hosted a significant portion of the AI OSS libraries: Google (10.3%), the LF AI & Data Foundation (9.8%), Meta (8.7%), Microsoft (5.9%), and the Apache Software Foundation (5.9%). The most downloaded were scikit-learn (621,408,470), Google's TensorFlow (341,220,381), and Databricks' MLFlow (156,621,153). Meta's PyTorch ranked 21st with 4,373,593 downloads, which was surprising because by then it had already overtaken TensorFlow as the most adopted DL framework in academic AI research (PaperswithCode, 2023) and for training OMs (Foster, 2022; Gururaja, Bertsch, Na, Widder, & Strubell, 2023).[1] Google's TensorFlow, released in 2015, and Meta's PyTorch, released in 2016, are credited for the widespread adoption of DL in AI R&D and commercial applications (Langenkamp & Yue, 2022; Law & Krier, 2023).

[1] I speculate that this low download count understates PyTorch's actual usage, which may be due to downloads through alternative avenues, such as Docker containers or cloud platforms like Google Colab.

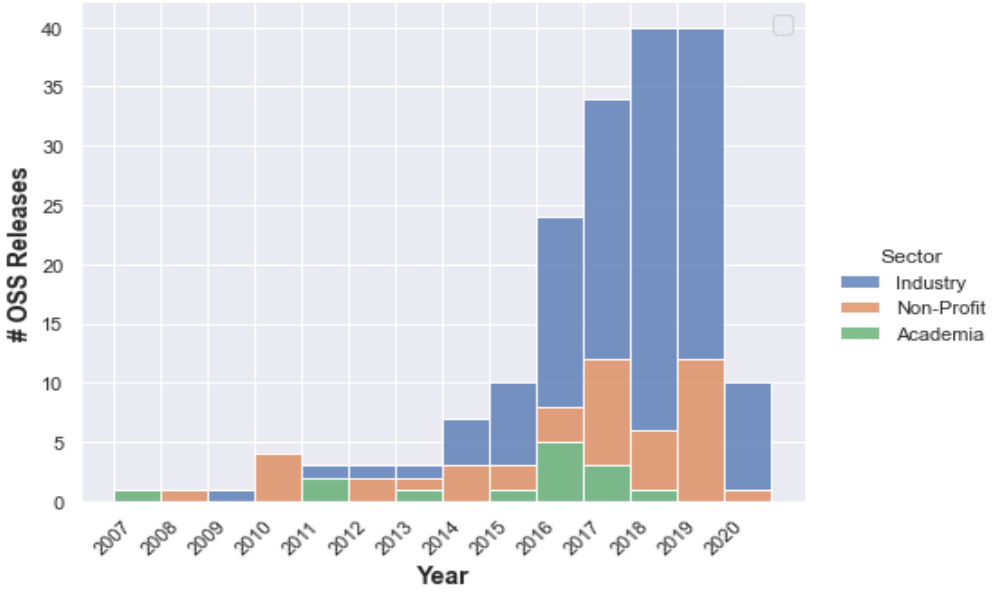


Figure 2.2: Timeline of AI OSS Releases
Source: LF AI & Data Foundation (2022)

Commercial releases of AI OSS are often framed in altruistic language. For example, upon the release of TensorFlow, Jeff Dean from Google stated, "We're hoping that the community adopts this as a good way of expressing ML algorithms of lots of different types and contributes to building and improving [TensorFlow] in lots of different and interesting ways" (Metz, 2015). One year later, Meta followed by releasing PyTorch, and a competitive rivalry over the adoption of their respective DL frameworks ensued (Langenkamp & Yue, 2022). For example, OpenAI's choice of PyTorch over TensorFlow for its AI development was celebrated as a triumph for Meta (Thomas, 2020). As Srnicek (2022, p.255) argues, one should not be fooled by the the semblance of ethical capitalism, as "the seemingly non-capitalist practice of releasing their AI software for free in fact obscures a significant capitalist battle between the[se] companies." These capitalist struggles have been revealed both through leaks and public statements, with a leaked memo in May 2023 highlighting Google's concern about having "no moat around closed-source AI development" and therefore advocating to "own the ecosystem and let open source work for us" (Patel & Ahmad, 2023).

2.1.2.2 The Arrival and Proliferation of Open Models

The year 2021 was a turning point for the OSAI commons, which saw the emergence of the first OMs being shared publicly on the Internet. OMs are characterised by the public release of pre-trained models, specifically the model parameters (i.e. weights and biases), which others can access, download, fine-tune, and use via platforms like HF Hub or APIs (White et al., 2024). This trend began with EleutherAI, a grassroots non-profit research group, which formed on a Discord server

with the intention of developing an "open source" variant to OpenAI's GPT. It did so by releasing The Pile, a library of datasets for training LLMs, in December 2020 and GPT-Neo, its first model, in March 2021. Subsequently, OMs gained more visibility with the development of BLOOM by over 1,000 volunteers in the BigScience project in July 2022, Stable Diffusion by Stability AI in August 2022, and or the leak of Llama by Meta in March 2023, among others (Tarkowski, 2023). Since then, it has become common for industry giants to release open weight FMs and LLMs; for example, 98 (66%) of the 149 FMs released in 2023 were released as OMs (Maslej et al., 2024).

This practice arose following years of debate about whether and the degree to which powerful AI models should be released openly. The choice to openly release AI models has been framed as a balancing act between facilitating research and innovation on the one hand, and mitigating harms on the other hand (Shevlane, 2022). Industry leaders have sought to shape policy discussions by positioning OSAI as either beneficial to innovation and democracy or detrimental to safety (Widder, Whittaker, & West, 2024). For the time, an avant-garde approach was taken by researchers at OpenAI, who in "Release Strategies and the Social Impacts of Language Models" documented OpenAI's staged release strategy of four variants of GPT-2, which varied in size from small (124 million parameters) to large (1.5 billion parameters), to monitor and address concerns about risks of misuse (Solaiman et al., 2019) . This approach diverged from the likes of Google, which demonstrated its non-appetite for critical inspection of the risks posed by its LLMs when it fired the researchers Timnit Gebru and Meg Mitchell over their "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" paper (Bender, Gebru, McMillan-Major, & Shmitchell, 2021). Solaiman (2023) has since proposed a framework for the safe release of generative AI systems (see Figure 2.3), contending that AI systems (N.B. not models) are not either fully open and fully closed; rather, the openness of AI systems can be plotted along a gradient with six degrees of openness, ranging from fully closed to fully open, which each involve trade-offs between concentrating power and mitigating risks (Solaiman, 2023).

The emergence of OMs—both their release and their development in the open—are viewed as a new approach to AI R&D that may challenge the dominance of industry leaders (Ahmed et al., 2023). According to Tarkowski (2023), "the field of AI development, previously dominated by a few companies that developed closed technologies, is now portrayed as one in which closed and open solutions compete." For example, the leaked Google memo warned there is "no moat around closed-source AI development" and "open source solutions will out-compete companies like Google or OpenAI" (Patel & Ahmad, 2023). However, scholars have argued that we should not lose sight of how companies seek to benefit from releasing their AI models openly. Widder et al. (2023) contend that Big Tech companies release OMs as a strategy to entrench their dominance in AI by attracting researchers and developers to adopt, test, and improve their OMs. This argument was evidenced

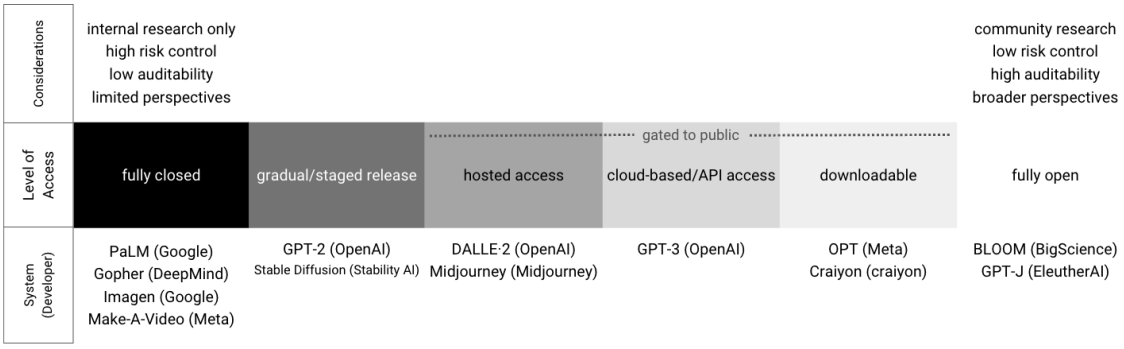


Figure 2.3: The Gradient of AI System Openness and Related Considerations
Source: Solaiman (2023, p.4).

by the aforementioned Google memo, which underscored the need to "own the ecosystem and let open source work for us" (Patel & Ahmad, 2023). Furthermore, the ability to build competitive OMs depends heavily on access to specialised and costly resources concentrated among a few large tech companies, as discussed in Section 2.1.1. Given the resource requirements, commercial AI companies with access to compute power, datasets, and talent increasingly dominate the OM ecosystem. For example, Chapter 6 shows that industry giants are already benefitting from this strategy: much of the activity on HF Hub revolves around researchers and developers building on top of a small number of OMs released by industry giants (Osborne, Ding, & Kirk, 2024).

In contrast to most companies that release OMs, Meta has been outspoken about its OM strategy. Upon releasing Llama 2, Nick Clegg, Meta's President of Global Affairs, explained in the *Financial Times* that, "openness isn't altruism—Meta believes it's in its interest. It leads to better products, faster innovation, and a flourishing market, which benefits us as it does many others" (Clegg, 2023). He argued that releasing Llama 2 would make it "safer and better" because it will benefit from the "wisdom of the crowds." He added that, "A mistaken assumption is that releasing source code or model weights makes systems more vulnerable. On the contrary, external developers and researchers can identify problems that would take teams holed up inside company silos much longer." Meanwhile, Mark Zuckerberg has explained the company's objective is to build an ecosystem around their Llama models. Upon the release of Llama 3, he explained that they are not doing open source:

> *"because we are, like, altruistic... We're doing it because... the reality is that this is an ecosystem... it gets better when you have all the silicon providers optimising all of their stacks for the thing that we're doing, and you have all these other companies and start-ups building different distillation tools or inference tools to make it go faster and more efficient...*

I just want everyone to be using it because the more people who are using it, the more the flywheel will spin for making Llama better" (South Park Commons, 2024).

Zuckerberg has also explained the value of ecosystem-building to Meta shareholders, citing the example of how the widespread use of PyTorch has "been very valuable for us" because it has facilitated the integration of external research and innovations into their systems (Meta, 2023).[2] Beyond industry giants with the resources to invest in ecosystem-building, it remains to be studied why other actors, including start-ups, release OMs. The commercial incentives of various kinds of companies for AI OSS donations presented in Chapter 5 provide a framework for such future research.

An exception to the trend of commercial dominance in the development of OMs are a handful of grassroots initiatives, including EleutherAI and the BigScience workshop, that have embraced open collaboration methods to develop OMs from the ground up (Ding, Akiki, Jernite, Steele, & Popo, 2023). For example, the development of BLOOM, a 176 billion parameter LLM that can generate text in 46 natural languages and 13 programming languages, was the largest collaboration on an OM to date, involving over 1,000 volunteers from over 70 countries and over 250 institutions (Akiki et al., 2022). This project was supported by the French government, which subsidised the use of its supercomputer, Jean-Z, for the training of the BLOOM model. Jernite, an employee of HF and a coordinator of the BigScience workshop, commented: "Much like the tensions between proprietary and OSS in the early 2000s, AI is at a turning point where it can either go in a proprietary direction, where large-scale state-of-the-art models are increasingly developed internally in companies and kept private, or in an open, collaborative, community-oriented direction, marrying the best aspects of open source and open science" (Jernite et al., 2021). Crucially, the BigScience Project coordination of the open development of an open FM illustrates an "alternative pathway for how AI can be developed beyond well-resourced technology companies and who can be a part of the process" (Ding et al., 2023, p.1). This project provides a useful lesson for Public AI strategies.

As OM development has gained momentum, new platforms have emerged to support the growing OSAI developer community, with HF Hub becoming the principal platform for sharing and developing OMs and datasets. HF Hub is a Git-based collaborating coding platform created by the start-up HF, which has emerged as the principal platform used by individuals and companies to share, download, and collaborate on OMs, datasets, and "spaces" (web applications for demonstrating models). In light of the growing popularity of HF Hub, scholars have examined its potential as a data source for empirical studies on OM development (Ait, Izquierdo, & Cabot, 2023b, 2023a; Castaño,

[2] One can draw parallels to the technology giant's prior self-serving "gift-giving" programmes such as Free Basics, which provided free internet access within the walled garden of Facebook and its partner websites to the populations of developing countries. The free but limited internet, effectively, locked users into its ecosystem and turned them into into commodities in its surveillance capitalism machine (Sen et al., 2017).

Martínez-Fernández, Franch, & Bogner, 2024). For instance, Castaño et al. (2024) observe skewed distributions in both commit patterns and contributor counts in model repositories. They find that while most repositories see limited activity and contributions from only a few developers (with a median of 1 contributor per repository), a small number of repositories, such as `bigscience/bloom` and `bigcode/santacoder`, stand out with significant activity and contributions from 22 and 17 developers, respectively. In Chapter 6 (RP3), I build on this work by showing that only a minority of model repositories on HF Hub receive most activity and engagement—e.g. 1% of models account for 99% of downloads—and that a minority of models, mainly ones developed by a few industry leaders, are used in "spaces." These findings substantiate claims that OMs are not delivering the equalising effects that are championed in the AI democratisation narrative (Widder et al., 2023).

The release of AI OSS and OMs are widely described as acts of AI democratisation. However, this term is used ambiguously as a catch-all term. According to Seger, Ovadya, et al. (2023), this phrase is typically used to refer to four kinds of goals: the democratisation of AI use, AI development, AI profits, and AI governance. In most cases, it refers to facilitating the use and development of AI, i.e. accessibility of software or models. Seger, Ovadya, et al. (2023, p.8) conclude that "democratisation" is, therefore, an unfortunate term and, "If by 'AI democratisation' all [one] means is 'make available to everyone', then I would suggest less normatively loaded language (something like 'broad accessibility') be used." Building on Seger, Ovadya, et al. (2023)'s demystification of the term, Chapter 5 investigates commercial incentives driving AI democratisation via AI OSS donations, and discusses the extensibility of its findings to other AI democratisation efforts, such as OM releases.

2.1.3 The Contested Definition of "Open Source AI"

The proliferation of OMs, which are widely promoted as "open source," has sparked debates about what truly constitutes "open source AI." The calling of OMs "open source" has been fiercely contested because, strictly speaking, "open source" refers to the source code of software that anyone can inspect, use, modify, or redistribute (e.g. White et al., 2024; Nolan, 2023). Yet this is not the case for OMs, which encompass various components, such as the training data and software used to train the model, which are not made available (White et al., 2024; Basdevant et al., 2024). The terms "open" and "open source" are used elastically to describe "systems that offer minimal transparency or reusability. . . alongside those that offer maximal transparency, reusability, and extensibility" (Widder et al., 2023, p.2). A review of OMs found that, "While there is a fast-growing list of projects billing themselves as 'open source', many inherit undocumented data of dubious legality, few share the all-important instruction-tuning (a key site where human annotation labour is involved), and careful scientific documentation is exceedingly rare" (Liesenfeld, Lopez, & Dingemanse, 2023, p.1).

Beyond definitional imprecision, Widder, Whittaker, and West (2024) argue that claims about "open" AI often neglect scrutiny of substantial industry concentration in AI development and deployment. In particular, claims about openness frequently focus on only one stage of the development-to-deployment lifecycle, failing to consider the broader context of how AI systems are developed and deployed at scale. Similar to how traditional OSS projects were co-opted by technology giants, they contend that "rhetoric around 'open' AI is frequently wielded in ways that exacerbate rather than reduce concentration of power in the AI sector" (Widder, Whittaker, & West, 2024, p.827).

The description of OMs as "open source" has been criticised as an act of "open washing" (Tarkowski, 2023). For example, Meta received criticism for releasing Llama 2 under a license that allows use of the model for R&D purposes by anyone except for by companies with over 700 million users (i.e. its main competitors) and a list of harmful use cases. For example, Maffulli (2023) argued, "Unfortunately, the tech giant has created the misunderstanding that Llama 2 is 'open source' – it is not. Meta is confusing 'open source' with 'resources available to some users under some conditions,' two very different things." Similarly, Tarkowski (2023) argues these restrictions "support collaborative development only within the bounds of the Llama 2 ecosystem. In this way, Meta is looking for a way to bet on open source – and its advantages concerning research and innovation – while building its own 'moat' against its competitors."

Another objection stems from the licensing of model weights using OSS licenses, such as Apache v.2 or MIT, which grant rights for the use, study, modification, and redistribution of software source code, but are not adapted to model weights, which are high-dimensional data (White et al., 2024). At the time of writing, efforts are underway to develop bespoke licenses for model weights and the Model Openness Framework (see Table 2.2) has been developed as a gradient approach to the openness and completeness of models, requiring the release of various software, data, and documentation components of models per classification level (White et al., 2024). Furthermore, it recommends open licenses for the respective software, data, and documentation components (White et al., 2024).

These efforts to develop more precise definitions for OSAI are crucial. As Widder, Whittaker, and West (2024) note, claims about openness in AI often incorrectly apply understandings imported from OSS to AI systems, which have distinct development processes and resource requirements. For this reason, they avoid the term "open source" altogether when discussing AI systems, opting instead for more precise terminology that specifies which components are open and under what conditions.

Between 2022 and 2024, the Open Source Initiative organised a co-design process with global volunteers to develop a definition of OSAI systems. In October 2024, it released its v1 defintion:

> *An AI system made available under terms and in a way that grant the freedoms to: use the system for any purpose and without having to ask for permission; study how the system*

MOF Class	Components Included
Class III. Open Model	1. Model Architecture 2. Model Parameters (Final Checkpoints) 3. Technical Report or Research Paper 4. Evaluation Results 5. Model Card 6. Data Card 7. Sample Model Outputs (Optional)
Class II. Open Tooling	1. All Class III Components 2. Training, Validation, and Testing Code 3. Inference Code 4. Evaluation Code 5. Evaluation Data 6. Supporting Libraries & Tools
Class I. Open Science	1. All Class II Components 2. Research Paper 3. Datasets 4. Data Preprocessing Code 5. Model Parameters (Intermediate Checkpoints) 6. Model Metadata (Optional)

Table 2.2: The Model Openness Framework: 3 Tiers of Model Openness and Completeness
Source: White et al. (2024).

> *works and inspect its components; modify the system for any purpose, including to change its output; and share the system for others to use with or without modifications, for any purpose (OSI, 2024c).*

The definition specifies which components of AI models must be released, drawing on the components identified in the Model Openness Framework (see Table 2.2). As a volunteer in this co-design process (OSI, 2024a) and as an author of the Model Openness Framework (White et al., 2024), I witnessed first-hand how contested the discussion is over which components ought to be released. For example, the v1 definition made a compromise on the release of training datasets, only requiring sufficiently detailed information about the datasets. This compromise has received fierce criticism from various sides, from safety experts who call for maximal openness to enable thorough auditing of models to industry actors who would prefer a more lenient definition that does not require them to release their "secret sauce." It looks like the v1 definition will undergo further debate before it gains traction. However, it is important to remember that this debate is not a matter of philosophical differences: it is important to not bend the open source principles in favour of commercial interests (Liesenfeld & Dingemanse, 2024); and regulations for AI, particularly the conditional exceptions for AI systems released under free and open source licenses in the EU's AI Act (EU, 2024), underscore the importance of standardising a definition of OSAI sooner rather than later.

2.1.4 The Benefits and Risks of OSAI

The proliferation of OMs has ignited heated debate about their benefits and risks (e.g. Kapoor et al., 2024; Law & Krier, 2023; Seger, Dreksler, et al., 2023; Srikumar, Chmielinski, & Chang, 2024). On the one hand, proponents argue that OMs present three key affordances for research, innovation, and competition: transparency through published weights and documentation, reusability through open licensing, and extensibility that enables building upon existing models through fine-tuning (Widder, Whittaker, & West, 2024; Cihon, 2024; Acquisti et al., 2024). In addition, OMs enable distributed testing (e.g. via auditing) and further development (e.g. via finetuning) of such models, which can result in improvements to their performance and safety (Wladawsky-Berger, 2023). More specifically, OMs lower the barriers for adaptability and customisation for diverse domains and languages (Pipatanakul et al., 2023; Kapoor et al., 2024), as well as applications in science (Yang et al., 2023; Kirchenbauer et al., 2023) and software development (e.g. Peng, Kalliamvakou, Cihon, & Demirer, 2023; M. Hoffmann, Boysel, Nagle, Peng, & Xu, 2024). On the other hand, scholars have warned that OMs pose many risks by enabling the creation of deepfakes (Nguyen et al., 2022; Thiel, Stroebel, & Portnoff, 2023), disinformation (Goldstein et al., 2023; Musser, 2023), and malware (Tsamados, Floridi, & Taddeo, 2023; C. David & Paul, 2023), among others.

Due to the applicability of FMs in a wide range of contexts, the "marginal risks" of open FMs—that is, risks of open weight FMs relative to closed weight FMs—have received particular attention (e.g. Kapoor et al., 2024; NTIA, 2024). Open weight FMs are considered to have five distinctive properties that present both benefits and risks: broader access, greater customisability, local adaptation and inference ability, the inability to rescind model access, and the inability to monitor or moderate model usage (Kapoor et al., 2024). These simultaneous benefits and risks were exemplified by the Alpaca 7B model, developed by researchers at Stanford University. On the one hand, an argument in favour of OMs is that since AI models can be expensive to train, OMs lower entry barriers and widen access to cutting-edge AI technologies. For example, with access to Meta's Llama model, these researchers were able to develop Alpaca 7B for less than $600, which demonstrated similar performance to OpenAI's GPT-3.5 (Taori et al., 2023). On the other hand, the dual-use capabilities of models increase the potential for harmful use, even by well-intended actors. Case in point: the Stanford researchers had to take down the demo of their Alpaca model after one week because of model hallucinations, which they attributed to "limitations associated with both the underlying language model and the instruction tuning data," as well as the hosting costs (Taori et al., 2023).

To counter concerns about risks, proponents make the argument that in the spirit of *Linus' Law*—i.e. that "Given enough eyeballs, all bugs are shallow" (Raymond, 2001a, p.30)—the open source

development model offers security advantages over proprietary development (Wladawsky-Berger, 2023). For example, when Meta released its Llama 2 models, it argued that the open weights would enable the "wisdom of the crowds" to improve the safety of the models (Clegg, 2023). However, this argument is constrained by the limited openness of such models, where key components (e.g. see Table 2.2) are not publicly released, making it difficult to study and audit OMs (White et al., 2024). Striking a balance between the benefits and risks of OSAI remains a critical challenge (Bdeir & François, 2024). For now, the National Telecommunications and Information Administration in the USA concludes that, "The current evidence base of the marginal risks and benefits of open weight FMs is not sufficient either to definitively conclude that restrictions on such OMs are warranted, or that restrictions will never be appropriate in the future" (NTIA, 2024).

This first section of the overarching literature review has provided an overview of the political economy of AI, with a focus on the supply chain capitalism of AI and the emergence of OSAI development. In the next section, I discuss prior work on the political economy of OSS, providing a background on the history of commercial participation in OSS development and theoretical frameworks that can inform the nascent research agenda on the political economy of OSAI.

2.2 From a "Programmers' Paradise" to a "Commons of Capital": A History of OSS

2.2.1 Humble Beginnings: The Emergence of a "Programmers' Paradise"

2.2.1.1 Origins of the OSS Movement

The origins of OSS are generally associated with two key figures: Richard Stallman and Linus Torvalds. In the early 1980s, Stallman, a computer scientist at MIT, was disgruntled with the proprietary nature of the *Unix* operating system, whose intellectual property was owned by the telecommunications giant AT&T and which he could not use beyond pre-authorised purposes (Birkinbine, 2020). In protest, in 1983, Stallman began working on his own operating system, *GNU*, which he released in 1985 along with the *GNU Manifesto* and the *GNU Public License* (Birkinbine, 2020). In doing so, more than protest against AT&T, Stallman began a political movement which became known as Free, Libre, Open Source Software (FLOSS), where the "free" and "libre" signify the political freedom of software; that is, "Free as in freedom, not as in free beer" (Stallman, 2002).

This moment is often seen as the beginning of the digital commons, including open knowledge initiatives such as Wikipedia, OpenStreetMap, and the Creative Commons, whose collective cause is the fight against the privatisation of information goods (Broca, 2013, 2021; Coleman, 2012). Draw-

ing on Ostrom's (1990) formulation of commons as resources or resource systems that are shared and governed by a group of people, digital commons are digital resources, such as OSS and open data, and their corresponding resource systems that are developed and governed by an online community. "Unlike tangible commons (such as urban gardens, forests or meadows), the digital commons (such as free software or Wikipedia) are not affected by overuse or material exclusivity. However, their existence can still be threatened by undersupply, inadequate legal frameworks, pollution, lack of quality or findability" (Dulong de Rosnay & Stalder, 2020, p.2).

Torvalds entered the scene in 1991. He was a 21-year-old student at the University of Helsinki, who set out to build an open source operating system for his personal computer, inspired by the proprietary Unix operating system (Birkinbine, 2020). Torvalds announced his plans in the online newsgroup Usenet in August 1991 and he released version 0.01 of this operating system, which he called Linux, the following month. Contrary to Stallman's explicitly political project, Torvalds's rationale focused on the process of software development; that is, he believed distributed development would reduce the workload of developers and lead to higher quality software (Birkinbine, 2020). His hunch turned out to be right: the project quickly attracted developers from across the world, who shared an interest in building an open source operating system. Raymond later coined *Linus' Law*—"Given enough eyeballs, all bugs are shallow" (Raymond, 2001a, p.30)—to characterise the benefits of this model of software development. This vision did not rule out commercial contributions to or commercial derivatives of Linux, leading to a bifurcation from the FLOSS camp (Birkinbine, 2020).

In the late 1990s to early 2000s, the emergent open source philosophy and mode of software development was celebrated as a "programmers' paradise," comprising geeks, hackers, and hobbyists, who contributed to this emerging digital commons due to intrinsic motivations, such as their political ideals (Kelty, 2008), their altruism and desire to share their "tinkerings" (Markus, Manville, & Agres, 2000), or their passion for working with peers on technical problems (Loebbecke & Angehrn, 2003). In the *Cathedral and Bazaar*, Raymond (2001a) juxtaposed the bottom-up and top-down organisational structures of the Linux community and companies like Microsoft: the former was a bustling "bazaar" vibrating with the activity of hackers, while the latter was a "cathedral" that employed a hierarchical structure for profit-driven production. The openness of OSS allowed individuals to perform a myriad of activities simply based on their skill and interest, from writing code and documentation to bug-spotting to "serving the hacker culture itself" (Raymond, 2001a). It was a "hackerdom," where "anyone is welcome–the more people, the louder the clamor, the better it is" (Kuwabara, 2000). This process, Raymond (2001a, p.52) argued, "produces a self-correcting spontaneous order more elaborate and efficient than any amount of central planning could have achieved."

2.2.1.2 Theorising Commons-Based Peer Production

In both *Coase's Penguin* (2002) and *The Wealth of Networks* (2006), Yochai Benkler formalised Raymond's observations of the social dynamics of the Linux community into a theoretical framework. In particular, Benkler (2006, p.60) developed the concept of "commons-based peer production," which he defined as a "radically decentralized, collaborative, and non-proprietary process of commons-based peer production." This model contrasts with the theory of the firm (Coase, 1995; Holmstrom & Tirole, 1989), which posits that companies exist to minimise transaction costs that arise in market exchanges through contractual designs. Benkler contended that commons-based peer production was more efficient than classical production and more aligned with values such as freedom and individual autonomy because it relies on voluntary contributions and collaboration among individuals who "cooperate with each other without relying on either market signals or managerial commands" (Benkler, 2006). As such, commons-based peer production offers an alternative to market-based production, but Benkler also noted that these spheres are not mutually exclusive, with peer production being able to influence market dynamics (Benkler, 2006).

While Benkler's theoretical framework has been instrumental in conceptualising OSS development as a commons-based peer production model, it appears idealistic in retrospect. It overlooks the necessary conditions for the social reproduction of commons-based peer production communities (Broca, 2021) and underestimates the adaptability of digital capitalism to integrate peer production within its structures (Birkinbine, 2020). The next sections delve into prior work that critique and problematise these early theories: first, I discuss literature on social hierarchies and digital divisions of labour that problematise the image of a "programmers' paradise"; then, I discuss literature on commercial participation in OSS development and its implications for the political economy of OSS.

2.2.2 Problematising the "Programmers' Paradise" Image of OSS

The early characterisations of OSS must be understood in their time and place. Many commentators sincerely hoped that the advent of the Internet would lead to more equitable knowledge economy (Lessig, 2003). However, social hierarchies and divisions of labour have formed in the online communities that have sought to "democratise" knowledge (e.g. Brooke, 2021; Graham, 2014).

The first aspect to problematise is the image of one coherent OSS community. Contrary to Raymond's organisational metaphors, there is not one OSS community and they are not all organised like bazaars. Most OSS projects are developed by a singular developer, which are therefore more akin to "caves" (Krishnamurthy, 2005b). Those that do have communities have different structures. Many projects are structured like "layered onions" with graduated layers of activity, ranging from

a dense core maintainers to one-off contributions from users (Crowston & Howison, 2005). The presence of power developers is common on social coding platforms, such as the Q&A forum Stack Overflow, where only 5% of users have answered 60% of questions and only 19.2% of users have ever answered a question (Parnin, Treude, Grammel, & Storey, 2012; Ford, Harkins, & Parnin, 2017).

Beyond those that contribute code to OSS projects, a social model of OSS projects encompasses the core maintainers, contributors, and users (Ferraioli, 2022). For example, Eghbal (2020) suggests four social structures OSS projects based on contributor-to-user ratios: federations (high contributor and user growth), clubs (high contributor but lower user growth), stadiums (high user but low contributor growth), and toys (personal projects with low growth). Furthermore, "all contributors" frameworks recognise both technical and non-technical contributors (Young et al., 2021).

What is more, contrary to the image of a welcoming paradise, many OSS developer communities are infamous for their unfriendly cultures towards developers who do not fit the "hacker" profile (Vasilescu, Capiluppi, & Serebrenik, 2014; Ford et al., 2017). A survey of 5,500 GitHub users paints a dismal picture of the OSS developer population: 95% of respondents identified as men, compared to only 3% who identified as women and 1% who identified as non-binary, and only 16% identified as ethnic or national minorities in their country (GitHub, 2017). The population of GitHub users is even less diverse than the software industry in the USA, where 22.6% of software engineers identify as women and around 34% identify as an ethnic minority (Finley, 2017; Gruman, 2020).

Moreover, OSS developer communities tend to be hyper-masculine spaces where women experience greater levels of hostility (Vasilescu et al., 2014; Terrell et al., 2017; Dunbar-Hester, 2019) and receive lower recognition for their contributions compared to their male peers (Terrell et al., 2017; Brooke, 2021). This hostility towards women is part of a broader "perception of a geek mythology culture that promotes expectations of male success and continual questioning of women's abilities" in the software industry (Margolis & Fisher, 2001; Ashcraft, McLain, & Eger, 2016). As Brooke (2019, p.177) contends with reference to Stack Overflow, "with an estimated population of only 6-11% women, the popular platform is only paradise for some."

Another aspect to problematise is the tendency to describe open source activity as "contributions" and "collaborations," rather than explicitly recognising them as work or labour. Yet, as OSS development concerns the creation and maintenance of information goods, it is digital labour (Birkinbine, 2015). In fact, many kinds of labour go into the development and maintenance of OSS, including writing code and documentation, bug-spotting, and maintenance. It is important to note that many types of labour in OSS projects are not digital, not technical, and not always visible (Hossain, 2021; Geiger, Howard, & Irani, 2021; Eghbal, 2020), including the range of tasks that maintainers do to ensure the long-term sustainability of their project, from community outreach to mentoring new-

comers to hosting hackathons (Eghbal, 2020). In fact, a significant portion of OSS development does not actually take place in the code repository, including social tasks like community animation which are critical to the social reproduction of OSS communities.

An ethnographic study of the *Mozilla* and *Brave* communities in Dakar, Bangladesh illustrates the role of social and often unpaid labour in the social reproduction of these communities. The study shows that these communities in large part are made up of students, who perform localisation tasks like documentation translation and brand ambassadorship tasks, such as organising workshops on their university campus (Hossain, 2021). This labour would go unnoticed if one only looked at the code itself. In these communities, the contributors are primarily university students who are eager to socialise with like-minded peers, build professional networks, and gain leadership and OSS development experience which can help them get jobs after their university degrees (Hossain, 2021). Subramanyam and Xia (2008) similarly find that while OSS developers in the USA show a relatively strong interest in "geek culture," developers in India and China tend to be motivated primarily by the professional development opportunities that are made possible through OSS experience. These findings underline that "researchers studying open source should be mindful of geographic variation in what motivates participation and what forms participation may take, particularly outside of the code repository" (Hossain, 2021, p.30).

Digital divisions of labour that follow geographic lines characterise participation in OSS development (Graham, 2014; Lehdonvirta, Kässi, Hjorth, Barnard, & Graham, 2019). The largest proportion of core developers on GitHub are based in North America and Western Europe, while the largest proportion of bug spotters are in the Global South (Takhteyev & Hilts, 2010). Similarly, activity on Stack Overflow is concentrated in metropolitan areas in the Global North (Braesemann, Stoehr, & Graham, 2019). This has led Takhteyev (2012) to argue that OSS's seeming independence from geography yet the centralisation of its production in both a few places and companies exemplify the paradoxes of globalisation. Rather than undoing the dominance of a few places and companies, digital divisions of labour are reproduced and reinforced on online platforms (Graham, 2014, 2019). What is more, his study of the developers of the Lua programming language in Rio de Janiero, Brazil illustrates that "participation in open source projects involves a complex negotiation of culture, language, and geography" for those who participate from the "wrong places" that far removed from English-speaking Global North centres of innovation, primarily Silicon Valley (Takhteyev, 2012, p.9).

The social hierarchies and digital divisions of labour in OSS developer communities complicate the early characterisation of OSS as a "programmers' paradise." The next section explores how OSS has shifted from a non-market, commons-based production model to one increasingly intertwined with commercial logics and structures of digital capitalism.

2.2.3 From a Digital Commons to a "Commons of Capital"?

2.2.3.1 Commercial Interests and Involvement in Open Source

The history of OSS has not been favourable to Benkler's (2006) theory of OSS as a commons-based peer production model that does not follow market signals or managerial commands. Taking the header title from an essay by Calimaq (2018) that asks if the digital commons are condemned to become a "commons of capital," this section reviews prior work on commercial participation in OSS development and introduces Birkinbine's (2020, p.3) framework for understanding how OSS has become "dialectically situated between capital and the commons"; that is, how OSS is continually shaped by the interactions and tensions between these opposing value systems.

OSS has evolved significantly since its humble beginnings in hacker communities and academic labs (Cao & Chesbrough, 2022). Since the late 1990s, it has been common for companies to employ staff to participate in OSS projects (Dahlander & Magnusson, 2005; Dahlander & Wallin, 2006) and even for companies, which may be market rivals, to collaborate on OSS development (e.g. Germonprez, Allen, Warner, Hill, & McClements, 2013; Nguyen Duc, Cruzes, Terje, & Abrahamsson, 2019; Teixeira & Lin, 2014), which has turned many OSS developer communities "from networks of individuals into networks of companies" (Ågerfalk & Fitzgerald, 2008, p.396). Furthermore, companies have adopted open-core business models, where their core software product is open source on top of which they offer commercial services and products (Streicher, 2020).

The example of Microsoft illustrates the shift in commercial posturing towards OSS better than no other (Birkinbine, 2020). Around 2000, it dominated the operating system market through its Windows suite. It responded to the rise of Linux as an open source alternative with a smear campaign, calling it "un-American" (Raymond, 2001c). In 2001, Steve Ballmer, its CEO, denounced Linux as "a cancer that attaches itself in an intellectual property sense to everything it touches" (Warren, 2020). Yet, in 2018, Microsoft bought GitHub, the most popular platform for hosting OSS used by over 90 million developers, for $7.5 billion (Microsoft, 2018) and is now the most prolific corporate contributor to OSS (Warren, 2020). This change in posture towards OSS is illustrative of the recognition that the OSS development model is an efficient and cheap model for developing high-quality software (Birkinbine, 2020), and of the strategic value of owning the infrastructure upon which OSS is hosted, accessed, and developed (Srnicek, 2017). Beyond this example, industry has emerged as the dominant contributor to OSS development through both in terms of code contributions and financial investment (Warren, 2020). According to estimates, companies in the EU invested €1 billion in OSS in 2018, which in turn generated €95 billion in GDP value (Blind et al., 2021), while companies in the USA invested $37.8 billion in 2019 (Korkmaz, Santiago Calderón, Kramer, Guci, & Robbins,

2024). While these amounts may seem high, a significant disparity exists between the supply value ($4.15 billion) and demand value ($8.8 trillion) of OSS (M. Hoffmann, Nagle, & Zhou, 2024), as well as in the supply of maintenance labour and usage for OSS (Champion & Hill, 2021). As I discuss in more detail below, this imbalance animates a heated debate about the sustainability of OSS.

An extensive literature discusses how and why companies participate in OSS development. While somewhat dated now, Bonaccorsi and Rossi (2006) offer a useful taxonomy of economic, social, and technological incentives of companies and how they compare to individuals. They show that while individuals are driven by personal interests and needs (von Krogh, Haefliger, Spaeth, & Wallin, 2012), companies are above all driven by economic and technological incentives (Bonaccorsi & Rossi, 2003). The OSS development model is associated with lower in-house development and maintenance costs, production efficiency, as well as high-quality software (Birkinbine, 2020; Lindman, Juutilainen, & Rossi, 2009). The commercial embrace of OSS collaboration has been understood as an extension of outsourcing, where companies externalise internal R&D processes to external contributors (Tapscott, 2011; Ågerfalk & Fitzgerald, 2008). Participation in an OSS project can help a company improve its market position by undercutting the product of a rival company (Fink, 2003). A common technological incentive for companies is to influence open standards, especially when they seek to resist standards developing around the technology of a rival (Lerner & Tirole, 2001). Meanwhile social incentives include the improvement of corporate image by publicly conforming to OSS values (Bonaccorsi & Rossi, 2006) and using this branding to hire OSS developers for whom these values are important (Lindman et al., 2009; Ågerfalk & Fitzgerald, 2008). Tables 5.1-5.2 in Chapter 5 (RP2) provide an overview of prior work on individual and commercial incentives.

Incentives vary between company and per type of engagement with OSS. Broadly speaking, companies engage with OSS development through three main models: supporting independently hosted projects through contributions or funding, collaborating with other organisations to share control, and hosting projects under single-company control (M. Zhou, Mockus, Ma, Zhang, & Mei, 2016). The most common is the supporting model. Companies commonly support OSS by funding developers' work hours for project contributions—whether as formal responsibilities or voluntary programmes (Dahlander & Wallin, 2006; Xia, Wang, Zhao, Bian, & Wang, 2023)—and by paying membership fees to foundations that host OSS projects (O'Mahony & Bechky, 2008).

Company-hosted OSS projects, also called "company-managed projects" (O'Mahony & Ferraro, 2007), are becoming increasingly common. Companies often spin out proprietary software as company-hosted OSS to increase adoption, gain external contributions, or reduce competitors' market share (West & O'Mahony, 2005). However, commercial dominance—i.e. the presence of a company that "makes the majority of contributions and thus has a strong influence on the development roadmap of

the project" (Y. Zhang, Stol, Liu, & Zhou, 2022, p.1049)—can deter external contributors, especially volunteers, who prefer open governance and do not want to perform free labour for the dominant company (M. Zhou et al., 2016; Y. Zhang, Tan, Zhou, & Jin, 2018). The collaborating model is the least common, but examples from the AI industry include the joint release of the Open Neural Network Exchange (ONNX) by Meta and Microsoft in 2017. At this point in time, there were many DL frameworks and ONNX facilitated interoperability between them (Candela, 2017).

Across all three models, it has become common for companies to collaborate with other companies, including market rivals or companies they are engaged in patent wars with (Teixeira & Lin, 2014). The term "open source co-opetition" has been coined to convey this tangle of cooperation and competition between companies (Teixeira, 2014). Drawing on management science (Brandenburger & Nalebuff, 1997), co-opetition explains how competitors form strategic alliances in areas far from their customers like R&D while competing on products (Bengtsson & Kock, 2000). These collaborations aim to access resources (Stuart, 2000), facilitate learning (Powell, Koput, & Smith-Doerr, 1996), and shape industry standards (Gnyawali & Park, 2011). The ecosystem concept frames these interdependencies between companies (Jansen et al., 2009), which are particularly common in high-technology sectors due to rapid innovation cycles and rising R&D costs (Gnyawali & Park, 2009). Ågerfalk and Fitzgerald (2008, p.396) note that many OSS developer communities have evolved "from networks of individuals to networks of companies." Chapter 4 examines this phenomenon in detail through a study of open source co-opetition in three company-hosted AI OSS projects: Meta's PyTorch, Googe's TensorFlow, and HF's Transformers.

2.2.3.2 Between Capital and the Commons: The Present and Future of Open Source?

Widespread commercial participation has fundamentally altered the political economy of OSS. According to Birkinbine (2020, p.3), "OSS is [now] dialectically situated between capital and the commons." On the one hand, it originated as a social movement that opposed the privatisation of software and it strives to sustain OSS as communal software goods that can be accessed, used, modified, and shared freely. On the other hand, recognising the affordances of the peer production model for their commercial aims, companies began to sponsor, use, and contribute to OSS, blurring the lines between the commons-based versus profit-driven model of software development. Rather than standing in opposition to market-driven production as suggested by Benkler (2006), OSS has become the "roads and the bridges" of the digital capitalism (Eghbal, 2016). In addition, open source has become a business model for companies to assert their market dominance, from investing in OSS to challenge the proprietary products of market rivals to releasing open source frameworks whose adoption enables the integration of external products into proprietary systems (Widder, Whittaker,

Strategy	Example
Invest in open source to challenge market competitors	**IBM and Linux:** In 1999, IBM invested $1 billion in Linux, operating software positioned as an open source alternative to the then-dominant Microsoft.
Release OSS to control a platform	**Google and Android:** In 2007, Google open-sourced and heavily invested in Android OS, allowing them to achieve mobile operating prominence over competitor Apple and attracting scrutiny from regulators for anticompetitive practices.
Re-implement and sell OSS as Software as a Service (SaaS)	**Amazon and MongoDB:** In 2019, Amazon implemented its own version of the popular open-source database MongoDB, known as DocumentDB, and sold it as a service on its AWS platform. In 2022, it transitioned to a revenue-sharing agreement with MongoDB.
Develop framework that enables the integration of external OSS products into company's systems	**Meta and PyTorch:** Meta CEO Mark Zuckerberg has described how open-sourcing the PyTorch framework has made it easier to capitalise on new ideas developed externally and for free.

Table 2.3: Methods of Asserting Market Dominance Through—Not In Spite of—OSS
Source: Widder, Whittaker, and West (2024, p.828)

& West, 2024). See historical examples of commercial strategies in Table 2.3.

However, Broca (2021) remarks that the last 30 years are not simply a story of digital capitalism subsuming the digital commons; instead, there has been "a game of reciprocal influences" between digital capitalism and the digital commons, where the former has adjusted to the norms and practices of the latter (Broca, 2021). Perhaps the best illustration thereof is the fact that the global digital economy could not run without OSS, which is is estimated to feature in 96% of proprietary codebases (Synopsys, 2023) and constitute up to 90% of proprietary software stacks (OpenSSF, 2022).

Recognising the way in which OSS is situated between these two opposing circuits of value, Birkinbine (2020) proposes a critical framework to interrogate the social relations, particularly the power relations, that mutually constitute the development, distribution, and use of OSS. This framework is rooted in the critical political economy of communications tradition, which focuses on the material operation of a sector and how products move through a value chain, from producers to distributors to consumers (Mosco, 2009). In this tradition, scholars focus on "uncover[ing] connections between ownership, corporate structure, finance capital, and market structures to show how economics affects technologies, politics, cultures, and information" (Meehan, Mosco, & Wasko, 1993, p.347). As Birkinbine (2020, p.29) notes, "most often, the inquiries of critical political economists of communication are directed at large corporations that hold extensive market power and the ability to influence the production, distribution, exhibition of, or access to, communication resources." In this tradition, power is understood both as a preventative and potential force; that is, "power manifests itself not just as a resource to achieve goals, but also as a form of control that is embedded within a broader set of social relations", and "power relations are present throughout the social system; they structure relationships and tend to reproduce those structures over time" (Birkinbine, 2020, p.34).

This lens is complementary to the supply chain capitalism of AI for understanding the political economy of OSAI, recognising power that may be wielded by actors within the OSAI commons as well as in both up- and downstream stages of the AI supply chain.

Through historical case studies of how companies have engaged with OSS developer communities, including Microsoft, RedHat, and Oracle, Birkinbine contends that for companies the core value of OSS stems from the collective labour power of OSS developer communities that expand the labour force that can develop and test software, which in turn results in cheaper and more efficient software production processes (Birkinbine, 2020). In short, it is the not the *products* but the *processes* of OSS development that companies capitalise on (Birkinbine, 2018). Birkinbine (2020, pp.24-25) describes this as the "incorporation of the digital commons" rather than its "enclosure" because it more accurately reflects the nuanced relationship between companies and OSS developer communities. While "enclosure" implies restrictions or barriers to access, companies have developed ways to commercialise OSS products and processes without necessarily limiting community access to shared resources. Birkinbine contends that "incorporation" better captures this dynamic, suggesting the inclusion of OSS elements into corporate structures without implying exclusion. It also alludes to the legal process of establishing a corporation. This terminological choice, thus, allows for a more precise description of how companies extract value from OSS projects, whilst keeping them open.

Commercial involvement in OSS projects has changed the nature of voluntary OSS development. First, the co-presence of paid and unpaid developers means that volunteers cohabit with conventional wage labour in OSS projects (Schoonmaker, 2018; O'Neil, Muselli, Raissi, & Zacchiroli, 2021). Second, the meaning of volunteer labour has been distorted because it is no longer just a contribution to a cause but also a free input into a production process which end users free-ride on (Broca, 2021). Third, commercial participation has resulted in the industrialisation of OSS development—i.e. its transformation into a "commons of capital" (Calimaq, 2018)—and the acceptance that "everybody knows there are business reasons why people are there" (Ågerfalk & Fitzgerald, 2008, p.395). It has also resulted in prejudices among volunteers against paid developers, who do "do boring work", "rarely care [for] documentation", and "lack personal attachment" to the project (Y. Zhang, Qin, Stol, Zhou, & Liu, 2024). The hybridisation of paid and unpaid labour "raises the question of the fair distribution of the profits firms derive from OSS" (O'Neil et al., 2021, p.1) and the unsustainability of commercial free-riding on the labour of OSS developers (H. Li et al., 2022; Champion & Hill, 2021).

This problem animates a debate about the sustainability of OSS and the social reproduction of the digital commons as a value system (Birkinbine, 2020). It underscores the need for a step-change in our understanding of the roles and responsibilities of OSS users, especially large private and public sector organisations, in supporting the developers of OSS that they use and depend on. Furthermore,

in light of the recognition of OSS as digital public goods, it is timely that governments move from rhetoric to action and unlock public funds to support both open source innovation and maintenance (Osborne, Sharratt, Foster, & Boehm, 2024). Fundamentally, it highlights the need "not just [for] investment in institutions, organisations, technologies, or innovations, but long-term and sustainable investment in the true source of their value: people" (Birkinbine, 2020, p.119). Chapter 7 (RP4) responds to this call through an examination of the role of public and private funding in supporting a community-led OSS project in the industry-dominated OSAI ecosystem.

It should be noted that two value circuits are not always in conflict (Broca, 2021); for example, commercial sponsorship provides invaluable funding for OSS developers (Varoquaux, 2021) and often times volunteers join OSS projects precisely because companies are active in them (Santos, Kuk, Kon, & Pearson, 2013). Furthermore, OSS developers have designed mechanisms to "negotiate their relationship with capitalist firms and, when necessary, defend their commons-based resources from unwanted influence" (Birkinbine, 2020, p.34). One strategy has been to establish "boundary organisations" as protective mechanisms which define the governance, membership, ownership, and control of production of or in OSS projects (O'Mahony & Bechky, 2008, see Figure 2.4). In essence, such organisations delineate the limits of cooperation between stakeholders with different or competing interests ("unexpected allies"), including market rivals, and shield contributors from unwanted external interference, including from a dominant company that may seek to influence the roadmap of an OSS project. Although boundary organisations seek to prevent commercial dominance in OSS projects, their effectiveness in doing so varies in practice (Wagstrom, 2009; Y. Zhang et al., 2022); for example, 10% of companies account for 80% of commits to OSS projects in the OpenStack ecosystem (Y. Zhang, Zhou, Mockus, & Jin, 2021).

In summary, funding and open governance, among others, are critical levers for balancing diverse interests in OSS developer communities, supporting development and maintenance, and ultimately sustaining them over time. These lessons are particularly salient for emerging discussions about pathways toward public interest OSAI. The experiences of OSS communities in negotiating their relationship with commercial interests provide valuable insights for building sustainable, community-governed OSAI commons that serve the public good rather than solely commercial interests.

2.2.4 Bridging Streams: Toward a political economy of OSAI

The prior work discussed in this literature review provide complementary lenses for the four RPs that examine commercial interests and involvement in OSAI development, and the implications thereof for norms, practices, and the potential trajectories of the OSAI ecosystem.

First, the supply chain capitalism of AI framework (Valdivia, 2024) maps the material dependen-

Role of a Boundary Organization in Enabling Collaboration

Organizing Practices Adapted	Interests Satisfied: Community-managed open-source software projects	Interests Satisfied: Firms
Governance		
Establishing project representation	Provides open access and participatory processes	Reduces ambiguity and provides some degree of discretion
Pluralistic control	Ensures independent and collective control without undue firm influence	Provides some voice on project direction without direct control
Membership		
Defining rights of members	Preserves individual basis of membership and independence of the community	Firms cannot gain formal rights, only sponsor contributors
Sponsoring contributors	Provides additional resources to help project improve	Offers firms a means of direct access to development process
Ownership		
Obtaining work assignment rights	Reinforces individual autonomy and independence	Ensures clear provenance of code
Developing contribution agreements	Ensures clear provenance of code	Ensures clear provenance of code
Managing code donation	Enhances technical quality and reach of the project	Improves efficiency: no separate code base to manage
Control of production		
Community control of code contribution	Allows community to preserve autonomy and independence	Sponsored contributors provide firms with visibility and access to code development
Managing technical direction	Allows community to preserve autonomy and independence	Sponsored contributors provide firms with informal influence on code development

Figure 2.4: The Role of Boundary Organisations in Enabling OSS Collaboration
Source: O'Mahony and Bechky (2008, p.441)

cies of AI development and provides a framework for organising prior work on power structures that characterise the AI industry, revealing how monopolies and oligopolies at various stages of the supply chain influence the development, distribution, and consumption of OSAI technologies. This holistic perspective enables us to challenge narratives about AI democratisation, highlighting that the mere availability of OSAI technologies does not fundamentally alter industry concentrations across the supply chain. Furthermore, it draws our attention to the environmental and human rights impacts throughout the supply chain of AI development, emphasising that praxis for public interest OSAI must recognise that OSAI technologies are not solely digital resources that are developed by online communities but ones that have material consequences that cannot be neglected. Then, the critical political economy of the digital commons (Birkinbine, 2020) provides insights into how commons-based peer production like OSS development has become entangled with commercial interests and digital capitalism. This framework illuminates how companies seek to harness the collective labour power of OSS developer communities, incorporating not only the *products* but more importantly the *processes* of OSS development. It also provides insights into how OSS developer communities negotiate their boundaries with companies as well as the roles of open governance and funding for

guarding the independence and long-term sustainability of community-led OSS projects.

Together, these frameworks inform a critical examination of both the promise and limitations of OSAI as a democratising or equalising force in AI development. They suggest that while OSAI may offer benefits through enabling transparency, reusability, and extensibility, its promise for democratising AI R&D is constrained by the power structures of the wider AI supply chain and that it may also provide an avenue for industry leaders to strengthen their market positions, such as by leveraging the collective labour power of global OSAI researchers and developers. A complementary theoretical lens has been developed by Widder, Whittaker, and West (2024), who scrutinise how open source may be used to entrench, rather than challenge, concentrations in AI R&D, drawing on lessons from the history of commercial interests and involvement in OSS development. Ultimately, this theoretical synthesis suggests that advancing public interest OSAI requires not only supporting OSAI development but also addressing structural conditions and material impacts across the wider AI supply chain.

3. Research Design

In this chapter, I discuss ex-ante considerations and ex-post reflections regarding my research design. While each RP discusses its respective research design choices and limitations, this chapter offers an opportunity to elaborate on the research design of the RPs in more detail and provide a reflexive account of my journey as a doctoral researcher. Some of the considerations and reflections have cross-cutting relevance to all RPs, while others are specific to the respective RPs.

3.1 Overarching Reflections on Research Design

This thesis seeks to advance the nascent research agenda on the political economy of OSAI by empirically examining why and how companies participate in the development of OSAI technologies, and the implications thereof for development practices, governance norms, and the potential trajectories of the OSAI ecosystem. As discussed in Chapters 1-2, the theoretical lens to this MRQ is rooted in political economy, with a focus on the social relations, in particular the power relations, that mutually constitute the development, distribution, and consumption of OSAI technologies. This lens informs the research questions (RQ) that I focus on in the respective RPs, from how dominant corporations like Meta and Google wield power in the OSAI ecosystem through their unilateral control of widely used AI OSS projects (RP1) to the avenues for community-led OSAI projects to provide credible alternatives to corporate offerings and sustain themselves in the long term (RP4). I approach this MRQ and the RQs of the respective RPs empirically by employing research principles and designs from the disciplines of social data science and software engineering research, which I discuss below.

3.1.1 Beginning with the PODIKW Framework from Social Data Science

As a doctoral researcher in the social data science programme at the Oxford Internet Institute, my starting point was to familiarise myself with core principles and research designs in the discipline of social data science. A key textbook in this discipline is "From Social Science to Data Science" by Hogan (2022), who formulates social data science as an emerging discipline, which still lack core tenets. Hogan (2022, p.6) suggests that "social data science is about the measurement of peo-

ple and their behaviour. This means that what is most relevant is considering whether what we measure links to what we seek to measure." Hogan proposes a framework for social data science called PODIKW (phenomena-operationalisation-data-information-knowledge-wisdom), adding P for phenomena and O for operationalision to the DIKW framework from the field of information visualisation. The reason for these additions, Hogan (p.5) explains, is that "the world is not filled with data. It is filled with phenomena, which we convert to data through operationalisation. Then once operationalised through measurement or encoding, we can see how it first becomes data and then serves as the basis for information, knowledge, and ultimately, wisdom." Thus, "in many senses, social data science is the *science of the operationalisation of social life*" (Hogan, 2022, p.10).

The power of this approach to social data science is its stepping back from data, which is often taken as a given in fields such as software engineering research or social data science, and its encouragement of a critical approach to the choices and limitations involved in the operationalisation of social life into researchable problems. My appreciation for this approach to social data science was enhanced during my research visit at the OSS Data Analytics Lab at Peking University, where my colleagues, who were highly talented computer scientists, approached research problems through a positivist lens, focusing on the measurement and analysis of "objective" data that was retrievable through web-scraping and Application Programming Interfaces (APIs). I may not have been able to convince all my lab colleagues to adopt a more critical approach in their own research, but this approach guided my research, acknowledging the limitations of quantitative methods in OSS research as well as its blindness to activity that is not digitally recorded or visible to the public (Hossain, 2021; Geiger et al., 2021). I discuss these considerations in more detail in section 3.3.2. These limitations also enhanced my appreciation for mixed-methods approaches that seek out the sweet spot between the relative strengths and weakness of qualitative and quantitative research methods respectively.

3.1.2 The Socio-Technical Framing of Software Engineering Research

Given that my research concerns the social dynamics of OSAI development, which broadly falls under the practice of software engineering, I consulted literature on research designs in the discipline of software engineering research. The "Guide to Advanced Empirical Software Engineering" by Shull, Singer, and Sjøberg (2008) had a major influence on my research design. In particular, the chapter on "Selecting Empirical Methods for Software Engineering Research" by Easterbrook, Singer, Storey, and Damian (2008) gave me a comprehensive overview of the strengths and weakness of research methods and the types of research questions that they can answer. Furthermore, I found their framing of software engineering research as a socio-technical discipline, spanning technological phenomena (e.g. the development of software artifacts) and social phenomena (e.g. software engineering as

a social process), useful for building connections between research designs in this discipline and theoretical questions in the political economy of OSAI literature.

This handbook also informed my pragmatic epistemology and realist ontology. Pragmatism, which evaluates knowledge based on its practical utility in addressing real-world problems (Menand, 1997), is common in software engineering research, where scholars tend to prioritise practical problem-solving over strict adherence to a single philosophical tradition. As Runeson, Höst, Rainer, and Regnell (2012, p.7) observe, "The community does not pay any larger attention to the inherent conflict between the positivist foundation for experiments and the interpretive foundation for case studies. This conflict has caused life-long battles in other fields of research." Instead, pragmatic researchers select methods based on their effectiveness in solving a research problem, and often apply mixed-methods approaches to examine a phenomenon from multiple angles (Johnson & Onwuegbuzie, 2004). This approach resonated with my research aims. Beyond the academic literature, it was important for me that my empirical findings contributed to practitioner and policy debates about OSAI, from demystifying the commercial interests in "AI democratisation" (RP2) to documenting OSS developers' views on the merits and drawbacks of public funding (RP4).

My realist ontology complements my pragmatic epistemology. Specifically, this means that I view OSAI development as an objective phenomenon that exists independently of human perception, whilst recognising that its practice is influenced by social factors (Easterbrook et al., 2008). This dual perspective is particularly valuable when studying commercial participation in OSAI, where objective metrics like lines of code or downloads exist alongside subjective elements, such as developer motivations, corporate strategies, and community values, among others.

The combination of my pragmatic epistemology and realist ontology directly influenced my research design choices. In particular, I employed various research designs that suited the research problem at hand: I employed mixed-methods in RP1 and RP2, quantitative methods in RP3, and qualitative methods in RP4. In the case of RP1 and RP2, the mixed-methods research designs enabled me to capture both quantitative data about development activities and qualitative insights into individuals' perspectives. Meanwhile quantitative methods were suitable to the exploratory nature of RP3, which sought to understand development activity patterns at scale on HF Hub as a novel but increasingly important platform for OSAI development; while qualitative methods were suitable for RP4 which sought to understand the personal views of OSS developers about funding in depth.

3.2 Reflections on Case Study Research Design

3.2.1 The Argument for Case Studies

In RP1, RP2, and RP4, I conducted case studies on OSS projects within the wider OSAI ecosystem that comprises "over 300 critical open source projects offering over 500 million lines of code, contributed by over 35,000 developers" (Haddad, 2022). In the field of software engineering research, a case study is "an empirical enquiry that draws on multiple sources of evidence to investigate one instance (or a small number of instances) of a contemporary software engineering phenomenon within its real-life context, especially when the boundary between phenomenon and context cannot be clearly specified" (Runeson et al., 2012, p.12). Case studies are a flexible research design, which support various research aims, from rich description and illustration of individual cases to explanation and confirmation of prior theory (Yin, 2018). Novelty, testability, and empirical validity are strengths that case study research offer for theory-building, especially in new research areas (Eisenhardt, 1989). Furthermore, case study research design is amenable to the mixing of methods and data sources, suiting my epistemological stance and enabling triangulation (Yin, 2018).

While I explain and evaluate the case study research design of each RP in their respective sections, I would like to address two common criticisms of case studies here. The first concerns the problem of generalisation; that is, case studies struggle to produce statistically generalisable findings. I temper the first criticism by acknowledging that this criticism only holds if statistical generalisation is one's aim (Runeson & Höst, 2008). However, like experiments, case studies typically do not aim for statistical generalisation. On the contrary, case studies aim for analytical generalisation, which means they aim to generalise findings to theoretical propositions, which operate "at a conceptual level higher than that of the specific case" (Yin, 2018, p.73). Furthermore, case study research can be conducive to theory-building because it can produce a detailed understanding of a phenomenon, validated and triangulated through multiple data sources and methods, which in turn can serve as an empirical foundation for generating, rejecting, or extending theories (Eisenhardt, 1989).

The second criticism concerns the risk of greater researcher bias in case study research than in other research designs, such as surveys and experiments, insofar as case selection and data selection are concerned (Easterbrook et al., 2008). As with every research design, this criticism can be tempered by specifying and documenting a systematic procedure for selecting and analysing case studies, from research design to data collection and analysis (Easterbrook et al., 2008; Yin, 2018).

3.2.2 Case Selection Procedure

I took the following two-step approach to the selection of cases for RP1, RP2, and RP4:

First, I referred to the LF AI & Data Foundation's database of OSS projects to establish boundaries for case selection (LFAI&Data, 2022). In January 2022, this database contained 318 OSS projects, complete with information about the projects' host organisation, licence, community size, and star count on GitHub, among others. As my research concerns OSAI, I removed "data" projects, which included OSS for database management (e.g. MySQL), data wrangling (e.g. pandas), data operations (e.g. Apache Hive), and data visualisation (e.g. seaborn), among others. This removal resulted in a dataset of 184 AI OSS projects. It became clear upon inspection that the dataset was incomplete; for instance, all OSS libraries within the JAX ecosystem by Google DeepMind were missing. Given this absence, it is unclear how many other OSS were missing from this database. Based on these gaps, the dataset was used as an indicator, rather than complete reference, of the AI OSS landscape. Subsequently, I labelled every OSS according to the sector of its host organisation (company, non-profit foundation, or university/public research institute) and its licence type (permissive or copyleft). Next, as a measure of popularity and adoption, I collected download data for each project from their respective ecosystem managers from 1 January 2017 to 31 December 2021.

The second step concerned the selection of cases from within these boundaries according to the specific study propositions of the respective RPs (Runeson & Höst, 2008). Variables such as the host organisation sector, GitHub stars (a feature that allows developers to both save and show appreciation for a repository), the community size (i.e. code contributors to the GitHub repository), and downloads were considered in the selection. The download statistics contained surprises; for example, Meta's PyTorch ranked 21st with 4,373,593 downloads, even though it had already emerged as the most adopted DL framework in academic AI research (PaperswithCode, 2023) and for training OMs (Foster, 2022; Gururaja et al., 2023) by then. I speculate that this low download count, which was the sum of pip and conda downloads, understates actual downloads and may have been due to the exclusion of downloads through alternative avenues, such as Docker containers or cloud platforms like Google Colab. Of course, this raises questions about the reliability of download counts for all OSS projects. For this reason, I did not select cases solely based on downloads; rather, I considered downloads together with other variables, such as GitHub repository community size and stars.

To give you one example: in RP4, I selected scikit-learn for the single case study on the role of funding in supporting successful, community-led OSS projects because I thought it was remarkable that a community-led OSS project, which is developed by researchers based at the French Institute for Research in Computer Science and Automation (Inria) and a global community of 2,250 contributors,

ranked first by downloads (621,408,470) in the industry-dominated OSAI ecosystem. In addition, in terms of GitHub stars, it ranked fifth (46,177) after four company-hosted OSS projects: Google's TensorFlow (156,813), Google's Keras (51,724), Meta's PyTorch (48,964), and HF's Transformers (47,486). For wider comparison, among the 38 OSS projects with over 1 million downloads, the mean community size was 520 (SD: 812) and the mean star count was 18,156 (SD: 27,769). Thus, in addition to having the most downloads, scikit-learn has a relatively large contributor community and star counts on GitHub. For these reasons, I selected it as a single case study that could reveal practical lessons for the success and sustainability of community-led OSS projects.

3.3 Reflections on Empirical Research Methods

In this section, I discuss the mixed-methods research designs of the RPs. For the reader's convenience, I'll briefly summarise the methods employed in each RP. Chapter 4 (RP1) employs a combination of data mining, SNA, and interviews to examine the structures, patterns, and practices of open source co-opetition in three company-hosted AI OSS projects. Chapter 5 (RP2) employs document analysis, questionnaires, and semi-structured interviews to analyse commercial incentives for OSS donations. Chapter 6 (RP3) employs data mining and SNA to explore development practices in over 500,000 repositories on HF Hub. Finally, Chapter 7 (RP4) employs semi-structured interviews to investigate the role of funding in sustaining scikit-learn as a community-led OSS project.

3.3.1 Reflections on Mixed-Methods Approaches

In Chapter 4 (RP1), I employ a mixed-methods approach that combines data mining, SNA, and semi-structured interviews. This research design was informed by two factors. First, the choice of research methods and the sequence of their application was informed by prior work on open source co-opetition (Nguyen Duc et al., 2019; Teixeira, Mian, & Hytti, 2016). Specifically, I employed a sequential research design, beginning with data mining and SNA to understand development activity patterns and the social network structure of inter-company collaboration in OSS projects, followed by semi-structured interviews with developers, which thickened the quantitative findings with the personal anecdotes and opinions. In addition, I presented findings from the quantitative analysis to the interviewees to guide the interview discussions and elicit their reactions. The reactions were mixed—some interviewees confirmed the findings, while others contested them—, enhancing my confidence in the validity and limitations of the findings (Drost, 2011). Overall, the triangulation of methods proved an effective strategy for the RP's theory-building aims (Eisenhardt, 1989).

The second factor was prior work that has spearheaded the complementarity of qualitative and

quantitative methods involving digital trace data; that is, "records of activity (trace data) undertaken through an online information system (thus, digital)" (Howison, Wiggins, & Crowston, 2011, p.769). In general, trace data have three key characteristics: they are found rather than produced for research, event-based, and longitudinal. What distinguishes digital trace data from (non-digital) trace data is that they are "both produced through and stored by an information system" (Howison et al., 2011, p.769). For example, Lehdonvirta et al. (2019, p.580) illustrate in their analysis of the global platform economy that digital trace data and interviews provide complementary affordances: "The trace data provide unbiased observational measures and formal generalizability to a limited population, while the qualitative data provide rich, noisy measures covering a wide range of cases but lacking formal generalizability." They explain that the aim is "not simply to strengthen the reliability of our conclusions through convergent validation (triangulation) but to increase the 'analytic density' of the research (Fielding, 2012), helping to extend the scope and depth of the conclusions by 'offsetting' one data source's weaknesses with the other (Bryman, 2006)." Similarly, Hand, Hillyard, Pole, and Love (2014, p.14) argue that digital trace data analyses and interviews "might complement one another or be recombined in interesting and productive ways that problematize previous divides between 'surface' (quantitative) and 'deep' (qualitative) data on large and small populations respectively. New patterns might be spotted at one 'scale' then lead to novel questions at another."

Drawing on these insights, I viewed quantitative methods, such as data mining and SNA, as appropriate for exploratory (what) and descriptive (how) questions about open source co-opetition patterns and qualitative methods, such as semi-structured interviews, as appropraite for descriptive (how) and analytical (why) questions about incentives and collaboration practices. The combination of these methods allowed for different kinds of questions to be combined (Robson & McCartan, 2016) and for enhancing research validity through triangulation (Creswell, 2018; Jick, 1979).

3.3.2 Reflections on Quantitative Methods

In this section, I discuss and reflect on my approaches to quantitative methods, with a focus on data mining and SNA undertaken in RP1 and RP3. For these RPs, I programmatically collected data via application programming interfaces (APIs), the GitHub REST API and the HF Hub API API, to analyse collaboration in OSAI developer communities on these platforms. As noted above, digital trace data does not just simply exist as research data. In both cases, I produced research data by making a number of choices and navigating API affordances and limitations. My approach was informed both by the aforementioned PODIKW framework (Hogan, 2022) and guidelines concerning validity issues when using digital trace data for SNA (see Figure 3.1). Howison et al. (2011, p.790) advise scholars to consider how platform affordances shape what types of digital trace data one can access and what

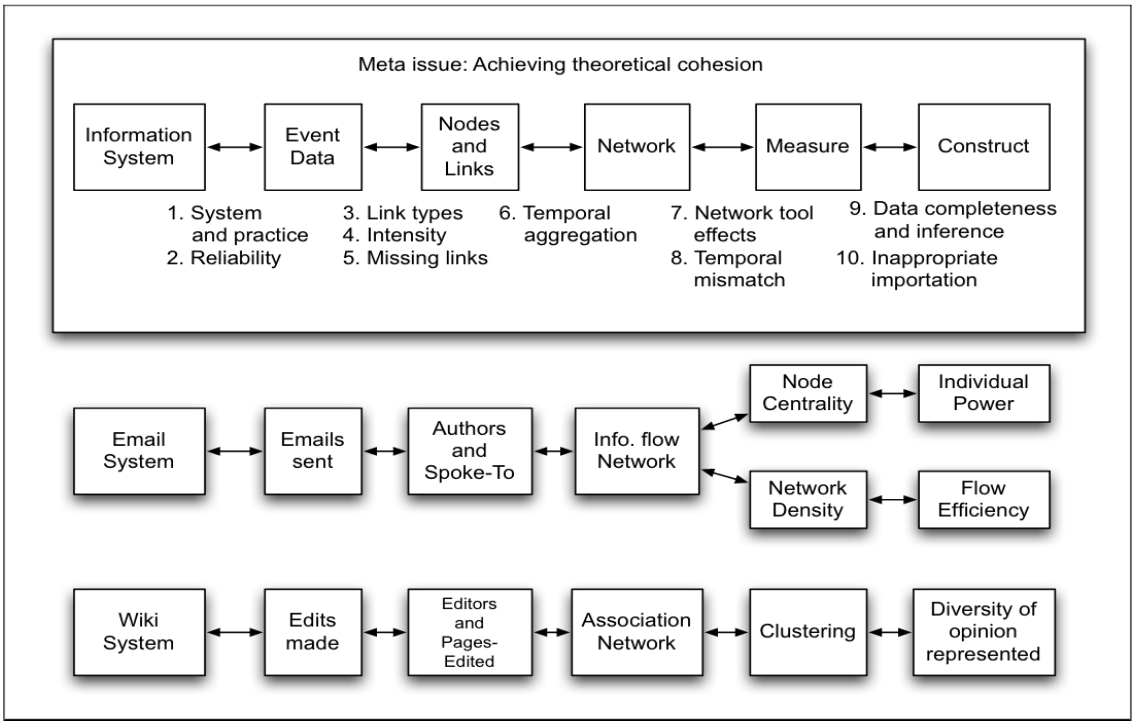


Figure 3.1: Validity Issues of SNA with Digital Trace Data
Source: Howison et al. (2011, p.772)

types of research are made possible, from what features exist to what data is stored and accessible.

3.3.2.1 Mining Research Data from GitHub

GitHub is the most popular OSS platform with around 89,000,000 active users (GitHub, 2022), ahead of SourceForge with around 30,000,000 active users (SourceForge, 2022). GitHub was founded in 2008 and bought by Microsoft in 2018. It combines the distributed version control and source code management functionalities of Git with proprietary features for issue tracking, task management, and discussion, among others. The types of interactions that take place on GitHub and the digital trace data these interactions leave behind are shaped by the platform features, which make certain types of interactions possible on the platform. It is now the dominant platform for OSS development and has shaped a "GitHub generation" of OSS developers who rely on the platform's user-friendly, standardised features, as opposed to the comparatively messy and disorganised online forums and mailing lists that were used by older generations of OSS developers (Eghbal, 2020). It should be noted that a considerable amount of activity within OSS projects takes place outside the repository, such as hackathons (Hossain, 2021) or "invisible labour" like community management tasks that keep OSS projects running (Eghbal, 2020; Geiger et al., 2021), GitHub should therefore be understood as an important, but limited, data source about OSS development.

3.3.2.2 Mining Research Data from HF Hub

HF Hub is a Git-based social coding platform that was launched in 2021 by the start-up HF. It has emerged as the principal platform for sharing and using OMs, datasets, and other AI resources. It is considered a counterpart to GitHub for AI models and datasets. HF Hub provides a user-friendly interface for exploring and comparing models, as well as tools for model evaluation and deployment. In light of the platform's growing popularity, scholars have begun to examine its potential as a data source for empirical studies on ML model development as a new area of software engineering (Ait et al., 2023b, 2023a; Castaño et al., 2024). RP3 contributes to this effort, showing how HF Hub is a fruitful, yet limited, data source. In particular, the API is not yet optimised for research purposes, which makes data collection slow (e.g. it requires unique API calls per model and the handling of rate limits) and limited (e.g. user metadata is not available).

As I discuss in detail in RP3, API limitations created some threats to validity. For example, unlike with the GitHub REST API, the lack of user metadata via the HF Hub API made the process of merging usernames difficult, resulting in a potential oversampling of developers that have multiple accounts and thus threatening the validity of the findings. More specifically, username merging is a necessary step in OSS research due to the presence of multiple usernames per unique developer, which arises when individuals use different accounts and/or how Git records a user's information based on how their credentials are recorded in the local Git repository that they submit commits from (Bird, Gourley, Devanbu, Gertz, & Swaminathan, 2006; Goeminne & Mens, 2013; Kouters, Vasilescu, Serebrenik, & van den Brand, 2012; Robles & Gonzalez-Barahona, 2005). Unfortunately, there is no perfect solution to merging usernames. While the GitHub API provides metadata about users, such as email addresses or locations, which can be used to improve the accuracy of automated username merging (Zhu & Wei, 2019), the HF Hub API does not provide metadata about users, which made it challenging to merge usernames either automatically or manually with confidence.

3.3.2.3 Common Challenges

Challenge 1: Operationalising Collaboration

Regardless of platform, an important step in the data mining process concerns how to operationalise collaboration and what unit of analysis to select. One advantage is that the digital traces of interactions among the full population of contributors to an OSS project are publicly available, for example via the Git version control systems used in repositories on platforms like GitHub, which allows for complete sampling of activity by the research population on such platforms.

There are many approaches to operationalising collaboration in OSS development in terms of in-

teraction type, intensity, and periods. Prior work has operationalised collaboration as interactions in mailing lists, issue and pull request trackers, and code authorship via code commits (Y. Zhang, Zhou, Stol, Wu, & Jin, 2020; Teixeira & Lin, 2014; Teixeira, Robles, & González-Barahona, 2015; Snarby, 2013). While prior work focuses in large part on activity in OSS repositories, as sites where the software is discussed and developed, it is important to acknowledge that collaboration occurs through various channels, both public and private, online and offline, beyond the repository (Casari, Ferraioli, & Lovato, 2023). In light of this historical focus on repositories, Casari et al. (2023, p.14) encourage researchers to broaden their view to the wider ecosystem—that is, "the collection of repositories, the community, their interactions, incentives, behavioral norms, and culture"—around a repository.

In RP1, contrary to prior work on open source co-opetition that has operationalised collaboration based on instances where two company-affiliated developers have interacted (usually through comments) on the same issue or bug report (Linåker, Rempel, Regnell, & Mäder, 2016; Nguyen Duc et al., 2019), I opted to operationalise collaboration as instances where two company-affiliated developers have submitted a commit to the same file during the same release cycle because commits are easily measurable, represent "validated" contributions to the codebase of an OSS project, and represent an accurate audit trail of collaborations on code authorship (Orucevic-Alagic & Host, 2014; Y. Zhang et al., 2020). Furthermore, I opted to record relationships between pairs of developers as reciprocated, directed ties which were weighted according to each developer's relative lines of code (LOC) contribution to the file during a release cycle in order to account for the intensity of collaborations as well as the temporal dimensions of collaboration (Basole, 2009). Similarly, in RP3, I operationalised collaboration as instances where developers had made a commit to a model repository on HF Hub, with direct edges recorded between developers that were weighted by the number of times a developer contributed a commit to the same repository or repositories as the other developer.

Challenge 2: Choosing the Unit of Analysis

The choice of unit of analysis is also key. In RP1, I followed prior work and chose the company as the unit of analysis rather than the individual developer (Nguyen Duc, Cruzes, Hanssen, Snarby, & Abrahamsson, 2017; Linåker et al., 2016; Orucevic-Alagic & Host, 2014). While this choice makes sense given that open source co-opetition concerns collaboration between companies, it was not straightforward, as the different levels of analysis provide different insights. Aggregation of individual developers to the company-level loses rich information about different social positions and practices of individuals and it assumes that company-affiliated developers are abiding ambassadors to the interests of their companies. Though, with frequent job-hopping, it is common that developers are active in OSS projects for longer periods than they are employed by a company (Lumbard & Ger-

monprez, 2017). Their activity in a OSS project may be encouraged or directed by their companies, but we cannot assume such developers have no personal loyalty to or interests in the project.

Challenge 3: Identifying Company Affiliations

Regardless of the unit of analysis, as I explain in RP1, it is cumbersome to identify the company affiliations of developers. Domain-mining, which is considered the most accurate approach (Nguyen-Duc, Cruzes, Terje, & Abrahamsson, 2019; Mehra, 2011), only identified the affiliations for 37.0%, 48.9%, and 9.2% of contributors to the PyTorch, TensorFlow, and Transformers repositories respectively. In turn, the manual labelling of affiliations by searching developers on the Internet was a cumbersome, inaccurate (e.g. it was difficult to match timestamped commits with self-reported affiliations that often had coarse timestamps), and time-intensive (e.g. the process took myself and two co-authors one working week to complete). What is more, the interviews with developers revealed further difficulties in attributing affiliations to developers: two interviewees commented that their contributions were non-strategic bug fixes and thus felt that they should not be considered as company-affiliated developers, rather volunteers. These comments underline the difficulty of accurately attributing affiliations to users, which causes a major threat to the validity of this kind of research.

Challenge 4: Static versus longitudinal networks

Another consideration concerns the choice between static or longitudinal representations of collaboration. For example, in RP1, I took a longitudinal approach because it enabled the analysis of the evolving social structures of collaboration in the OSS projects (Teixeira et al., 2016; Linåker et al., 2016) and painted a more dynamic picture of the evolution of collaborations between various companies (Crowston & Howison, 2005). This was made possible through the collection of timestamps for each interaction, which GitHub records and makes available through its API. I separated activity into annual snapshots based on the dates of software releases to enable meaningful comparative analysis of the respective collaboration networks (Linåker et al., 2016; Long & Siau, 2007). By contrast, in RP3, I did not take time into consideration in the representation of collaboration in 350,000 OM repositories. This choice was made simply due to convenience. However, I acknowledge that studying the structural evolution of collaboration networks would have been of interest to the research aim to describe the social network structure of collaboration on this novel platform.

This section has provided a brief discussion of some key considerations in the operationalisation of collaboration in OSAI developer communities on various online platforms in line with best practice guidelines (Howison et al., 2011; Hogan, 2022). RP1 and RP3 provide more detailed discussions of

these methodological considerations and the respective threats to validity.

3.3.3 Reflections on Qualitative Methods

In this section, I discuss and reflect on my approaches to qualitative research, with a focus on research interviews undertaken for RP1, RP2, and RP4.

3.3.3.1 Approach to Research Interviews

Research interviews are a "conversation with a purpose" (Burgess, 1984, p.102), which are a suitable method for gathering information about individuals' opinions, thinking, and motivations, among others (Kvale, 1996; Gubrium, Holstein, Marvasti, & McKinney, 2012). Semi-structured interviews are an effective strategy to "combine depth of understanding with purposeful, systematic, analytic research design to answer theoretically motivated questions" (Lamont & Swidler, 2014, p.159).

Since interviews are employed to offer insights into the world, one's epistemological stance is crucial at every stage of the interview-based research process, from planning interviews to analysing the data collected (Eynon, 2022). Two popular approaches in interviewing research are those of the interviewer as a miner (modernist approach) or the interviewer as a traveller (post-modern approach) (Brinkmann & Kvale, 2015). In the former, the interviewer seeks to unearth knowledge and to bring it to the surface; whilst in the latter, the interviewer is a traveller who embarks on a journey and constructs knowledge interactively with interviewees and through reflective interpretation of how the "sights" from these travels are turned into knowledge (Brinkmann & Kvale, 2015).

My approach was closer to the former: for all RPs that involved interviews, I conducted semi-structured interviews with carefully prepared questions to collect insights on particular topics, be it company-affiliated developers' motivations for donating their OSS project in RP2 or developers' opinions of the relative merits and drawbacks of different funding sources in RP4. However, similar to the PODIKW framework for social data science (Hogan, 2022), I did not assume that knowledge simply exists "out there" as objective facts and that it was my task to "extract" it and analyse it. Reflexivity, which I discuss in detail below, is critically important in this regard. Furthermore, I employed semi-structured interviews and open-ended questions to allow for spontaneity and unexpected insights beyond what the questions I had planned beforehand (Gorman, Clayton, Shep, & Clayton, 2005; Kvale, 1996). In other words, I did not assume to know everything that could be "mined" in the interviews, and this approach expanded the boundaries of insights in my data. What is more, in contrast to "mining" objective facts about the world, for me subjective and contesting opinions were acceptable or even desired, as they shed light on important nuances (Beitin, 2012b; King, 2009).

Some interviews were conducted in-person and some were conducted digitally, which had respec-

tive benefits and drawbacks. I conducted digital interviews for RP1 and RP2. The benefits were the immediacy and speed of reaching OSS developers, who were scattered around the world (Gorman et al., 2005). However, it is true that in digital interviews it is more difficult to build rapport with interviewees and to see visual cues or to read vocal tone and body language that interviewees use to express themselves (Edwards & Holland, 2013). As my approach to qualitative data analysis (QDA) followed an approach to thematic analysis that paid attention to recurrence, repetition, and forcefulness in interviewee responses (Lawless & Chen, 2019), it was important for me to record these details during the interviews in short-hand notes or in extended reflections after interviews (Strauss, 1987). Unsurprisingly, it was relatively easier to capture these details in the in-person interviews I conducted for RP4. While the period of data collection was long (November 2021 and April 2023), which involved significant amount of my time for preparation, travel, execution, and transcription (Gorman et al., 2005), it enabled a deep immersion in the scikit-learn project and trust-building with the maintainers (Conti & O'Neil, 2007; Howlett, 2021; James & Busher, 2009).

3.3.3.2 Reflexivity

In social science research, it is critical that one as a researcher proactively considers and takes into account the manifold ways in which one's positionality and disciplinary conventions may influence one's research, from its initial design through to its completion (Jacobson & Mustafa, 2019). Throughout the research process, I adhered to guidelines for exercising reflexivity, which concerns an awareness and ongoing reflection on "the position from which we see the world around us [and how it] impacts our research interests, how we approach research and respondents, the questions we ask, and how we interpret data" (Jacobson & Mustafa, 2019, p.1). It is important to begin this process of reflection at the outset of a research project, rather than once it is finished (Finlay, 2002).

I adopted social identity maps as a tool to help me identify and reflect on how my intersectional social identity influenced research conducted for this thesis (Jacobson & Mustafa, 2019). Figures 3.2 and 3.3 show illustrative summary visualisations of more extensive hand-written social identity maps that I had developed at the Transfer of Status milestone (i.e. prior to commencing research) and throughout the interview research processes of RP1, RP2, and RP4. While the visualisations of the social identity maps show identifiers in separated identity boxes, the intention is to illustrate the intersectionality of the various identifiers. I note that such maps are "not meant to be used as a rigid tool but rather as a flexible starting point to guide researchers to reflect and be reflexive about their social location" (Jacobson & Mustafa, 2019, p.1). Their use "goes beyond only naming one's own social identity. [The] purpose is to reflect on one's positionality and how it becomes implicated in action during research as well as to hold researchers accountable to a high standard of social and

moral responsibility before, during, and after their research" (Jacobson & Mustafa, 2019, p.8).

There are a number of considerations that I will discuss in brief here. First and foremost, I recognise that my position as a middle-class, white, Northern European man pursuing a fully-funded PhD at the University of Oxford is one of significant privilege, which undoubtedly shapes the lens through which I see the world. This intersection of social identities affords me numerous advantages, such as access to people and resources, perceptions of credibility and authority, and easier navigation in a predominantly white and male-dominated industry. Furthermore, my fully-funded scholarship has enabled me to conduct research, including international fieldwork, attend international conferences, as well as participate in a six month research visit at Peking University's OSS Data Analytics Lab in Beijing, China. This privilege also gives me potential blind spots, particularly in understanding the experiences of underrepresented groups—such as women, non-binary individuals, and people of colour—and stakeholders from outside Western contexts, especially the global majority.

The intersection of my nationality (UK/Germany) and European values influenced my understanding of public interest and the responsibility of the state in supporting the development, provision, and sustainability of OSS as digital public goods and digital infrastructure (Eghbal, 2016). For example, this political perspective motivated my research on the role of public funding in supporting non-commercial OSS projects in Chapter 7 (RP4) as well as research on public sector support for OSS and the digital commons outside of my thesis (Osborne, Sharratt, et al., 2024; Osborne, Boehm, & Jimenez Santamaria, 2023). While I have lived in Beijing, China and Bogotá, Colombia during the writing of this thesis, I acknowledge that my worldview is Europe-centric, which risks overlooking alternative viewpoints or state capacities from non-European regions.

My political views have also shaped my approach to the political economy of OSAI. I am a European social democrat who holds political views about the inequalities and harms of global (digital) capitalism, as well as the promise of OSS and the digital commons as a value system for promoting social justice, digital democracy, and economic fairness. For example, I was naturally inclined towards critiques of commercial interests in OSS, and strategies for protecting the collective interests of OSS developer communities against capture—or "incorporation" (Birkinbine, 2020)—by corporate actors. Furthermore, I am concerned about the impacts of the normalisation of free-riding on OSS developers' labour and commercial hosting of OSS projects, which lack open governance, on OSS development and governance norms, which certainly influenced my perspective, including my study of collaboration practices in company-hosted OSS projects in RP1.

What is more, work experience in AI policy at the UK Government and as a researcher at theLF also undeniably had an influence on my perspective on the political economy of OSAI, including the roles and responsibilities of states and non-profit foundations in the OSAI commons. First, I acknowledge

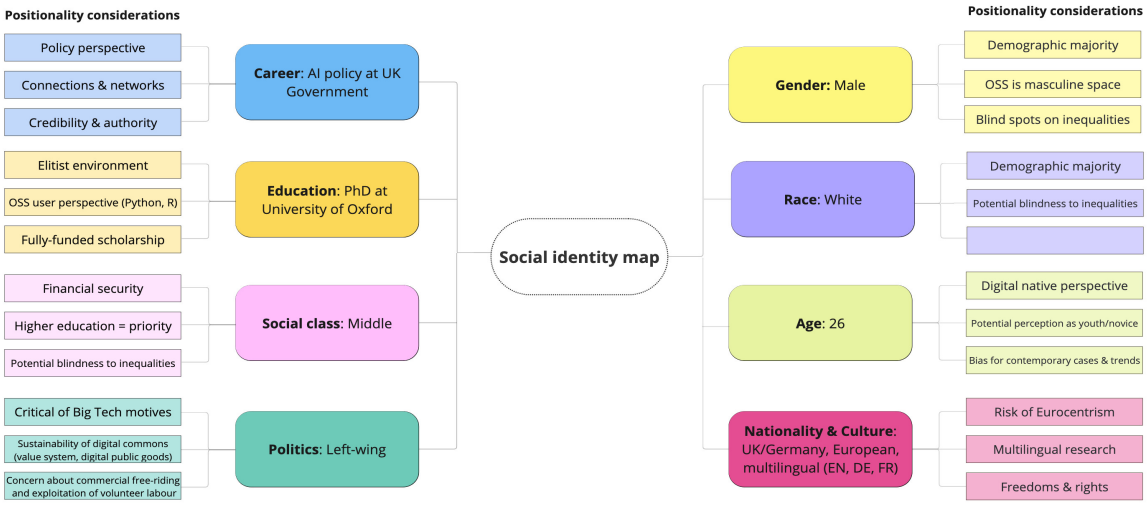
that this work experience provided me with certain privileges, including insider insights on AI policy-making and OSS collaboration, governance, and funding. They may also have contributed to a perception by research subjects that I am an OSS or AI policy insider, potentially influencing their responses. These roles also facilitated access to research subjects. For example, for RP4, I was able to recruit the coordinator and the deputy coordinator of the French AI strategy for interviews with relative ease, owing to an existing relationship from my time at the UK government.

My role as a researcher at the LF has influenced my research in several ways. For the reader's understanding, since June 2022, I have worked at the LF as a researcher on research projects concerning public sector adoption of and funding for OSS (e.g. OpenForum Europe, 2023; Osborne et al., 2023). This work included co-authoring a toolkit on measuring public funding impacts on OSS development with Germany's Sovereign Tech Fund (Osborne, Sharratt, et al., 2024). The *leitmotif* of this toolkit is that by understanding the (positive or negative) impacts of public funding on OSS development, we can work towards designing funding that work both for OSS developers and the public interest. This affiliation affected my research in at least three key ways. First, attending LF conferences and workshops provided valuable practical insights into OSS development, from technical aspects to community dynamics. They provided a practical education about and immersion into both the exciting and the banal of OSS development. Second, I often heard critiques, especially among Europeans, that the LF, as a US-based 501(c)(3) entity, primarily serves its corporate members' interests through a "pay to play" model (Butler et al., 2018). While my research focus on the public sector limited direct corporate exposure, this environment may have subconsciously tempered my critique of corporate involvement in OSS. Third, as discussed in RP2, my LF affiliation likely increased participation in interviews and may have introduced biases regarding the hosting of OSS projects by vendor-neutral foundations versus companies (RP1, RP2) as well as response or social desirability biases in the interviews (RP2). To maintain the independence of research, I relied on my doctoral funding and used social identity mapping for regular positionality reflection.

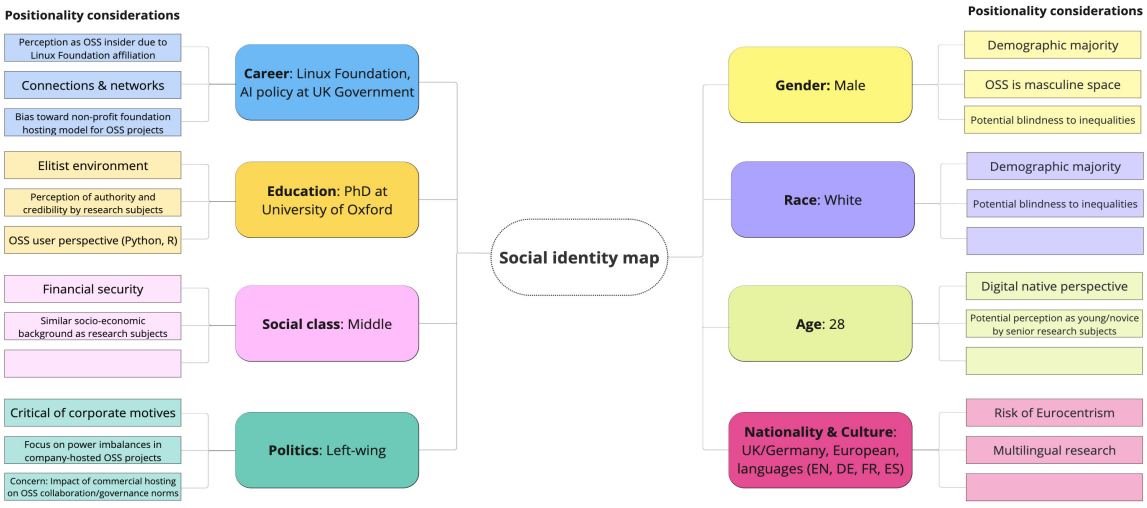
These considerations represent only a few of the many ways in which my positionality has influenced my research. Throughout the process of writing this thesis and the RPs, I sought to remain cognisant of these potential biases and maintain a reflexive stance to enhance the rigour of my work. However, I acknowledge that it is impossible to control for all biases and as such as the reader should interpret my approach to reflexivity as a best-effort but inevitably imperfect one.

3.3.3.3 Common Challenges

In this section, I reflect on three major challenges that I faced: gaining access to interviewees, dealing with corporate secrecy and narratives during interviews, and interviewing elites.

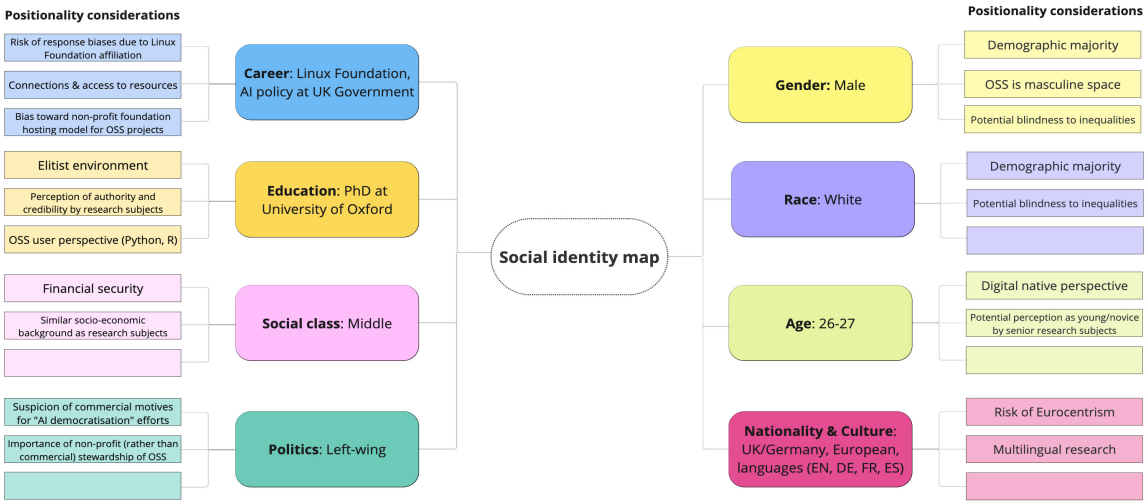


(a) Illustrative Visualisation of Social Identity Map for Transfer of Status

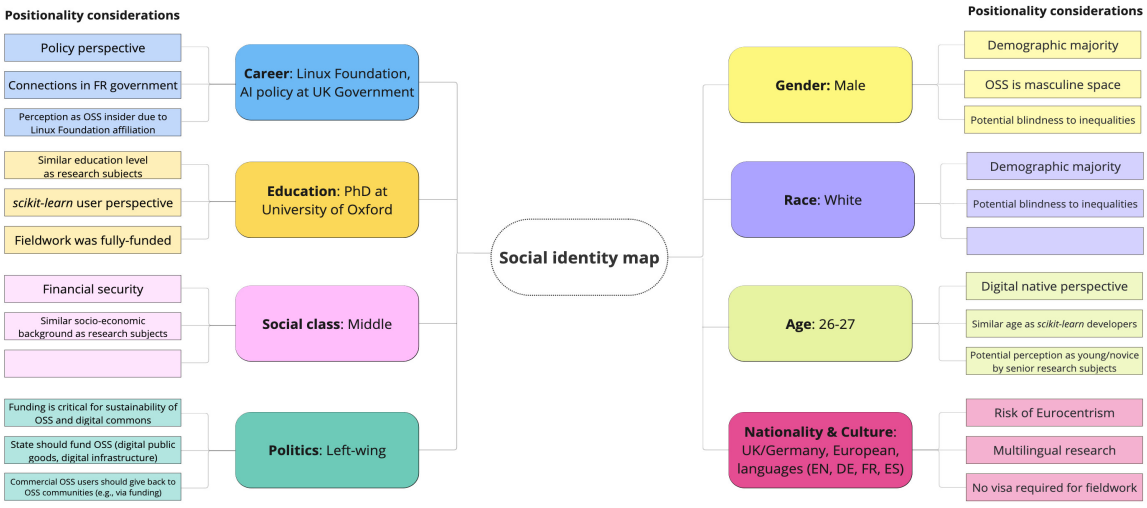


(b) Illustrative Visualisation of Social Identity Map for Chapter 4 (RP1)

Figure 3.2: Summary Visualisation of Social Identity Maps (1-2)



(a) Illustrative Visualisation of Social Identity Map for Chapter 5 (RP2)



(b) Illustrative Visualisation of Social Identity Map for Chapter 7 (RP4)

Figure 3.3: Summary Visualisation of Social Identity Maps (3-4)

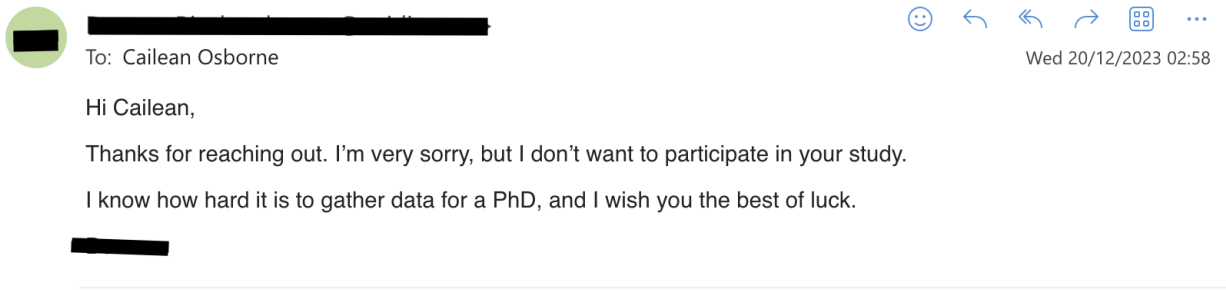


Figure 3.4: Rejection to Interview Invitation

Challenge 1: Gaining Access to Interviewees

A challenge I experienced for RP1, RP2, and RP4 was gaining access to stakeholders within companies, such as developers (RP1 and RP2) or corporate representatives in the scikit-learn consortium (RP4), which was significantly more difficult than gaining access to non-company-affiliated developers. As I discuss below, I suspect a major reason for this challenge was either the presence of company policies (e.g. non-disclosure agreements) that restrict employees from speaking about their commercial strategy or a fear among employees that they would say something that had to be approved by their legal team. Other reasons may have been a lack of time, interest, or compensation.

For RP1, I purposely sampled developers who had contributed code to TensorFlow, PyTorch, and Transformers, who were not affiliated with the respective host companies. In total, I sent 350 personalised emails and received 13 responses (3.71%), including 10 acceptances (2.86%). As shown in Figure 3.4, one contributor, who was a senior developer at a major AI hardware company, rejected the invitation but empathised that he understood how difficult it can be to collect data for one's PhD. Curiously, he added me on LinkedIn and wished me good luck with my PhD. This interaction suggested good will and the possibility of self-censorship due to factors, such as a company policy. Furthermore, attempts to snowball sample interviewees proved unsuccessful, as interviewees were unwilling to personally facilitate introductions. Two interviewees explained that this was due to the need to stay anonymous in light of corporate non-disclosure policies.

I faced similar challenges whilst working on RP2, despite enjoying privileges, such as an affiliation with the LF, that in theory should have facilitated access to company-affiliated developers who had donated OSS projects to the LF. Specifically, this study was inspired when I learned that Meta was going to donate PyTorch, its DL framework, to the LF before it was publicly known. Being affiliated with the LF, I became familiar with Meta's strategic incentives which were not disclosed in public announcements that came out a few months later. In the interest of reproducibility, my sampling strategy relied on contacting my research population—that is, maintainers who had donated the

respective OSS projects to the LF AI & Data Foundation or the PyTorch Foundation and staff at these foundations—by sending an email to the respective public mailing lists of technical projects or via publicly available contact information. Overall, this resulted in responses from 14 maintainers, the Executive Director of both foundations, and a project coordinator at the LF AI & Data Foundation.

While I primarily relied on publicly-available contact information and faced difficulties in recruiting interviewees, I recognise that being affiliated with theLF nonetheless enables access to some interviewees. For example, the staff already knew who I was, which surely increased the likelihood of their willingness to participate in my study. In addition, when I did not receive a response from the PyTorch maintainers via publicly available contact information, I asked the Executive Director if he could help me reach out to maintainers on my behalf. Despite his support, I was not able to recruit any of the PyTorch maintainers for interviews. Similar to RP1, I suspect this unwillingness was related to corporate policies (e.g. NDAs or the need for Legal to review statements), in addition to a lack of time or interest. Furthermore, as I discuss in RP2, I recognise that my affiliation with theLF may have influenced the willingness of interviewees to participate in interviews as well as social desirability bias in their responses. I sought to minimise these biases by explaining at the beginning of the interviews that the research was being independently conducted and funded for my PhD, and none of the data would be shared with the LF. Several interviewees ended the interviews with expressions of good luck for my PhD and an interest in reading my paper once published, which reassured me that my interviewees understood the interviews were conducted for my PhD.

For RP4, I encountered both privileges and challenges with regards to gaining access to interviewees. The study began with an immense privilege. Before commencing the PhD, I had already acquired access to officials in the French government's AI Directorate through my previous employment at the UK Government's Centre for Data Ethics and Innovation. Specifically, I had built a working relationship with the coordinator of the French AI strategy, Renaud Vedel, through our respective participation in the British and French delegations in the Global Partnership on AI (GPAI). For example, I had told Renaud that my PhD would concern the political economy of OSAI and he shared that he had just secured internal buy-in to fund scikit-learn via the French AI strategy. Through this relationship, I was able to interview him and his colleagues with relative ease, which otherwise would have been difficult. Furthermore, due to our pre-existing rapport, the interviews were frank and bypassed typical challenges involved in interviewing elites, such as uneven power dynamics. Upon leaving my job to commence my PhD, Renaud invited me to the GPAI Annual Summit in Paris, France in November 2021 to meet Gaël Varoquaux, a co-founder of scikit-learn. At the summit, I met Gaël and we discussed his understanding of scikit-learn as a "political project" against the dominance of Big Tech in AI R&D. After this meeting, Gaël gave me his business card and invited me to meet the

team at their office at the Paris-Saclay campus, a science park in the outskirts of Paris.

On 23 March 2022, I visited the scikit-learn team for the first time. I was welcomed by the community manager, François Goupil, and we spoke for an hour about scikit-learn project and their funding model. We bonded quickly when we established common ties to my hometown in the south of Germany, which he had visited many times during his PhD in Vaduz, Liechtenstein. After our conversation, he invited me to join the team for lunch. Our conversations were casual and based on open-ended questions related to my research. They confirmed that I would be welcome to return for research interviews. In the year and half that followed, François played a crucial role in enabling access to their corporate sponsors by introducing me to them and by inviting me to events, such as a hackathon hosted by one of their sponsors. Despite his generous support, it was still hard to gain access to most of their corporate sponsors. Multiple efforts were made with limited success. On two occasions, François sent an email on my behalf to the consortium members, inviting them to participate in a research interview. The second time, I offered three Amazon vouchers worth 50€ each, thinking it may incentivise participation in the interviews. However, this outreach campaign resulted in only one response. One sponsor was reached by virtue of the fact that they had hosted a hackathon at their office, where I met him and asked him in-person to participate in a research interview. Another sponsor was recruited by François when they joined the consortium.

Challenge 2: Dealing with Corporate Secrecy and Narratives

Given that most interviewees for RP1, RP2, and RP4 were from companies, they may have been selective in what they revealed to me due to organisational restrictions, privacy, and/or their belief that they were not best placed to answer the question. For example, for RP2, two interviewees told me they could not answer certain questions on behalf of their company because it was not their area of expertise. Moreover, the audio recordings may have reduced interviewees' willingness to share sensitive information, considering that "[a] person chooses a perspective to present based on who is listening and the culture around them" (Beitin, 2012a, p.11). It is also possible that the stakeholders who agree to participate in an interview are the ones who have the least to hide (Beitin, 2012b).

Another challenge was dealing with narrative environments when company-affiliated stakeholders evidently told PR narratives rather than sharing their personal views (Beitin, 2012b). For example, for RP2, two interviewees from a management consultancy repeated arguments about "democratising" the project and demonstrating the credibility of their engineering teams that were also made in the company's blog post that announced the donation. For RP4, two company representatives emphasised their corporate social responsibility to support OSS and the digital commons.

Challenge 3: Interviewing Elites

Interviewing elites comes with its unique methodological challenges when compared to non-elite interviews, including gaining access, controlling the interview process, and accounting for reflexivity, truth, and power in the interviewee's account (Mikecz, 2012). Elites can broadly be understood "as individuals who might broadly have a more high-ranking status than the researcher" (Shaw, 2020, p.53). In my case, elites encompassed managerial stakeholders at companies and the LF for RP2, and managerial stakeholders at companies in the scikit-learn consortium and senior policymakers in the French government for RP4. In general, these interviews were characterised as experiences of "interviewing up" rather than "interviewing down" (Smith, 2006).

While gaining access to elites is hard enough, as discussed above, gaining their trust and building rapport is arguably even harder (Mikecz, 2012) because elites purposefully erect barriers (Laurila, 1997), typically have limited availability (Conti & O'Neil, 2007; Atkinson & Delamont, 2010), and may not take a researcher seriously (Zuckerman, 1972). These challenges can be compounded by cross-cultural differences (Mikecz, 2012). When interviewing elites, one must be flexible (Conti & O'Neil, 2007; Atkinson & Delamont, 2010), well-prepared to demonstrate one's seriousness (Zuckerman, 1972), and project "a positive image in order to gain their respect" (Harvey, 2011, p.434). It is important to begin on the "right note" by being straightforward about one's research aims and intentions (Ostander, 1993). It is also recommended to ask elites open-ended questions because in general they "do not like being put in a straightjacket of close-ended questions. They prefer to articulate their views, explaining why they think what they think" (Aberbach & Rockman, 2002, p.674).

I followed this guidance when interviewing elites. I made sure I was well-prepared with background knowledge and sent interviewees interview guides with information, charts, and images that were intended to be used as stimuli for discussion. In addition, while interviewing developers, including senior engineers and C-Suite executives in the case of start-ups, for RP1, I shared my screen to show findings from the quantitative analysis of collaboration in the PyTorch, TensorFlow, and Transformers repositories, and found that the interviewees were always willing and often enthusiastic to take a look at and discuss my findings. Some challenged the findings (e.g. low commit counts from volunteers), and shared their views about how companies collaborate on OSS development or on how I should empirically study it. For RP2, I asked interviewees questions specifically about their project and their companies' public announcements. In some cases, I shared my screen to highlight certain phrases. This attention to detail demonstrated that I had "done my homework." As mentioned above, for RP4, I interviewed the coordinator of the French AI strategy and his deputy. Due to our existing professional relationship, I was able to bypass some of the challenges involved in interviewing such senior stakeholders. For example, we organised the interviews through personal email

addresses and WhatsApp, and he invited me to the GPAI Annual Summit to meet him and his team.

3.4 Ethical Considerations and CUREC Approval

To ensure my research was conducted both ethically and legally, I completed the University of Oxford's research integrity course for social science researchers and the Central University Research Ethics Committee (CUREC) 1A form prior to commencing the research. In this form, I extensively documented my research design, the ethical considerations for each RP, and the research protocols that I would follow to ensure compliance with CUREC's Best Practice Guidance for Interviews and Internet-Mediated Research. My CUREC 1A form was reviewed and approved by the Oxford Internet Institute's Research Ethics Committee on behalf of the University of Oxford's Social Sciences and Humanities Interdivisional Research Ethics Committee. Below, I detail some of the key ethical considerations and protocols implemented across the RP.

3.4.1 Ethical Considerations for Research Interviews (RP1, RP2, and RP4)

In my interview-based research, I paid attention to ethical considerations like respondents' autonomy, informed consent, and data protection. For all interviews, respondents were provided with detailed information sheets explaining the research objectives and their rights. Informed consent was obtained in writing before each interview, with respondents given explicit choices about attribution and quotation preferences, including options for anonymisation, pseudonymisation, or direct attribution. A key ethical consideration was the sensitive nature of discussing topics like corporate strategies with company-affiliated developers who may not have been able to share certain information due to non-disclosure agreements or would not want information attributed to them for this reason. To address this, respondents were assured they could decline to answer any questions and were given control over how their responses would be attributed.

All interview recordings were captured on a PIN-protected device and promptly transferred to encrypted storage. Audio recordings were deleted immediately after transcription (within 1-3 days), with transcripts stored in encrypted folders on a password-protected computer. A master list linking respondent names to pseudonyms was maintained in a separate encrypted file and destroyed once no longer needed for analysis. For the questionnaire data for RP2, responses were initially collected through Google Forms but immediately downloaded and stored in encrypted files, with the original survey and data deleted from Google's servers. Email addresses and other contact information were deleted after the publication of the respective RPs.

3.4.2 Ethical Considerations for Data Mining (RP1 and RP3)

I mined data about development activity from GitHub and HF Hub. As discussed above, while the digital traces of development activity are recorded and made publicly available by the platforms, one still needs to mine and operationalise these digital traces into research data, and the choices made in this process involve ethical considerations. For example, for RP1, I collected development activity data from GitHub repositories (specifically, commits) as well as personal data (specifically, usernames and email addresses) in order to merge duplicate usernames and determine corporate affiliations. Once duplicate usernames were merged and affiliations identified, I deleted the personal data, ananonymised usernames o prevent individual identification, and ultimately aggregated individual developers to their company affiliations (e.g. developer@google.com to Google) in the analysis of company-to-company collaboration networks. For RP3, the research design minimised ethical risks by focusing on aggregate development patterns on HF Hub rather than individual behaviours. While usernames were collected to merge duplicate usernames and construct the collaboration network, the analysis focused on overall network structure without reporting individual identities. The study of model dependencies examined only public repository metadata and aggregated usage patterns.

3.4.3 Publication Ethics and Research Integrity

Throughout the research process, I paid careful attention to protecting respondent confidentiality in publications while maintaining research transparency. All interviewees were given the opportunity to review quotes attributed to them, and all respondents were pseudonymised (e.g. Respondent A). In reporting findings from the data mining components, care was taken to present results at a level of aggregation that prevented individual identification while preserving analytical validity. Furthermore, by requesting respondents to review their quotes and attributions, I ensured that the respondents were comfortable with the information I included in the respective RPs.

This chapter has provided a discussion of the ex-ante considerations and ex-post reflections regarding my research design. Next, Chapters 4–7 present the four empirical RPs. While they have been ordered to ensure thematic flow, please note that they were written and published independently and, therefore, they do not have a continuous narrative and should be read as independent RPs.

4. Characterising Open Source Co-opetition in Company-hosted OSS Projects: The Cases of PyTorch, TensorFlow, and Transformers

Peer review status: *This paper was published by the Proceedings of the ACM on Human-Computer Interaction (Track: Computer-Supported Cooperative Work and Social Computing).*

Co-authorship statement: *I am the first-author of this RP alongside Farbod Daneshyan, Runzhi He, Henzhe Ye, Dr. Yuxia Zhang, and Prof. Minghui Zhou. I document our relative contributions in Section 4.7.3. Please note that I use "we" rather than "I" due to the co-authorship of this RP.*

Abstract

Companies, including market rivals, have long collaborated on OSS development, resulting in a tangle of co-operation and competition known as "open source co-opetition." While prior work investigates open source co-opetition in OSS projects that are hosted by vendor-neutral foundations, we have a limited understanding thereof in OSS projects that are hosted and governed by one company. Given their prevalence, it is timely to investigate open source co-opetition in such contexts. Towards this end, we conduct a mixed-methods analysis of three company-hosted OSS projects in the AI industry: Meta's PyTorch (prior to its donation to the Linux Foundation), Google's TensorFlow, and HF's Transformers. We contribute three key findings. First, while the projects exhibit similar code authorship patterns between host and external companies (~80%/20% of commits), collaborations are structured differently (e.g. decentralised vs. hub-and-spoke networks). Second, host and external companies engage in strategic, non-strategic, and contractual collaborations, with varying incentives and collaboration practices. Some of the observed collaborations are specific to the AI industry (e.g. AI model integrations), while others are typical of the broader software industry (e.g. bug fixing or task outsourcing). Third, single-vendor governance creates a power imbalance that influences open source co-opetition practices and possibilities, from the host company's singular decision-making power (e.g. the risk of license change) to their community involvement strategy (e.g. from over-control to over-delegation). We conclude with recommendations for future research directions.

4.1 Introduction

Companies have participated in the collaborative development of OSS since the late 1990s (Broca, 2021), capitalising on a myriad of benefits, from cheaper development costs (Bonaccorsi & Rossi, 2006; Birkinbine, 2020) to open standards and interoperability (X. Li et al., 2024; Lerner & Tirole, 2005). In light of the increasing prevalence of commercial participation in OSS development, researchers have been encouraged to investigate the incentives, roles, and effects of such commercial activity on the norms, practices, and future trajectories of OSS developer communities (Germonprez et al., 2018). Within this area of research, one line of inquiry focuses on why and how companies collaboratively develop OSS (Germonprez et al., 2013; Y. Zhang et al., 2020), including market rivals and even companies that are engaged in patent wars against each other (Teixeira et al., 2016; Nguyen Duc et al., 2019). The term open source co-opetition has been coined to convey this tangle of co-operation and competition between companies (Teixeira, 2014).

To date, prior work on open source co-opetition primarily focuses on OSS projects that are hosted by vendor-neutral foundations, such as the OpenStack Foundation[1] (Y. Zhang et al., 2021; Teixeira et al., 2015, 2016), the LF (Germonprez et al., 2013), the Apache Software Foundation (Linåker et al., 2016), and the Eclipse Foundation (Wagstrom, 2009). Yet, prior work identifies the vendor-neutrality of such foundations as well as their open governance protocols as structural enablers for collaboration between "unexpected allies" (O'Mahony & Bechky, 2008), which limits the generalisability of prior work to OSS projects that lack such vendor-neutral governance, such as ones that are released, hosted, and governed by one company.[2] Given the prevalence and impact of company-hosted OSS project across software domains and industries, from web development to AI, it is timely to address this research gap and advance our understanding of open source co-opetition as one increasingly common type of commercial participation in OSS development.

We address this research gap through a mixed-methods analysis of open source co-opetition in three company-hosted OSS projects in the AI industry: Meta's PyTorch (prior to its donation to the LF in September 2022), Google's TensorFlow, and HF's Transformers. Many of the most popular OSS projects in the AI industry are company-hosted projects, making them suitable for this analysis. We examine three targeted RQ, guided by the objective of extending theory on open source co-opetition strategies and practices to the context of company-hosted OSS projects. First, through repository mining and SNA, we investigate the **(RQ1)** patterns and structures of collaboration be-

[1] N.B.: The OpenStack Foundation was renamed the Open Infrastructure Foundation in 2021.

[2] N.B.: The literature uses different nomenclature to describe this kind of OSS project: "company-hosted" (M. Zhou et al., 2016), "company-managed" (O'Mahony & Ferraro, 2007), and "company-sponsored" (West & O'Mahony, 2005). Throughout this study, we use "company-hosted" OSS projects to refer to OSS projects that are initiated, hosted, and governed by one company (Schaarschmidt, Walsh, & von Kortzfleisch, 2015).

tween companies, providing a baseline understanding of commonalities and differences between the three cases. Subsequently, through 10 semi-structured interviews, we investigate **(RQ2)** the types of collaborative relationships that host and external companies pursue and why, as well as **(RQ3)** what similarities and differences characterise open source co-opetition practices in company-hosted OSS projects compared to foundation-hosted projects, as identified by prior work.

We make three key contributions to the literature on open source co-opetition. First, while code authorship patterns by host and external companies are consistent across the projects over time (e.g. ~80%/20% of commits respectively), we observe varying structures of collaboration between companies on code authorship in project files (e.g. decentralised networks vs. hub-and-spoke networks), which hints at different configurations of control and collaboration in company-hosted projects. Second, we identify and characterise three distinct relationship types between host and external companies in company-hosted OSS projects: strategic, non-strategic, and contractual collaborations. Each type differs in the relevance of business strategy, competitive dynamics, and personal incentives for the involved developers. Some of the observed collaborations are specific to the technology and competitive dynamics in AI industry (e.g. hardware-software optimisation or AI model integrations), while others are typical of the broader software industry (e.g. bug fixing, code adoption or outsourcing of development). Third, single-vendor governance in company-hosted OSS projects introduces a power imbalance that influences open source co-opetition practices and possibilities, from the host's singular decision-making authority (e.g. the risk of license changes) to their community involvement strategy (e.g. from over-control of to over-delegation to the community).

The paper has the following structure. First, we discuss prior work on commercial participation in OSS development and open-source co-opetition (Section 4.2). Next, we present our research design (Section 4.3). Then, we report the results (Section 4.4). Then, we discuss the key implications of the findings and limitations (Section 4.5). Finally, we conclude the paper (Section 4.6).

4.2 Related Work

4.2.1 Commercial Participation in OSS Development

Companies have collaborated on the development of OSS since the late 1990s (Broca, 2021). The last decade, in particular, has seen a "rapid acceleration of corporate engagement with open source" (Germonprez et al., 2018), leading to what scholars have called "the incorporation of the digital commons" (Birkinbine, 2020) or the emergence of a "commons of capital" (Calimaq, 2018). In light of these developments, researchers have been encouraged to investigate the incentives, roles, and effects of commercial activity (Germonprez et al., 2018).

Companies participate in OSS development in various ways, which can be broadly categorised into three models: supporting, collaborating, and hosting OSS (M. Zhou et al., 2016). In the supporting model, a company assists an independently hosted project. This may include by deploying developers (Dahlander & Wallin, 2006), funding maintainers or the project (Osborne, 2024a), or joining project steering committees (Butler et al., 2018), among others. The collaborating model involves multiple organisations sharing control over the project's intellectual property. In the AI industry, this was exemplified by the joint release of the ONNX by Facebook and Microsoft in 2017, which focused on facilitating interoperability between multiple DL frameworks (Candela, 2017). In the hosting model, a single company exercises full control over a project's governance (Yue & Nagle, 2024) and intellectual property (M. Zhou et al., 2016), achieved by employing the maintainers (O'Mahony & Ferraro, 2007) and often requiring contributors to sign contributor license agreements (CLAs) (M. Zhou et al., 2016), among others. For this reason, O'Mahony refers to such projects as "company-managed projects" to underline that they are not "initiated and managed by a distributed group of individuals who do not share a common employer." Companies spin-out proprietary software projects into company-hosted OSS projects in order to increase adoption of the software, to benefit from external contributions, or to reduce a competitor's market share (cited in West & O'Mahony, 2005). As discussed in Chapter 2, spin-outs are common in the AI industry, including but not limited to Google and Meta's releases of their TensorFlow and PyTorch DL frameworks, which are often presented as acts of AI democratisation (Srnicek, 2022).

In company-hosted OSS projects, companies take different approaches to project governance and community involvement in line with their strategic goals (M. Zhou et al., 2016). While some companies maintain complete control of their project, which "resembles proprietary development conducted within a glass house," others strive to attract contributors, which involves significant investment into community development (West & O'Mahony, 2005). However, no matter how many resources a company may invest in its project, it is not guaranteed that a company can successfully build and retain a community of contributors in such projects because they are dominated by the company and lack aspects that attract external contributors, such as open governance or a meritocratic culture (Osborne, 2024c; Yue & Nagle, 2024). What is more, contributors may be suspicious of projects that "are viewed as transfers to the community to maintain code as opposed to collaborative partnerships" (West & O'Mahony, 2005). See Chapter 5 (RP2) for a deeper investigation of this phenomenon. Prior work also shows that the incentives and types of contributions to company-hosted OSS projects tend to be needs-based, compared to a combination of hobbyist and needs-based contributions community-led OSS projects (Shah, 2006).

Companies have diverse incentives for developing OSS, which differ from the primarily intrinsic

motivations of volunteer contributors (Benkler, 2006; von Krogh et al., 2012). These incentives encompass both strategic and social incentives. Strategic incentives include reducing development costs (Birkinbine, 2020; Crowston, Wei, Howison, & Wiggins, 2012), influencing open standards (Fink, 2003; Lerner & Tirole, 2002), and recruiting developers (Ågerfalk & Fitzgerald, 2008; Fink, 2003). Companies also seek vendor independence (Chesbrough, 2023; Lerner & Tirole, 2002), faster time to market (Ahlawat, Boyne, Herz, Schmieg, & Stephan, 2021; Chesbrough, 2023), and market competitiveness (Lindman et al., 2009; Loebbecke & Angehrn, 2003). Social incentives include reciprocating to the OSS ecosystem (Feller & Fitzgerald, 2002; Franck & Jungwirth, 2002) and improving corporate reputation as an OSS patron (Osterloh, Rota, & Kuster, 2003; Osborne, 2024a).

These diverse incentives highlight the multifaceted nature of commercial participation in OSS development. However, the collaborative nature of OSS development often leads to scenarios where companies, including market rivals, work together in OSS projects. This phenomenon, known as open source co-opetition (Teixeira, 2014), results in an interplay of co-operation and competition in the OSS context. In the following section, we delve deeper into the concept of open source co-opetition, exploring its definition, manifestations, and the current state of research in this area.

4.2.2 Open Source Co-opetition: Definition and Prior Work

Companies do not contribute to OSS development in isolation. They contribute to individual OSS projects as well as broader OSS ecosystems, which involve diverse contributors, from volunteers to companies (Teixeira & Lin, 2014). As a result, many OSS developer communities have evolved "from networks of individuals to networks of companies" (Ågerfalk & Fitzgerald, 2008, p.396).

The term "open-source co-opetition" has been coined to convey the tangle of co-operation and competition between companies in the OSS context (Teixeira, 2014). It draws on the "co-opetition" concept from management science (Brandenburger & Nalebuff, 1997; Dagnino, 2009), which contends that companies, which might be market rivals or even engaged in patent wars, form strategic alliances in areas that far from the customer, such as in R&D, whilst competing on revenue-generating products and services (Bengtsson & Kock, 2000). A key concept in this literature is that of ecosystems (Teixeira et al., 2016; Adner, 2006; Iansiti, 2004). In software engineering research, ecosystems are understood as "a set of businesses functioning as a unit and interacting with a shared market for software and services, together with the relationships among them" (Jansen et al., 2009, p.9).

The ecosystem concept views inter-company collaborations in OSS projects as an extension of common inter-organisational relationships, from joint ventures to contractual alliances, which aim to improve resource access and manage uncertainty (M.-J. Chen & Miller, 2015; W. H. Hoffmann, 2007). They are "access relationships" that provide access to the resources of companies (Stuart,

2000; Zineldin, 2004), to facilitate learning (Powell et al., 1996; Hamel, 1991), and to enable companies to improve their market position (Kogut, 1988; Stuart, Hoang, & Hybels, 1999) or shape industry standards (Gnyawali & Park, 2011; Brandenburger & Nalebuff, 2021). Co-opetition relationships are common in the high technology sector (Stuart, 1998; Dagnino, 2009) due to shorter product life cycles, the convergence of multiple technologies, and the rise in R&D expenditure (Gnyawali & Park, 2009). They often create strategic interdependencies between companies (Dagnino, 2009), which influence company behaviour and strategy (Gulati, 1998) and blur traditional organisational boundaries (Gilsing, Nooteboom, Vanhaverbeke, Duysters, & Van Den Oord, 2008).

Scholars of OSS development have adopted the co-opetition framework to investigate strategies and practices of inter-company collaboration in OSS development, both at the level of individual developers and companies (Nguyen-Duc et al., 2019; Teixeira et al., 2016). At the developer level, prior work finds that company-affiliated developers report little interest in which companies other contributors work for and typically view contributors from other companies as their peers (Nguyen-Duc et al., 2019). Furthermore, collaboration is characterised by low affiliation-based homophily, with frequent inter-company collaboration in OSS projects (Teixeira et al., 2015). Prior work also identifies two mechanisms of competition at the individual level. First, while multiple developers from a company may contribute to an OSS project, there is typically a gatekeeper—which may be one developer or a handful of developers—who coordinates a company's strategy, files issue reports, submits pull requests (PRs), and manages information flows (Nguyen Duc et al., 2017, 2019). The gatekeeper has the authority to decide what information or code the company will share with the project, and therefore he or she acts as a key lever for companies to engage in co-operative and competitive interactions simultaneously (Nguyen Duc et al., 2017; Nguyen-Duc et al., 2019). Second, the fork provides developers with the option to deviate at any time to pursue their own strategic goals or if they are not content with the management or direction of the project (Teixeira & Lin, 2014). The threat of the fork also encourages dominant contributors, such as companies, to restrain their influence and appease the interests of the wider community (Teixeira & Lin, 2014).

At the company level, a study on inter-company collaboration in OSS projects within the OpenStack ecosystem found that companies engage in intentional, passive, and isolated collaborations (Y. Zhang et al., 2020). Intentional collaborations are collaborations between companies that have a market relationship, such as the supply and consumption of OpenStack software or service provision (Y. Zhang et al., 2020). Passive collaborations occur when companies contribute to the same project without explicit coordination, while isolated collaborations take place when a company contributes to a project alone (Y. Zhang et al., 2020). Another study found that companies that share the same revenue model, such as offering complementary software or hardware, collaborated more than those

with different revenue models in the OpenStack ecosystem (Teixeira et al., 2015).

Prior work on open source co-opetition is primarily limited to OSS projects that are hosted by vendor-neutral foundations, including the OpenStack Foundation (Y. Zhang et al., 2021; Teixeira et al., 2015, 2016), the LF (Germonprez et al., 2013), the Apache Software Foundation (Linåker et al., 2016), and the Eclipse Foundation (Wagstrom, 2009). However, foundation-hosted projects have key characteristics that threaten the generalisability of prior findings to other project hosting and governance models. Through their vendor-neutrality and open governance protocols, foundations operate as "boundary organisations" that foster collaboration between diverse contributors, including volunteers and companies (O'Mahony & Bechky, 2008). They are reputed to foster "communities of competitors," where "market rivals...intentionally coordinate activities for mutual benefit in precise, market-focused, non-differentiating engagements" (Germonprez et al., 2013). We note that foundation-hosted projects do not shield projects from the dominance of companies (Wagstrom, 2009; Y. Zhang et al., 2021). For example, 10% of companies contribute 80% of commits and 20% of companies employ 80% of the contributors in the OpenStack ecosystem (Y. Zhang et al., 2021). Furthermore, IBM continued to dominate OSS projects hosted in the Eclipse ecosystem long after it established the Eclipse Foundation (Wagstrom, 2009). Commercial dominance can have negative consequences for the participation of volunteers (M. Zhou et al., 2016), which in part is due to the concern of performing free labour for the dominant company (Y. Zhang et al., 2018).

Beyond OSS projects that are hosted by vendor-neutral foundations, we have a limited understanding of open source co-opetition in OSS projects that are initiated, hosted, and governed by one company. To date, only two case studies have investigated open source co-opetition in such scenarios. A network analysis of collaboration in Google's Android simply highlights the dominance of Google developers and non-trivial contributions from its market rival Apple (Orucevic-Alagic & Host, 2014), while a study on Apple's WebKit demonstrates the methodological utility of temporal network visualisations for observing evolving collaborations (Teixeira & Lin, 2014). However, these studies fail to investigate open source co-opetition strategies and practices, nor do they consider how different governance approaches taken by host companies influence strategies and practices, from those that maintain control of development (M. Zhou et al., 2016) to those that adopt a more community-oriented approach (West & O'Mahony, 2005). Given the prevalence and impact of company-hosted OSS projects across the software industry, from web design to AI, it is timely to investigate the strategies and practices of open source co-opetition in such contexts, and ultimately to advance our understanding of the nature and impact of commercial participation in OSS development.

4.3 Study Design

4.3.1 Research Objectives and Research Questions

Our research objectives are two-fold: first, we seek to test prior theory on open source co-opetition practices in the context of company-hosted OSS projects; and second, to identify and characterise co-opetition practices that are unique to company-hosted OSS projects. Underlying these objectives is the following motivating RQ: How do companies collaborate on OSS development in OSS projects that are hosted and governed by a single company? We operationalise this motivating RQ by asking three targeted RQs, which we examine through a sequential, mixed-methods analysis of three case studies. First, through repository mining and SNA, we seek to understand **(RQ1)** typical structures and patterns, if any, of open source co-opetition in company-hosted OSS projects, providing a baseline understanding commonalities and differences between the cases. Subsequently, through 10 semi-structured interviews with company-affiliated contributors to the three projects, we seek to identify and characterise **(RQ2)** different types of collaborative relationships between host and external companies and their motivations, as well as **(RQ3)** the similarities and differences that characterise open source co-opetition in company-hosted OSS projects compared to foundation-hosted projects. This research design enables the testing of prior theory with mixed-methods findings from multiple cases (Easterbrook et al., 2008; Runeson et al., 2012), and thereby promises to enhance the convergence validity of the empirical findings (Runeson & Höst, 2008; Jick, 1979).

4.3.2 Multiple Case Study Research Design

4.3.2.1 Case Selection

We employed a four-step strategy to select cases. First, we defined the selection criteria: OSS projects had to be hosted by a company, involve external companies, and undergo active maintenance. Second, we selected the AI industry as the boundaries for case selection due to evidence of commercial investments (Ahmed et al., 2023; Whittaker, 2021) and involvement in the development of OSS and OMs (Langenkamp & Yue, 2022; White et al., 2024). Third, we prepared a "starting list" of company-hosted AI OSS projects by downloading a dataset of over 300 OSS projects from the LF AI & Data Foundation's website (LFAI&Data, 2022). We removed OSS projects with "data" labels, resulting in 184 AI OSS projects. Fourth, we labelled projects according to the type of its hosting organisation (company, non-profit foundation, or university/public research institute) and sorted the projects by the size of their contributor community. Then, we selected the three top-ranked projects in descending order: Google's TensorFlow, Meta's PyTorch, and HF's Transformers. Given Meta's donation of

PyTorch to the Linux Foundation in September 2022 (Meta, 2022), we limited data collection to this time point to ensure that all projects were company-hosted in our sample.

The selected cases—PyTorch, TensorFlow, and Transformers—represent different layers of the AI stack. TensorFlow (Google, 2023) and PyTorch (PyTorch, 2023b) are DL frameworks used to train models, while Transformers (HuggingFace, 2023b) provides higher-level APIs for downloading, fine-tuning, and sharing pre-trained models on HF Hub, a popular platform for hosting and developing AI models and datasets. Given their widespread usage in the AI industry and large contributor communities, they are promising cases for the study of open source co-opetition. In particular, TensorFlow and PyTorch make for interesting comparative cases because Google and Meta have long been in a heated rivalry over industry adoption of their respective DL frameworks (O'Connor, 2021). For example, OpenAI's choice of PyTorch over TensorFlow for its AI development was celebrated as a win for Meta (Thomas, 2020). Furthermore, the inclusion of Transformers enables a comparative analysis of projects hosted by industry giants (i.e. Meta and Google) and start-ups (i.e. HF), thus overcoming the limited focus of the two prior studies on OSS projects hosted by industry giants; i.e. Google's Android (Orucevic-Alagic & Host, 2014) and Apple's WebKit (Teixeira & Lin, 2014). We acknowledge the temporal cut-off in September 2022 and focus on major company-hosted OSS projects with >1,000 contributors as limitations, which we discuss in Section 4.5.

4.3.2.2 Case Presentation

We present the cases below (see summary information in Table 4.1).

TensorFlow by Google: TensorFlow is an open source DL framework that is widely used in academia and industry for creating and training ML models (Google, 2023). It was started by Google Brain in 2011 to facilitate the use of neural networks in Google research and products (Abadi et al., 2016). TensorFlow was publicly released in 2015, and TensorFlow 2 was released in 2019. After its initial release, Jeff Dean from Google stated, "We're hoping that the community adopts this as a good way of expressing ML algorithms of lots of different types and contributes to building and improving [TensorFlow] in lots of different and interesting ways" (Metz, 2015). Other reported incentives include increasing adoption, benefiting from crowdsourced innovation, and recruitment (Metz, 2015).

PyTorch by Meta: PyTorch is an open source DL framework widely used in academia and industry for training neural networks (PyTorch, 2023b). It was released in 2016 and maintained by Facebook AI Research at Meta until its donation to the LF's PyTorch Foundation in September 2022 (Zemlin, 2022). By 2022, PyTorch had overtaken TensorFlow as the most adopted DL framework in academic AI research (PaperswithCode, 2023) and for training OMs (Foster, 2022). Mark Zuckerberg has spoken publicly about the benefits that Meta has derived from the dominance of PyTorch in AI R&D,

in particular the crowdsourcing of AI innovations from researchers and developers in the PyTorch ecosystem and their integration in Meta's products and services (South Park Commons, 2024).

Transformers by HF: Transformers provides APIs and tools to download, fine-tune, and share pre-trained ML models (OMs) hosted on HF Hub (HuggingFace, 2023b). Transformers is integrated with TensorFlow and PyTorch, but operates at a higher level of the AI stack. It is developed by HF, a start-up, whose mission is to democratise AI by providing accessible tools and resources for researchers and developers. While HF initially developed a chatbot app, it is now better known for HF Hub, which hosts a fast-growing number of pre-trained models and datasets, and its OSS libraries (e.g. Transformers and Diffusers), which allow researchers and developers to download, modify, and share models and datasets hosted on HF Hub. In light of the emerging popularity of its tools and HF Hub, the start-up has raised hundreds of millions of US dollars in investment (Dillet, 2022).

4.3.3 Software Repository Mining

4.3.3.1 Data Collection

We mined data from each repository's commit logs on GitHub in order to analyse code authorship patterns and distributions. Specifically, we obtained historical commit data from each repository via the GitHub REST API, spanning from the date of the first commit in each repository until 12 September 2022, which we set as the data collection cut-off date to count PyTorch as a company-hosted project. Each commit dataset includes per commit: `sha`, `date`, `name`, `email address`, `modified source files`, and `lines of code (LOC)` added, `LOC deleted`, and `LOC changed (net)`. We acknowledge that while this data collection cut-off date predates major developments in the AI industry, it is defensible given that our study focuses on testing and extending theory on open source co-opetition in company-hosted OSS projects with insights from the AI industry, rather than focusing on trends in the AI industry per se. We discuss this temporal limitation in Section 4.5.2.

4.3.3.2 Username Merging

We merged multiple identities for unique developers, which is a common problem in software engineering research that arises when developers use multiple accounts on GitHub or due to how Git records their local credentials (Bird et al., 2006; Goeminne & Mens, 2013; Kouters et al., 2012; Robles & Gonzalez-Barahona, 2005). Following prior work (Zhu & Wei, 2019), we applied an automated approach. We built two bipartite networks that respectively mapped each username to all previously used corresponding email addresses and each email address to all previously used corresponding usernames. Then, we merged identities based on the linked username-email address

Table 4.1: Summary Information about PyTorch, TensorFlow, and Transformers

	PyTorch	**TensorFlow**	**Transformers**
Project Ownership			
Released by	Meta	Google	HF
Repository owned by	Meta	Google	HF
GitHub organisation	PyTorch	TensorFlow	HF
Company Information			
Company size	Large	Large	Small
Size (# employees)	~86,000 (Statista, 2022b)	~190,000 (Statista, 2022a)	~120 (Perez, 2022)
Market valuation (USD)	~450B (Cap, 2023b)	~1.4T (Cap, 2023a)	~2B (Dillet, 2022)
Project Information			
Initial release date	September 2016	November 2015	November 2018
License	BSD 3-Clause	Apache 2.0	Apache 2.0
# contributors	2,430	3,197	1,392
# GitHub stars	58,600	168,000	70,000
# commits	51,538	135,051	10,609
# forks	16,300	87,200	16,100
Resourcing by Company			
Employs maintainers	Y	Y	Y
Organises events	Y	Y	Y
Produces courses/tutorials	Y	Y	Y
Legal Control by Company			
Enforces CLA	Y	Y	N
Owns trademark	Y	Y	Y
Contribution Policy by Company			
Sets contribution policy	Y	Y	Y
Resolves CoC violations	Y	Y	Y
Encourages docs fixes	Y	Y	Y
Encourages issue reports	Y	Y	Y
Recommends first issues	Y	Y	Y
Encourages PRs	Y	Y	Y
Encourages PR reviews	Y	Y	NaN
Company Branding			
On documentation	N	N	Y
On project website	N	N	Y

N.B.: Project information as of 12 September 2022.

pairs and email address-username pairs. For analytical purposes, we created a unique user ID for each developer identity. This resulted in 3,434→3,058, 3,964→3,564, and 1,479→1,392 contributors in the PyTorch, TensorFlow, and Transformers contributor datasets respectively. Three authors cross-validated the accuracy of this approach, identifying 6, 3, and 2 errors in the respective datasets.

4.3.3.3 Bots Removal

In line with prior work (Y. Zhang et al., 2020; Robles & Gonzalez-Barahona, 2005; Lin, Robles, & Serebrenik, 2017), we removed commits by bots from the commit datasets. We identified bots through the following steps. First, we identified contributors with mentions of "bot" in their usernames and manually verified if they were a bot by searching them on GitHub. Second, we ranked contributors by their commit counts and identified bots among the 100 most active contributors per project. Through this process, we dropped 919 (1.78%), 38,503 (28.20%), and 23 (0.22%) commits from the PyTorch, TensorFlow, and Transformers commit datasets respectively.

4.3.3.4 Affiliation Identification

Following prior work (Y. Zhang et al., 2020, 2022), we applied a semi-automated approach to identify the affiliations of contributors at the time of each commit. We mined affiliations from the email addresses associated with each commit, which is considered the most accurate source of affiliation data (Nguyen-Duc et al., 2019; Mehra, 2011). The affiliations of commits with consumer email addresses, identified using a publicly available list (ihmpavel, 2022; Valiev, Vasilescu, & Herbsleb, 2018) or no email addresses, were left blank (i.e. NaNs). This identified the affiliations of 37.0%, 48.9%, and 9.2% of contributors in the PyTorch, TensorFlow, and Transformers datasets respectively. We discuss the divergence in affiliation identification per project in Section 4.5.2.

For contributors with missing affiliations, we mined affiliations from users' GitHub profiles. To address data quality concerns (e.g. self-reported affiliations are not time-sensitive to activity), three authors cross-validated affiliation data for contributors with 5 or more commits, and manually labelled missing values through Internet searches. We filtered out contributors who had submitted less than 5 commits to limit the analysis to contributors who met a minimum activity threshold and to reduce data labelling burden. Contributors who did not work for a company were recorded as "volunteers" and unidentifiable affiliations were recorded as "unknown." When contributors used both company and private email addresses, we linked all commits to their company affiliation.

Three authors reviewed 100 randomly sampled commits from each project to estimate agreement in the manual labelling. We found 5 errors, indicating 98.3% agreement. We could not cross-validate the entire dataset due to resource constraints, as finding the affiliation(s) of one contributor took up

to 10 minutes. We further evaluated the accuracy of the automated approach against the manually validated ground truth, finding the automated approach had labelled 83.9%, 92.0%, and 39.5% of commits correctly in the PyTorch, TensorFlow, and Transformers datasets respectively.

4.3.3.5 Descriptive Analysis

To partly answer RQ1, we report the provenance of commits per affiliation type (host company, external company, volunteer, unknown) in Figure 4.1; the dominance of host companies in Table 4.3; and the relative contributions of the top ten companies per project in Tables 4.4a-4.4c.

4.3.4 Social Network Analysis

4.3.4.1 Operationalising Collaboration on Code Authorship as a Social Network

We employed SNA to investigate open source co-opetition, drawing on prior work that used SNA to study collaboration among individuals (e.g. Crowston & Howison, 2005; Madey, Freeh, & Tynan, 2002; Singh, 2010) and companies (e.g. Linåker, Regnell, & Damian, 2020; Snarby, 2013; Teixeira et al., 2016). While some prior work focuses on discussions in issue trackers (Nguyen-Duc et al., 2019; Teixeira et al., 2016; Linåker et al., 2016), we operationalised collaboration as commits made by a pair of contributors to the same file during a release cycle for two reasons. First, commits represent an accurate, timestamped audit trail of code authorship (Orucevic-Alagic & Host, 2014; Y. Zhang et al., 2020); and second, examining code authorship per release enables longitudinal analysis of collaborations (Basole, 2009; Teixeira et al., 2016; Linåker et al., 2016). The second step concerned the choice of the unit of analysis. Since this study is concerned with co-opetition between companies, we chose the company as the unit of analysis in line with prior work (Nguyen Duc et al., 2017; Teixeira et al., 2016). However, we acknowledge that this aggregation loses crucial information about the activity of individuals (Dahlander & Wallin, 2006; M. Zhou & Mockus, 2010).

4.3.4.2 Network Construction

We recorded directed edges between pairs of contributors, who had contributed to the same file(s) during a release cycle, with edges weights corresponding to LOC changed in said file(s) by the respective contributor. For example, if contributor A modified 5 LOC in file F and contributor B modified 6 LOC in file F during the same release, the edge weights for A->B and B->A would be 5 and 6 respectively. The directed networks can therefore be formally represented as $G = (C, A_c, E, W_{ij})$, where C is the set of developers, E is the set of edges, A_c is the set of node attributes, and W_{ij} is the edge weight. We aggregated the release networks into annual snapshots to enable comparative

analysis (Long & Siau, 2007). Next, we assigned unique user IDs to nodes and excluded bots to limit the network to human-to-human collaboration. Then, we constructed company networks by merging developer nodes with the same affiliation, combining their edges, and summing their edge weights. We filtered out nodes with "volunteer" or "unknown" affiliations to focus on companies.

4.3.4.3 Network Analysis

We performed the network analysis in three steps. First, we measured three kinds of network centrality to understand different aspects about companies' roles in the collaboration networks (see Tables 4.4a-4.4c). Specifically, out-degree indicates a company's breadth of collaborations (NetworkX, 2023c), PageRank suggests its global importance (NetworkX, 2023d), and betweenness centrality reflects its brokerage role (NetworkX, 2023a). Second, we visualised annual network snapshots to observe changes in the collaboration relationships between companies (Teixeira & Lin, 2014) and the role of individual companies (Linåker et al., 2020). For readability, we filtered the networks to the 20 nodes with the highest degree centrality (Figure 4.2). Third, to account for network size effects in the former steps, we analysed three size-independent metrics of the complete networks over time (see Tables 4.5a-4.5c). Specifically, degree centralisation measures how much the network structure is organised around focal nodes (Freeman, 1978); degree skew indicates the asymmetry of the degree distribution, helping to identify the presence of hubs (Barabási & Albert, 1999); and the clustering coefficient quantifies the tendency of nodes to cluster together, helping to identify the presence of tightly-knit communities of collaborating companies (Watts & Strogatz, 1998). We report a single degree centralisation value rather than separate in- and out-degree centralisation values because the reciprocated edges result in identical centralisation values.

4.3.5 Semi-structured Interviews

4.3.5.1 Interviewee Sampling

We recruited 10 company-affiliated contributors for interviews (see Table 4.2). Our sampling approach involved sending personalised interview invitations to a subset of company-affiliated developers, who were not affiliated with the respective host company (Nguyen-Duc et al., 2019). In exchange for their time, we offered to donate 15 USD to a project of their choice. In total, we sent 350 emails to 150 TensorFlow contributors, 150 PyTorch contributors, and 50 Transformers contributors. We sent fewer emails to Transformers contributors due to fewer company-affiliated contributors in this project. We received 13 responses (3.71%) and 10 acceptances (2.86%).

Table 4.2: List of Respondents and Affiliations

ID	Sector	Company Size	Project(s)
A	Software consultancy	Small	PyTorch
B	Transportation, IT	Large, small	PyTorch, TensorFlow, Transformers
C	IT	Large	TensorFlow
D	IT	Medium	PyTorch
E	Transportation, IT	Large	TensorFlow
F	Software consultancy, IT	Small, large	PyTorch
G	E-commerce, IT	Small, small	PyTorch, Transformers
H	Software consultancy	Small	PyTorch
I	IT	Medium	TensorFlow
J	Software consultancy	Small, medium	PyTorch

4.3.5.2 Semi-structured Interviews

We conducted 10 digital, semi-structured interviews, which lasted between 30 and 60 minutes. The semi-structured interviews followed an interview guide with five topics: their personal and employer's incentives; their individual and employer's contribution strategies, if any; their experience of collaborating with developers employed by the host company and/or other companies; a discussion of the quantitative findings; and their views on the unique aspects of open source co-opetition in company-hosted OSS projects. During the interviews, we showed the network visualisations to the respondents to elicit responses about the evolving relationships between companies (Molina, Maya-Jariego, & McCarty, 2014; Hogan, Carrasco, & Wellman, 2007; Tubaro, Ryan, & D'angelo, 2016). We asked respondents to identify collaborations between companies that they were aware of, to explain their understanding of the nature and incentives for these collaborations, and to comment on temporal changes that were visible in the network visualisations (see Figure 4.2).

4.3.5.3 Thematic Analysis

We analysed the interview data following a systematic six-step procedure for thematic analysis; that is, the identification, analysis, and reporting of themes in qualitative data (Braun & Clarke, 2006). We adopted an integrated approach to code and identify themes in the interview data, combining deductive and inductive methods (Cruzes & Dybå, 2011). We used key findings from prior work as initial categories for the deductive coding, whilst inductively coding the interview data following grounded theory approaches to capture novel themes (Charmaz, 2006). This combination enabled us to both test prior theory and uncover new findings. The first author performed the initial coding until reaching saturation (Charmaz, 2006). A second author validated the codes to enhance the reliability of the analysis (Lincoln & Guba, 1985). Subsequently, the codes were merged into themes. Finally, we member-checked themes with respondents to ensure accuracy and relevance (Lincoln & Guba,

1985). When we quote respondents in Section 4.4, we identify them with their ID from Table 4.2 and mention their project(s) in abbreviated form in brackets for the reader's convenience.

4.4 Results

4.4.1 RQ1: What, if any, are typical structures and patterns of open source co-opetition in company-hosted OSS projects?

Key findings

The three projects reveal similar patterns of code authorship between host and external companies, yet distinct structures of collaboration. In each project, the host and external companies account for ∼80% and 20% of commits respectively. PyTorch and TensorFlow have decentralised network structures with lower degree centralisation, lower degree skew, and higher clustering coefficients, indicating strong inter-connections between companies. By contrast, Transformers has a hub-and-spoke network structure with higher degree centralisation and lower clustering, underlining HF's broker role between external companies in its project.

4.4.1.1 Distribution of Code Authorship by Host and External Companies

Host companies are dominant in their respective projects by several metrics (see Table 4.3). Meta, Google, and HF employ 61.25%, 47.61%, and 32.18% of contributors to their respective projects. These percentages increase in the maximal k-cores of the annual network snapshots, indicating the host companies' control over core development. For example, in the 2022 network snapshots, 31%, 50%, and 9% of contributors to PyTorch, TensorFlow, and Transformers respectively were affiliated to the host company, rising to 42%, 68%, and 38% in the network cores. Host company employees account for approximately 80% of annual commits, while external companies contribute 10-20% of annual commits (Figure 4.1). The Pareto principle is evident in each project, with less than 20% of contributors responsible for more than 80% of commits.[3] Transformers has the most imbalanced authorship, with 7.54% of contributors making 80% of commits, and has a low bus factor due to most commit activity coming from a few highly active contributors.

[3]The Pareto principle, commonly known as the 80/20 rule or the law of the vital few, states that approximately 80% of effects come from 20% of causes (M. Newman, 2005).

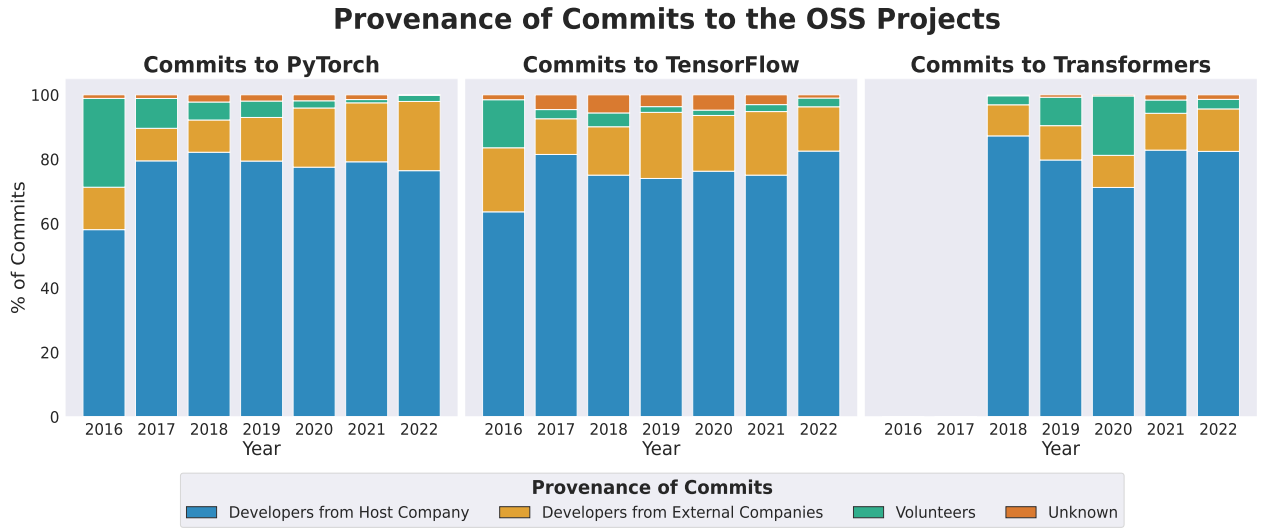


Figure 4.1: Provenance of Commits to PyTorch, TensorFlow, and Transformers

4.4.1.2 Collaboration Networks: Centralised vs Decentralised Collaboration

We observe distinct patterns of collaboration across the three projects, with PyTorch and TensorFlow exhibiting decentralised structures despite dominant code authorship by the host companies, while Transformers shows a hub-and-spoke structure. In PyTorch, while Meta contributes the majority of commits (84%) and lines of code (84%), many commits are from Nvidia, Intel, AMD, and Google (see Table 4.4a). Similarly, in TensorFlow, Google dominates in commits (85%) but contributes a smaller share of lines of code (34%), with significant contributions from Nvidia, Intel, and IBM, among others (see Table 4.4b). Despite this concentration of authorship, external companies have high out-degree centrality values in both projects, indicating active collaboration on project files among these companies. However, Transformers presents a contrasting picture. HF not only dominates in code authorship (91% of commits, 94% of lines of code) but also in network centrality measures (see Table 4.4c). External companies have low out-degree and PageRank centrality, indicating limited breadth of collaborations on project files and their global network importance, while HF has high betweenness centrality, indicating its pivotal role as a broker between external companies.

The network visualisations underscore these structural differences (see Figure 4.2). PyTorch and TensorFlow display decentralised collaboration with dense collaboration between various companies, while Transformers exhibits a hub-and-spoke structure, with HF playing a broker role between external companies on the periphery of the network. These differences persist despite variations in network size across projects and over time (see Tables 4.5a-4.5c), corroborating the observed differences in Figure 4.2. For example, the Transformers networks exhibit higher degree centralisation and degree distribution skew compared to PyTorch and TensorFlow, indicating a collaboration structure organised around a focal company (i.e. HF). By contrast, the PyTorch and TensorFlow networks

have higher clustering coefficients, suggesting the presence of more decentralised, interconnected communities of companies that collaborate on code authorship in these projects.

4.4.2 RQ2: What Types of Collaborative Relationships Do Companies Pursue in Company-hosted OSS Projects, and Why?

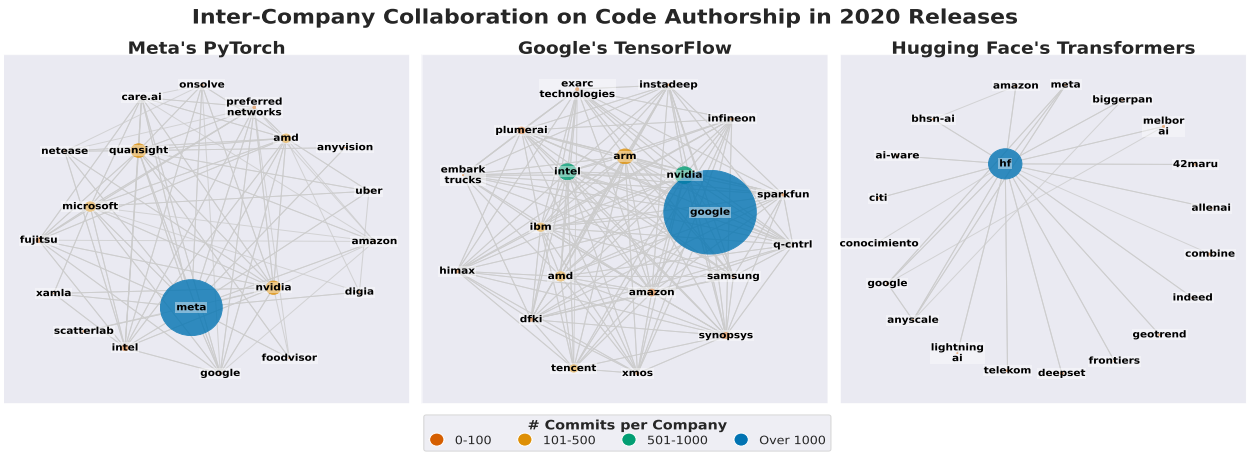
Key findings

Host and external companies collaborate in, at least, three types of relationships in company-hosted OSS projects: strategic, contractual, and non-strategic collaborations. Strategic collaborations are primarily dyadic relationships between the host company and an external company, often characterised by private collaborations and driven by competitive incentives. Contractual collaborations involve the outsourcing of development to third-party companies. Non-strategic collaborations encompass a range of contribution types and motivations, such as hobbyism, bug-fixing, code adoption, and corporate OSS initiatives, which blur the line between voluntary and work-based contributions by company-affiliated contributors.

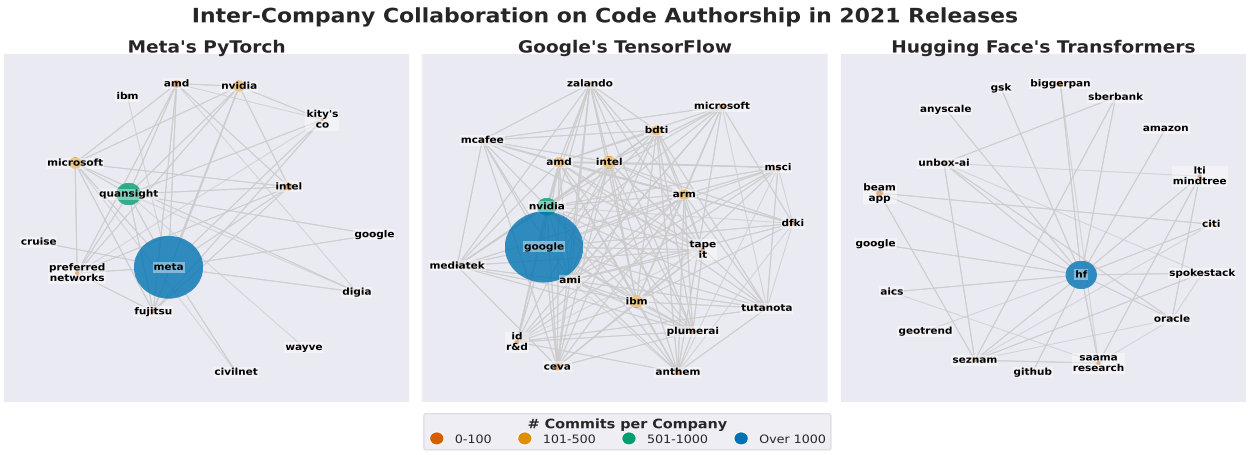
4.4.2.1 Strategic Collaborations

The first type of inter-company collaboration that we observe concerns strategic collaborations between the host company and an external company. In each of the three projects, the host company has engaged in strategic partnerships with external companies, such as AI accelerator manufacturers, cloud service providers, and AI model producers, where business objectives and competitive dynamics have been paramount. Contrary to prior work, we find that business strategy plays an important role at the developer level in such scenarios. For example, competing AI accelerator manufacturers and cloud service providers pursue strategic collaborations with Meta and Google to ensure compatibility between new releases of the frameworks and their respective hardware or cloud services. Respondent F (PT) explained that their company had clear goals on what they wanted to achieve in PyTorch, and their team's contributions were "narrowed down" accordingly. In regular closed-door meetings with the PyTorch maintainers, which were held privately for select company-affiliated developers to protect proprietary information, they would exchange information about their respective upcoming hardware and software releases and priorities. Since it was also in the interest of Meta to have PyTorch run efficiently on their AI accelerators, Respondent F (PT) contended that this relationship could be characterised as a strategic interdependence between the companies.

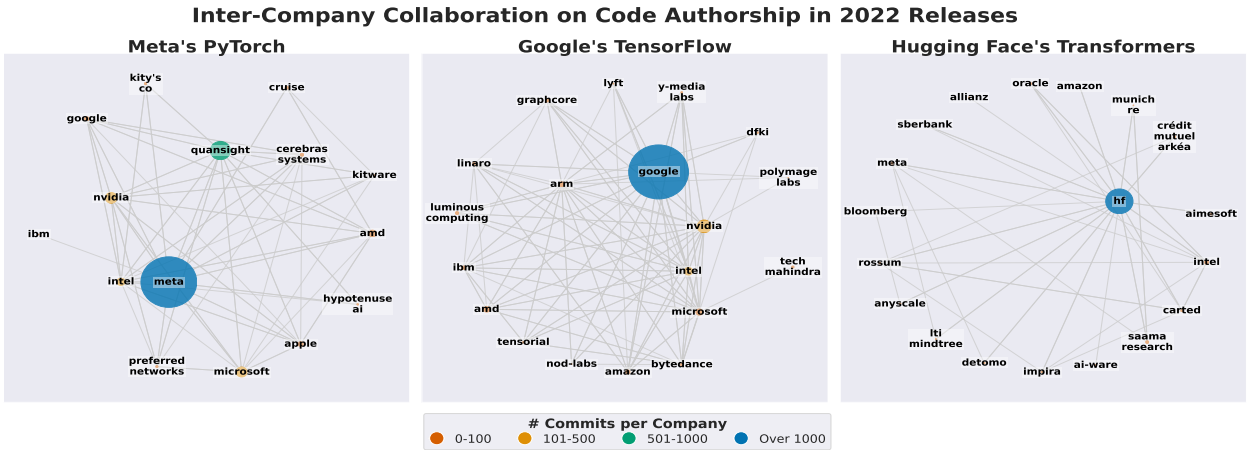
Competitive logics between market rivals are key to these strategic collaborations. Respondent B (PT, TF, TR) explained that it is a strategic priority for the AI accelerator manufacturers to minimise



(a) Inter-Company Collaboration on Code Authorship in Repository Files in 2020 Releases



(b) Inter-Company Collaboration on Code Authorship in Repository Files in 2021 Releases



(c) Inter-Company Collaboration on Code Authorship in Repository Files in 2022 Releases

Figure 4.2: Open Source Co-opetition Networks in PyTorch, TensorFlow, and Transformers

the risk of being undercut by a market rival if these popular DL frameworks do not run efficiently on their hardware. Respondent F (PT) agreed, explaining, "When you're contributing to PyTorch on behalf of the company, you have to think about how it makes profit for the company. So, our very first goal was to make sure that PyTorch works well on our GPUs." Similarly, respondents discussed the self-interest of Google to collaborate with AI accelerator manufacturers, despite also being in the business of selling AI accelerators (i.e. via its Tensor Processing Units) to ensure optimal performance of TensorFlow on other hardware offerings and vice versa, especially "Nvidia's GPUs which are the market leader," in order to avoid being undercut by PyTorch. Commenting on Figure 4.2, Respondent D (PT) asserted that these strategic collaborations are ultimately about revenue generation: "AWS have an incentive because everyone wants to buy their servers to train their models and people are buying Nvidia's cards to train their neural networks. So, their incentive is to indirectly boost their sales by helping make PyTorch run really well." Contrary to prior work, business strategy evidently influences the activity of developers who contribute on behalf of their employers.

While Figure 4.2 does not show comparable evidence of open source co-opetition in Transformers, Respondent G (PT, TR) explained that market rivals in the cloud market like Microsoft Azure, AWS, and Google Cloud have become active contributors to Transformers as maintainers of integrations with their respective cloud offerings since the end date of our data collection in late 2022. They explained that HF engages in strategic partnerships with AI companies like Meta and Stability AI to offer day-zero integrations for their OM releases. They explained, "That's something we need to be active upon, we can't really delegate that to the community to be able to, you know, quickly sprint on it and quickly deliver because there's a lot of advantages of being the first movers here."

These private collaborations create a layer of strategic interaction that exist alongside, but separate from, the broader community. As a result, it can be difficult for developers that are not affiliated with one of the companies to participate in these strategic collaborations. For example, Respondent F (PT) explained that they had found it difficult to contribute to issues or PRs related to request for comment (RFC) documents submitted by AI accelerator manufacturers since they referred to undisclosed proprietary information. They commented, "There are a lot of things that you will just not know as an outsider." However, Respondent C (TF) explained that due to the interdependence between the DL frameworks and AI accelerators, third-party companies do collaborate with the corresponding companies in some cases, albeit without access to their private communications or meetings, such as by reporting issues and submitting PRs on hardware-related features.

4.4.2.2 Contractual Collaborations

The second type of inter-company collaboration concerns contractual collaborations between the host company and an external company to which the host company outsources development tasks. In PyTorch, Meta outsources development to QuanSight, a consultancy that focuses on the Python data science stack. After Meta, QuanSight has made the most contributions to PyTorch (see Table 4.4a). Respondent A (PT) explained that their team contributes to a range of core modules in PyTorch: "We're not helping them build a particular product, we're just helping them build PyTorch as a project." Due to the contractual nature of the collaboration between Meta and QuanSight, QuanSight developers support Meta developers with the realisation of their strategic goals in PyTorch, "which are essentially for PyTorch to be cutting edge and performing well across the board." Respondent H (PT) speculated that Meta had contracted this work out to leverage their know-how and to reduce costs since outsourcing was cheaper than paying the salaries of Meta engineers.

While the contractual nature of the relationship between Meta and QuanSight involves closed-door communication and collaboration, Respondent A (PT) commented that Meta grants QuanSight developers the freedom to contribute as if they were organic contributors. Since their mandate is not focused on specific features, they collaborate with all kinds of contributors, including volunteers and company-affiliated contributors. They explained that QuanSight contributors mostly work independently: "Maybe once or twice a week, they'll check in and report up to the project managers what they've been doing, but essentially it's up to them to take the initiative to find higher priority issues and just make sure that they get resolved." They added that they also deploy small teams of developers to work with a lead engineer from Meta, who gives them direction. Given how closely QuanSight and Meta employees collaborate in PyTorch, they suggested that it can be hard to distinguish between Meta and QuanSight employees. "You know, if you were to sit over their shoulder and watch them day to day without any idea of what company they work for, you would probably just assume that they were all one big team." This contractual collaboration exemplifies a hybrid model, where contracted developers are granted autonomy while still needing to align with the host company's goals, thus blurring the lines between internal and external contributors.

4.4.2.3 Non-strategic Collaborations

The third type of inter-company collaboration that we observe concerns non-strategic collaborations between the host company and developers from external companies, who have a number of incentives for contributing to company-hosted OSS projects, including their personal interest, bug-fixing, and OSS contribution initiatives at work. For example, Respondent F (PT) explained they got in-

volved in PyTorch following a summer internship at an academic lab, which sparked their interest in DL and led them to "dive deeper into the PyTorch code base." Meanwhile, four respondents explained that they had contributed to projects to fix bugs they encountered at work. However, Respondent B (PT, TF, TR) explained that while they made fixes to TensorFlow for work purposes, they underlined that there was no direction from their employer. They contrasted these contributions with strategic contributions their start-up now makes to PyTorch when they encounter issues that are specific to the company's use cases and not "in the top of priority list for the PyTorch maintainers."

The lines are blurred about whether such contributions should be considered as voluntary or company-affiliated contributions. In fact, the distinction between voluntary and company-affiliated contributions was not a relevant consideration for several respondents, who had engaged in such non-strategic collaborations. For example, Respondent D (PT) explained that even though they use their company-affiliated GitHub account for all of their activity, they did not attribute this activity to their company because "OSS is a hobby of mine." Their principal contribution to PyTorch was the improvement of its implementation of xdoctests, a Python package for executing tests in documentation strings that they maintain. Since then, they explained that PyTorch maintainers reach out from time to time to get their advice on related problems, which they help out which on a voluntary basis. They explained, "If I'm not working on a project or something explicitly beneficial for my company, I don't charge my time." Cases as such demonstrate the importance of hobbyism among company-affiliated developers who contribute to company-hosted OSS projects.

Corporate OSS contribution initiatives and personal code adoption goals were also mentioned as non-strategic reasons. Respondent C (TF) explained that they began contributing to TensorFlow through an OSS contribution initiative by their employer, which "started out as a side hobby" but they got more involved when a maintainer invited them to collaborate on API migrations for tf.data. They used a restricted-access Google document with the list of APIs to migrate, but "the conversation was mostly through PRs on GitHub." While they contributed during working hours, they received no direction from their employer. Meanwhile, Respondent B (PT, TF, TR) contributed a modified version of research code, initially developed for a scientific paper and published with a non-commercial use license, to the Transformers repository to increase its adoption and impact. These non-strategic collaborations highlight the complex interplay between personal interests, professional development, and corporate initiatives in driving contributions to company-hosted OSS projects, often transcending clear-cut categorisations of voluntary or company-affiliated work.

4.4.3 RQ3: What Similarities and Differences Characterise Open Source Co-opetition in Company-hosted OSS Projects compared to Foundation-hosted OSS Projects?

Key findings

There are both similarities and differences in open source co-opetition in company-hosted OSS projects compared to foundation-hosted OSS projects, as identified by prior work. On the one hand, prior findings about the role of gatekeepers, the competitive use of forks, and non-competitive attitudes among developers apply to a certain extent in the three projects. On the other hand, single-vendor governance introduces a power imbalance that affects open source co-opetition practices and possibilities in company-hosted OSS projects, from the risks of a host company's unilateral decision-making power (e.g. to change the license) to their community involvement strategy (e.g. from over-control to over-delegation to the community).

4.4.3.1 The Limited Role of Gatekeepers as Facilitators of Open Source Co-opetition

Prior work highlights the role of developers, who act as a gatekeeper on behalf of their company, as key facilitators of open source co-opetition. However, we observe that developers perform gatekeeper roles to coordinate contributions made on behalf of their company in some but not all cases in the three projects. For example, in strategic collaborations, lead developers or engineering managers coordinate closed-door meetings and information sharing. However, respondents contested the universality of gatekeepers. Respondent F (PT) asserted, "There are always gatekeepers, right? The PyTorch team was just like any other team at [company], we always had a tech lead, we always had a manager, it's not like we were operating independently. They smoothen the communication between the two companies." They explained that while engineering managers were key coordinators and communicators in their biweekly meetings with Meta, junior developers were given ample opportunity to communicate their views "and they were heard." The gatekeeper role was even less important in contractual and non-strategic collaborations. For example, Respondent C (TF) "did not have a specific direction or requirements from [company]" during their contributions to TensorFlow, while Respondent H (PT) enjoyed freedom to work on wide-ranging PyTorch issues due to QuanSight's contractual agreement with Meta. These findings complicate prior theory on the indispensable role of gatekeepers in facilitating open source co-opetition, providing critical nuance on the relevance of gatekeepers in different types of collaborations (Nguyen Duc et al., 2017, 2019).

4.4.3.2 Attitudes towards "Competitors" among Developers

We find that prior findings about non-competitive attitudes among company-affiliated developers apply to a certain extent in the three projects (Nguyen Duc et al., 2017; Nguyen-Duc et al., 2019). Respondents, who had participated in non-strategic and contractual collaborations, did not view contributors from other companies, including market rivals, as competitors. Respondent H (PT) commented that the affiliation of contributors was not important to them, citing their philosophy that "everyone grows during collaboration, but not everyone grows during competition." Respondent C (TF) commented that their manager was not concerned that they were contributing to a Google project: "It was the other way around. They were appreciative that I was contributing to an open source project, which we used extensively." Furthermore, Respondent A (PT) argued that since PyTorch is a core library, they and their team do not have to navigate the competitive dynamics that a product team might face if they were to collaborate with a product team from another company. Several respondents suggested that competitive attitudes are the strongest when several market rivals are engaged in strategic collaborations with the host company at the same time. For example, Respondent F (PT) explained, "There was always competition between AMD and Nvidia... if AMD had a bug because of which our task was slowing down, we would pitch to Meta that whatever AMD was doing, they [should] only touch their codebase and [not] affect our tasks and slow us down." These examples illustrate the presence of competitive attitudes among developers, thus contesting prior theory (Nguyen Duc et al., 2017; Nguyen-Duc et al., 2019).

4.4.3.3 Different Approaches to Community Involvement by Host Companies

The respective host companies have taken different approaches to community involvement in their projects, which has shaped collaboration practices and possibilities. Numerous respondents remarked that Meta and HF take community-driven approaches. For example, Respondent B (PT, TF, TR) commended the "back and forth" discussions with PyTorch maintainers, and Respondent G (PT, TR) highlighted the open RFC process in Transformers where "anyone from the community is encouraged to take part in the discussions." Respondents contrasted the friendliness of PyTorch and Transformers maintainers with their transactional interactions with TensorFlow maintainers. Respondent D (PT) commented, "If I didn't know that PyTorch was from Facebook, I don't think I ever would have figured that out, whereas with Google it's kind of obvious." However, community-oriented approaches can also backfire for host companies, when they fail to distribute decision-making in their projects. For example, Respondent B (PT, TF, TR) shared that they thought HF were "not really espoused to open source philosophically, [rather] they're espoused to it as a means to an

end" and they were "reaping all the benefits" from the community. They explained that when the start-up changed the license of its text-generation-inference repository without community deliberation, their company decided to stop contributing to HF's OSS projects to avoid future risks.

By contrast, respondents explained that Google takes a top-down approach to TensorFlow, which creates obstacles for community involvement and collaboration. For example, Respondent C (TF) explained, "[Google] controls the level of granularity that other people should have access to. If there is something specific to Google, then they don't release it." Respondent C (TF) also mentioned that Google develops TensorFlow privately through an internal RFC process, which is out-of-bounds to the broader contributor community. Similarly, Respondent E (TF) stated, "I didn't know what was on Google's mind about what was most important and I certainly didn't have a community feeling." Moreover, Respondent C (TF) explained that during their time collaborating with a TensorFlow maintainer on tf.data, they had limited awareness of the project's overall roadmap: "For tf.data, I understood what we were doing and why these migrations were necessary, but not the roadmap for TensorFlow as a whole. That is too complicated and not visible to outsiders." This lack of transparency, in turn, creates barriers for external contributors, who are blind to long-term roadmap priorities and may feel subordinate in the project's social hierarchy.

4.4.3.4 Anticipating and Managing the Risk of License Changes

Respondents underlined that a host company's ability to unilaterally change the license of its project without community deliberation is a key differentiating factor of company-hosted OSS projects. Respondent G (PT, TR) explained that one must be prepared for the event that the host company will change the license, while Respondent F (PT) commented that projects hosted by start-ups generally have higher risks than those hosted by industry giants because start-ups are under greater pressure to generate revenue from their OSS projects. Respondent B (PT, TF, TR) explained that HF had changed the license of its text-generation-inference project from Apache v.2 to a restrictive license, which was problematic for their commercial use of the project. He explained that in response a handful of companies, including their start-up, forked the project. While the fork allowed external companies to continue their collaboration on the project, they explained that "this would never have happened at a foundation project." Similarly, Respondent H (PT) remarked, "the distribution of power is one thing that very much supports open source projects in general. However, if one company is in control, even if it's managed nicely, you're going to drive a lot of companies away just as a matter of principle or policy." Thus, the uneven playing field between host and external companies, including the risk of a sudden license change without community deliberation, is a unique aspect of company-hosted OSS projects that influences open source co-opetition practices.

4.5 Discussion

In this section, we discuss the implications of our findings for research on open source co-opetition (see Section 4.5.1), as well as the threats to the validity of our findings (see Section 4.5.2).

4.5.1 Theoretical Contributions

We set out to understand how companies collaborate on OSS development in the absence of vendor-neutral governance, specifically in OSS projects that are hosted and governed by one company. The findings from PyTorch, TensorFlow, and Transformers reveal collaboration strategies and challenges that both align with and diverge from prior theory on open source co-opetition.

4.5.1.1 Host Company Dominance and Divergent Collaboration Structures

The findings both confirm and contest findings from prior work. On the one hand, as expected, the host companies dominate code authorship in respective their OSS projects, contributing around 80% of commits across the projects and over time, confirming prior work on commercial dominance in company-hosted OSS projects (M. Zhou et al., 2016; Orucevic-Alagic & Host, 2014). On the other hand, our analysis of file-based collaboration reveals varying structures of inter-company collaboration in company-hosted OSS projects, from decentralised collaboration in PyTorch and TensorFlow to centralised collaboration in Transformers. The hub-and-spoke structure observed in Transformers is noteworthy, as HF publicly champions itself as "the AI community building the future" (HuggingFace, 2023a). However, we note that these variations simply reveal differences in how developers from various companies have interacted on project files through commits. These differences may reflect or be due to various project-specific factors, such as technical architecture, the host company's contribution policy, or specific collaboration dynamics among companies. We encourage future research to investigate reasons for different collaboration structures to better understand this phenomenon.

4.5.1.2 Contributions to the Open Source Co-opetition Literature

Our findings extend the literature on open source co-opetition by both testing prior theory and contributing empirical findings from the context of company-hosted OSS projects.

First, we find that prior theory regarding the key coordinating role of gatekeepers (Nguyen Duc et al., 2017; Nguyen-Duc et al., 2019), the strategic use of the fork as a competitive mechanism (Teixeira & Lin, 2014), and non-competitive attitudes among individuals developers from rival companies (Nguyen Duc et al., 2017; Nguyen-Duc et al., 2019) generalise to the three projects in our sample to a certain extent. However, our findings shed light on important nuances. First, we find

that the coordinating role of the gatekeeper is not universally salient to open source co-opetition. While stakeholders like team managers play a coordinating role in strategic and contractual collaborations, such as by coordinating meetings between companies, respondents emphasised that they have enjoyed significant freedom when they have participated in such collaborations. Gatekeepers are even less relevant in non-strategic collaborations, where the lines are often blurred between voluntary and work-based contributions by company-affiliated developers. Moreover, contrary to prior findings (Nguyen-Duc et al., 2019), we find that business strategy is indeed important at the developer level in some cases, in particular in strategic and contractual collaborations. What is more, while the respondents explained that in general they do not view contributors that are affiliated with market rivals as competitors (Nguyen-Duc et al., 2019), competitive dynamics between rival companies, such as AI accelerator manufacturers, do influence interactions and priorities at the developer level. These findings contribute to a more nuanced understanding of open source co-opetition.

In addition, we contribute to open source co-opetition theory by characterising strategic, contractual, and non-strategic collaborations as three distinct collaborative relationship types between companies, which vary in terms of their collaboration practices and incentives. This typology builds upon the prior categorisation of intentional and passive collaborations between companies (Y. Zhang et al., 2020), offering more granularity about the role of business strategy and collaboration practices in different types of collaborations between companies. In particular, our findings highlight that strategic collaborations, which often involve closed-door meetings and private communications between the host company and select external companies, create a private layer of interaction that exists alongside, but separate from, the broader community. This dual nature of collaboration— public and private— presents unique challenges for outsiders. Developers who are not affiliated with collaborating companies find it difficult to contribute to certain issues or pull requests, particularly those related to proprietary information or undisclosed features. This dynamic highlights a potential tension between the open source ethos of transparency and the strategic needs of companies engaged in open source co-opetition. Future research should explore how this balance between openness and strategic privacy affects community dynamics, contributor motivation, and overall project sustainability. Furthermore, as the Meta-QuanSight relationship was the only observed contractual collaboration in our sample, further research is needed to validate the characteristics of contractual collaborations across a broader range of OSS projects, company profiles, and sectors.

We also contribute novel insights into the unique aspects of open source co-opetition in the context of company-hosted OSS projects. In particular, single-vendor governance in company-hosted OSS projects creates a power imbalance between the host company and external contributors, leaving external contributors dependent and vulnerable to the decision-making and goodwill of the host

company. Furthermore, we find that a company's approach to community involvement influences external companies' willingness to contribute, with both over-control and over-delegation potentially hindering collaboration. This extends prior work on the effects of commercial dominance (M. Zhou et al., 2016; Y. Zhang et al., 2022), demonstrating that excessive task delegation, in addition to excessive dominance, can also deter contributors. While our findings advance understanding of open source co-opetition in company-hosted OSS projects, several questions remain. Since our findings are based on a limited sample of projects ($n = 3$), further research is needed to validate these findings across a wider range of projects, companies, and sectors. Future research could also examine the impact of governance changes when projects transition from commercial hosting to vendor-neutral foundations, providing insights into the relationship between project governance and collaboration dynamics between companies (Yue & Nagle, 2024; Osborne, 2024c).

4.5.1.3 Open Source Co-opetition in the AI Industry

The findings shed light on open source co-opetition dynamics specific to the AI industry, which manifest in, at least, three kinds of strategic collaboration. First, hardware-software optimisation is a critical area of strategic collaboration, with Google and Meta engaging in strategic partnerships with AI accelerator manufacturers such as Nvidia, Intel, and AMD. These collaborations aim to mutually optimise framework performance and hardware capabilities, even in cases where the companies are market competitors. Second, cloud integration is another area of strategic collaboration, with partnerships formed between framework maintainers and cloud service providers like AWS, Google Cloud, and Microsoft Azure. These collaborations seek to ensure seamless deployment and optimal performance of frameworks and AI models in cloud environments, reflecting the growing importance of cloud infrastructure for AI development and deployment. Third, AI model integration has become an important area of collaboration, exemplified by HF's collaborations with AI companies like Meta AI and Stability AI, which aim to offer day-zero integrations for new model releases, capitalising on first-mover advantages and maintaining competitive edges in the rapidly evolving AI industry.

We acknowledge that our quantitative findings are temporally limited to late 2022, predating major industry developments like the launch of ChatGPT (OpenAI, 2022), the release of open foundation models like Meta's Llama models (Touvron et al., 2023), and grassroots initiatives like the BigScience Workshop (Akiki et al., 2022). Given the pace of change in the AI industry, it is likely that the population of contributing companies and the collaboration structures within the projects have evolved since 2022. However, the nature and extent of these changes may vary across different layers of the AI stack. For example, at the DL framework level, since Meta's donation of PyTorch to the LF, the project has been governed by a governing board comprising representatives from AMD,

AWS, Google Cloud, HF, IBM, Intel, Meta, Microsoft, and Nvidia (Osborne, 2024c). This governance change has not resulted in significant net increases in contributions, but an increase in contributions by AI accelerator manufacturers and a decrease by Meta (Yue & Nagle, 2024).

By contrast, Transformers has likely seen more significant developments to its community of contributors. The growth of HF Hub as the *de facto* platform for sharing and hosting OMs, has likely influenced participation in the development of Transformers, which provides APIs to download, use, and share models on HF Hub. For example, the interviews revealed that since our quantitative data collection, HF has engaged in strategic partnerships with AI model providers like Meta AI and Stability as well as cloud compute providers like Microsoft Azure, AWS, and Google Cloud, who have become active contributors to the project as maintainers of their respective integrations. Furthermore, Transformers' integration with multiple DL libraries (Jax, PyTorch, and TensorFlow) and its role in facilitating access to OMs hosted on HF Hub likely position it at the centre of evolving collaboration between industry giants in the hardware and cloud sectors in the OSAI ecosystem.

4.5.2 Threats to Validity

4.5.2.1 Construct Validity

Construct validity concerns the extent to which the measurements accurately represent the phenomenon under study (Easterbrook et al., 2008). We faced the following challenges. First, while we operationalised commits to common files per release as a proxy for collaboration, we acknowledge this captures only one type of observable collaboration in OSS repositories among other types of collaboration (Casari et al., 2023). Moreover, by aggregating network data, we sacrificed granularity to enable a year-to-year comparative analysis of collaboration. Second, identifying the affiliations of individual contributors was challenging. Domain-name mining from email addresses only identified affiliations for 37.0%, 48.9%, and 9.2% of contributors in PyTorch, TensorFlow, and Transformers respectively. These low numbers may be due to the use of Github privacy features and personal email addresses, as well as the prevalence of contributors who are indeed volunteers (especially in Transformers). Manual labelling was resource-intensive and may have introduced errors (e.g. by assigning a company affiliation rather than volunteer affiliation). Third, the SNA may have been influenced by the different network sizes. To account for such effects, we examined size-independent network metrics across the projects and over time, which corroborated the SNA findings. Fourth, recruitment challenges limited our interview sample ($n = 10$ in total, $n = 2$ for Transformers). Nonetheless, the interviews provided insights from various seniority levels and company profiles, and positive feedback on the quantitative findings increased our confidence in their validity.

4.5.2.2 External Validity

External validity concerns the generalisability of findings. We acknowledge three key threats. First, the case study research design limits the generalisability of the findings. However, we underscore that the objective was to test theory, and the case study research design was suitable for this objective (Yin, 2018). Furthermore, the inclusion of Transformers expanded the analysis beyond industry giants' projects, a limitation of prior work. Second, the findings are based on a limited sample of projects ($n = 3$) and interviews ($n = 10$) and may not be exhaustive or universally applicable across all company-hosted OSS projects. Further research is needed to validate these findings across a wider range of projects, companies, and sectors. Furthermore, the interview sample does not capture the diversity of views of the broader population of contributors or senior decision-makers. Nonetheless, the sample included various seniority levels, from junior and mid-career developers to start-up founders, as well as various company profiles, from start-ups to industry giants, thus providing a diversity of perspectives. Third, the quantitative findings were temporally limited to September 2022, predating developments like the launch of ChatGPT (OpenAI, 2022), OMs like Meta's Llama models (Touvron et al., 2023), or the rising popularity of HF Hub (Osborne, Ding, & Kirk, 2024), among others. Given the pace of change in the AI industry, we assume that the projects have evolved since data collection. However, since our research objective was to test prior theory on open source co-opetition, this limitation is defensible. While some findings are specific to the AI industry, the identified collaborative relationship types, practices, and governance challenges likely remain relevant despite recent technological advancements in AI and beyond.

4.5.2.3 Reliability

Reliability concerns the consistency and replicability of the research. To enhance the reliability of the quantitative analysis, the Python scripts and Jupyter notebooks used for data collection and analysis are available on GitHub (Osborne, Daneshyan, et al., 2024b). Three authors participated in a systematic approach to manually label and validate contributors' company affiliations with the goal of enhancing data quality. However, this was a labour-intensive process that took three authors one week to complete, which decreases the repeatability of this approach. With regards to the interviews, they were recorded and transcribed by the first author to aid the analysis, and research guidelines were followed for the thematic analysis of the interview transcripts (Braun & Clarke, 2006). Moreover, the involvement of two authors in coding and validating codes reduced the potential biases that may arise when a single author performs QDA alone (Cruzes & Dybå, 2011). Finally, we member-checked findings with respondents to increase their practical relevance (Lincoln & Guba, 1985).

4.6 Conclusion

This study's investigation of how companies collaborate in company-hosted OSS projects via a mixed-methods analysis of Meta's PyTorch, Google's TensorFlow, and HF's Transformers makes three key contributions to the literature on open source co-opetition. First, while the projects exhibit similar code authorship patterns between host and external companies (~80/20% of commits respectively), collaborations between companies are structured differently (e.g. decentralised vs. hub-and-spoke networks). Second, host and external companies engage in strategic, non-strategic, and contractual collaborations, which vary in the relevance of business strategy, competitive dynamics, and personal incentives of developers. Some of the observed collaborations are specific to the AI industry (e.g. hardware-software optimisations or AI model integrations), while others are typical of the wider software industry (e.g. bug fixing or task outsourcing). Third, single vendor governance in company-hosted OSS projects creates a power imbalance that shapes open source co-opetition practices and possibilities, from a host company's singular decision-making power (e.g. risk of license change) to its community involvement strategy (e.g. from over-control to over-delegation). We concluded with recommendations for future research to advance our understanding of commercial participation and its impact on collaboration dynamics in OSS communities in the AI industry and beyond.

4.7 Appendix for Chapter 4 (RP1)

4.7.1 Commercial Dominance in PyTorch, TensorFlow, and Transformers

Table 4.3: Metrics for Host Company Dominance in PyTorch, TensorFlow, and Transformers

Metrics	**PyTorch**	**TensorFlow**	**Transformers**
# developers	738	1,172	174
# host company employees	452 (61.25%)	558 (47.61%)	56 (32.18%)
# companies	66	167	75
# developers that made 80% of commits	157 (12.54%)	248 (12.03%)	21 (7.54%)
# companies that made 80% of commits	1 (1.52%)	1 (0.59%)	1 (1.33%)
# bus factor (50% commits)	43	78	5
# bus factor (50% net LOC)	22	34	4

N.B.: Limited to contributors who have made $\geq$ 5 commits. The bus factor is the smallest number of people that make 50% of contributions (CHAOSS, 2023).

4.7.2 Top Corporate Contributors to PyTorch, TensorFlow, and Transformers

Table 4.4: Top Corporate Code Contributors to PyTorch, TensorFlow, and Transformers

(a) PyTorch ranked by `n_commits`

Affiliation	Commits	LOC (net)	Out-degree	PageRank	Betweenness
Meta	0.84	0.84	0.98	0.39	0.02
Quansight	0.05	0.04	0.67	0.14	0.25
Nvidia	0.03	0.04	0.78	0.11	0.31
Microsoft	0.02	0.04	0.51	0.02	0.11
AMD	0.01	0.01	0.59	0.04	0.16
Google	0.01	0.01	0.73	0.05	0.20
Intel	0.01	0.01	0.65	0.04	0.35
IBM	0.01	<0.01	0.35	0.01	0.04
Twitter	<0.01	<0.01	0.33	0.01	<0.01
DeepMind	<0.01	<0.01	0.16	0.01	<0.01

N.B. Excludes commits by volunteers (Commits=0.05, LOC=0.05) and unknown affiliations (Commits=0.02, LOC=0.01).

(b) TensorFlow ranked by `n_commits`

Affiliation	Commits	LOC (net)	Out-degree	PageRank	Betweenness
Google	0.85	0.34	0.97	0.25	0.04
Nvidia	0.03	0.15	0.64	0.04	0.07
Intel	0.03	0.12	0.86	0.17	0.09
IBM	0.02	0.09	0.72	0.05	0.06
AMD	0.01	0.06	0.49	0.01	0.04
Arm	0.01	0.03	0.64	0.10	0.04
Huawei	0.01	0.02	0.33	0.01	0.07
Microsoft	<0.01	0.02	0.39	0.01	0.15
Graphcore	<0.01	0.02	0.41	0.01	0.12
Offscale	<0.01	0.01	0.14	<0.01	0.01

N.B. Excludes commits by volunteers (Commits=0.10, LOC=0.06) and unknown affiliations (Commits=0.04, LOC=0.11).

(c) Transformers ranked by `n_commits`

Affiliation	Commits	LOC (net)	Out-degree	PageRank	Betweenness
HF	0.91	0.94	0.96	0.43	0.39
Fractal Ideas	0.01	0.01	0.17	0.14	0.08
Intel	0.01	0.01	0.17	0.01	0.08
Saama AI Research	<0.01	<0.01	0.13	0.01	0.01
Uber	<0.01	<0.01	0.02	<0.01	<0.01
Deepset	<0.01	<0.01	0.09	0.01	0.19
Telekom	<0.01	<0.01	0.06	<0.01	<0.01
Oracle	<0.01	<0.01	0.15	0.01	<0.01
LTIMindtree	<0.01	<0.01	0.17	0.01	0.25
Google	<0.01	<0.01	0.09	0.01	0.07

N.B. Excludes commits by volunteers (Commits=0.10, LOC=0.06) and unknown affiliations (Commits=0.01, LOC=0.01).

4.7.3 Metrics for Collaboration Networks

Table 4.5: Metrics for PyTorch, TensorFlow, and Transformers Collaboration Networks

(a) PyTorch Collaboration Networks

Year	Nodes	Edges	Deg Centr	Deg Skew	Clust Coef
2016	9	34	0.5938	0.16	0.5579
2017	20	130	0.6925	0.64	0.6202
2018	26	194	0.7296	1.44	0.7158
2019	24	178	0.6616	0.89	0.7096
2020	25	210	0.6337	0.65	0.7802
2021	16	108	0.5156	0.23	0.7827
2022	16	126	0.4356	-0.26	0.6956

N.B. Deg Cent = Degree Centralisation, Deg Skew = Degree Skew, Clust Coef = Clustering Coefficient.

(b) TensorFlow Collaboration Networks

Year	Nodes	Edges	Deg Centr	Deg Skew	Clust Coef
2016	22	104	0.8118	2.68	0.6843
2017	43	276	0.8673	2.89	0.6836
2018	43	370	0.7653	1.86	0.7801
2019	55	810	0.7407	1.65	0.8160
2020	63	1208	0.6855	1.33	0.7659
2021	46	692	0.6123	0.94	0.7882
2022	24	190	0.5028	0.78	0.7634

N.B. Deg Cent = Degree Centralisation, Deg Skew = Degree Skew, Clust Coef = Clustering Coefficient.

(c) Transformers Collaboration Networks

Year	Nodes	Edges	Deg Centr	Deg Skew	Clust Coef
2018	4	10	0.2222	0.00	0.8333
2019	23	70	0.9008	3.81	0.5155
2020	27	68	0.9379	4.62	0.3342
2021	27	100	0.8506	3.13	0.4387
2022	24	82	0.7977	3.39	0.4756

N.B. Deg Cent = Degree Centralisation, Deg Skew = Degree Skew, Clust Coef = Clustering Coefficient.

Co-authorship Statement

I am the first author alongside 5 co-authors from my research visit at Peking University's OSS Data Analytics Lab (2023-2024). As the first author, I carried out the majority of tasks, namely: research design (e.g. presented at the Transfer of Status milestone), literature review, developed initial Python scripts for data mining, guided co-authors' optimisation of Python scripts for data mining and processing, manual data labelling, quantitative data analysis, interviewee outreach, 10 interviews, interview transcription, interview analysis, paper writing, paper submission, and two rounds of revision and resubmission at ACM CSCW. My co-authors carried out the following tasks:

- Farbod Daneshyan adapted and improved the efficiency of a Python script that I had written for mining network data from the respective GitHub repositories and carried out data collection under my guidance. In addition, Farbod shared the task of manually labelling the affiliations of contributors, and validated the initial codes that I had produced through thematic analysis of the interview data and we discussed the results to ensure agreement on key themes.
- Runzhi He supported data collection and processing tasks. In particular, Runzhi carried out the automated classification of the affiliations of users and commits in the respective GitHub repositories. In addition, Runzhi shared the task of manually labelling the affiliations of contributors, as well as the task of validating random samples of contributors' company affiliations.
- Hengzhi Ye carried out the username merging process. In addition, Henzhe shared the task of validating random samples of contributors' company affiliations.
- Dr. Yuxia Zhang provided feedback on the research design prior to data collection as well as on drafts of the manuscript prior to its submission to ACM CSCW.
- Prof. Minghui Zhou provided feedback on the research design prior to data collection as well as on drafts of the manuscript prior to its submission and resubmission to ACM CSCW.

5. Why Do Companies Democratise AI? The Case of Software Donations to Foundations

Peer review status: *This paper was presented at the 2023 Creative Commons Global Summit on "AI & the Commons," and is under review for publication.*

Abstract

Companies claim to "democratise" AI when they release (i.e. open-source) or transfer ownership (i.e. donate) of AI OSS or OMs, but what does this term mean and why do companies "democratise" AI? While press releases celebrate a myriad of benefits that AI democratisation promises for research and innovation, the commercial incentives driving such efforts are often obscured from public view. The motivations and strategies behind such efforts warrant careful examination, as they shape not only which technologies are built but also their governance and beneficiaries. Towards this end, this study investigates the commercial incentives for 43 AI OSS donations to the LF as one common method of AI democratisation through a mixed-methods analysis of pre-donation technical pitches, post-donation blog posts, a questionnaire, and semi-structured interviews with OSS donors. This analysis contributes a taxonomy of individual and organisational social, economic, and technological incentives for OSS donations. It highlights the role of democratising the governance of an OSS project—from single-vendor governance by one company to vendor-neutral governance by a non-profit foundation—as a structural enabler for strategic goals, such as attracting external contributors, reducing development costs, and influencing industry standards, among others. Furthermore, OSS donations are often championed by individual developers within their companies, highlighting the relevance of bottom-up incentives for AI democratisation efforts. The RP discusses the generalisability of the findings to OSS donations across the software industry as well as to other open source AI democratisation efforts, such as the release of OMs. It concludes with recommendations for future research.

5.1 Introduction

In recent years, technology giants and startups alike have actively been "democratising" AI through open source initiatives, including by open-sourcing AI software and models as well as donations thereof to non-profit foundations. Wrapped in a rhetoric of ethical capitalism, such efforts that are championed for enabling state-of-the-art AI R&D beyond well-funded corporate labs (Widder, Whittaker, & West, 2024). This narrative has gained traction, with world leaders pledging public funds to OSAI development (Chatterjee & Volpicelli, 2023) and venture capital flowing into OSAI start-ups (Wiggers, 2023). However, AI democratisation remains an ambiguous term, encompassing a variety of goals and methods (Seger, Ovadya, et al., 2023). While corporate press releases celebrate the benefits of their AI democratisation efforts for research and innovation, the commercial incentives driving such efforts are often obscured from public view (Srnicek, 2022).

While commercial contributions have undeniably resulted in the growth of the OSAI ecosystem (White et al., 2024) and advances in AI R&D (Law & Krier, 2023; Langenkamp & Yue, 2022), the growing influence over OSAI development warrants careful examination, as it shapes not only which technologies are built but also their governance and beneficiaries. The risk exists that commercial interests may unduly influence the OSAI ecosystem' trajectory and potentially crowd out public interest alternatives (Whittaker, 2021; Varoquaux et al., 2024; Widder, Whittaker, & West, 2024). Understanding commercial incentives for AI democratisation efforts is, therefore, crucial for creating conditions that promote and safeguard public interests in the OSAI ecosystem. Towards this end, this RP presents an exploratory analysis of why companies democratise AI with a focus on OSS donations as one method of AI democratisation. Specifically, it investigates commercial incentives for 43 AI OSS donations to the LF through a mixed-methods analysis of pre-donation technical pitches, post-donation blog posts, a questionnaire, and semi-structured interviews with OSS donors.

This RP makes contributions to both research and practice. First, it contributes a novel taxonomy of social, economic, and technological incentives at the developer and organisation levels. In particular, it highlights the role of democratising the governance and control rights of an OSS project (i.e. from a company to a vendor-neutral foundation) as a structural enabler for strategic goals, such as attracting external contributors, reducing development costs, and influencing industry standards, among others. Furthermore, it sheds light on the role of individual developers, who champion and coordinate OSS donations within their companies, thus highlighting the relevance of the bottom-up incentives in addition to organisation level business strategy. Some incentives are unique to AI OSS (e.g. develop corporate reputation as an AI leader), while others generalise to OSS donations across the software industry (e.g. reduce development costs or shape industry standards). Beyond OSS

donations, the taxonomy provides a foundation for discerning incentives for other AI democratisation efforts, such as the release of OMs. I discuss which incentives may apply to other types of OSAI technologies (e.g. OMs) and their type of democratisation (e.g. release versus transfer of ownership).

The paper is structured as follows. First, it reviews prior work on the political economy of OSAI and OSS (Section 5.2). Second, it presents the research aims and the study design (Section 5.3). Third, it reports the key findings (Section 5.4). Fourth, it discusses the implications of the findings for research and practice, future research directions, as well as the threats to validity (Section 5.5). Finally, it concludes with a summary of the key contributions of the paper (Section 5.6).

5.2 Related Work

5.2.1 Democratising AI: Narratives and Practices

In light of concerns about industry concentrations and influence on AI R&D, it has become *en vogue* for companies to "democratise AI"—an altruism-laden term that is notoriously ambiguous. Prior work finds that AI democratisation is used as a catch-all term to encompass a variety of goals and practices (Seger, Ovadya, et al., 2023), including the following:

- **Democratising AI use:** Lowering entry barriers for the use of AI technologies, including commercial products like OpenAI's ChatGPT or GitHub's Copilot, access to OMs through APIs or publicly available OMs, and the release of OSS like PyTorch and TensorFlow.
- **Democratising AI development:** Lowering the entry barriers for the development of AI technologies, including but not limited to the release of AI OSS (e.g. DL frameworks) and OMs.
- **Democratising AI profits:** Redistributing the economic value accrued to companies from their use of AI technologies to the respective users and impacted populations.
- **Democratising AI governance:** Distributing the decision-making power in the development or use of AI technologies to a wider community of stakeholders and impacted populations.

In most cases, AI democratisation is used to refer to the lowering of barriers for the use or the development of AI technologies, leading Seger, Ovadya, et al. (2023) to conclude that, "'AI democratisation' is a (mostly) unfortunate term." Open source technologies and collaboration methods have been integral to AI democratisation efforts, enabling both access to and participation in the development of AI. Widder, Whittaker, and West (2024, p.828) identify three primary advantages of OMs: transparency through the publication of model weights and documentation; reusability via open licensing for third-party utilisation; and extensibility, which facilitates the adaptation and fine-tuning

of existing models for specific applications. The authors highlight these advantages apply to most, but not all, OMs because some, such as the Llama models by Meta, are shared under restrictive licenses that limit reusability by certain actors (e.g. companies with more than 700 million users; i.e. market rivals) or for certain harmful use cases. As I discuss below, the providers of OMs under such licenses are accused of "open-washing" (White et al., 2024; Liesenfeld et al., 2023).

Commercial releases of AI OSS (Langenkamp & Yue, 2022; Srnicek, 2022) and OMs (Osborne, Ding, & Kirk, 2024; Widder et al., 2023) have contributed to the rapid growth of the OSAI ecosystem, which now comprises over 300 AI OSS libraries (Haddad, 2022), over a million AI OSS repositories (GitHub, 2024a), and over a million OMs (HuggingFace, 2024h). The prevalence of commercial OSAI releases begs the questions of why companies release their AI software and models, and what are the impacts thereof on the norms, practices, and potential trajectories of AI developer communities?

Prior work hints at a number of incentives. Scholars contend that industry giants promote open source as an alternative to their concentrated power in the AI industry, whilst using it as a means to shape industry standards, benefit from user innovation, and ultimately extend their influence the norms and tools used by researchers and developers around the world (Widder et al., 2023; Widder, Whittaker, & West, 2024). With reference to Google's TensorFlow and Meta's PyTorch, two DL frameworks released in 2015 and 2016 respectively, Srnicek argues that "the seemingly non-capitalist practice of releasing their AI software for free in fact obscures a significant capitalist battle between the major companies" (Srnicek, 2022). This was evident in a leaked Google memo, which claimed that "open source solutions will out-compete companies like Google or OpenAI" and for this reason they should "own the ecosystem and let open source work for us" (Patel & Ahmad, 2023).

Meta has been outspoken about its OSAI strategy: by releasing AI software like PyTorch and LLMs like their Llama models, Meta seeks to increase adoption of its AI technology, improve their performance and safety through crowdsourced feedback, and ultimately benefit from ecosystem effects. Upon releasing Llama 2, Nick Clegg, Meta's President of Global Affairs, explained that "Openness isn't altruism—Meta believes it's in its interest. It leads to better products, faster innovation, and a flourishing market, which benefits us as it does many others" (Clegg, 2023). He argued that releasing Llama 2 would make it "safer and better" because it will benefit from the "wisdom of the crowds." He added that, "A mistaken assumption is that releasing source code or model weights makes systems more vulnerable. On the contrary, external developers and researchers can identify problems that would take teams holed up inside company silos much longer." Meanwhile, Mark Zuckerberg, Meta's CEO, has explained publicly that Meta seeks to build an ecosystem around its AI software and models as a source of strategic advantage. For example, he explained to shareholders that the widespread use of PyTorch has "been very valuable for us" because it has facilitated the integration

of PyTorch-based research and innovations into Meta's systems (Meta, 2023). Similarly, upon the release of Llama 3, he explained that they are not doing open source "because we are, like, altruistic... I just want everyone to be using it because the more people who are using it, the more the flywheel will spin for making Llama better" (South Park Commons, 2024).

These statements illustrate how companies seek to build moats through building and owning software ecosystems. Software ecosystems are "a set of businesses functioning as a unit and interacting with a shared market for software and services, together with the relationships among them" (Jansen et al., 2009, p.9). Ecosystem strategies mirror strategies for industry platforms, which are the "products, services, or technologies that act as a foundation upon which external innovators, organized as an innovative business ecosystem, can develop their own complementary products, technologies, or services" (Gawer & Cusumano, 2014, p.417). Platform ownership confers significant competitive advantages through strategic positioning as intermediaries between users and their activities as well as being the rule-setter (Srnicek, 2017). Platforms also facilitate innovation on complementary products and services. "The more innovation there is on complements, the more value it creates for the platform and its users via network effects, creating a cumulative advantage for existing platforms: As they grow in adoption, they become harder to dislodge by rivals or new entrants, with the growing number of complements acting like a barrier to entry" (Gawer & Cusumano, 2014, p.421). Network effects often "mean that early advantages become solidified as permanent positions of industry leadership" (Srnicek, 2017, p.95). Just as platforms leverage network effects and complementary innovations to establish market dominance, corporate software ecosystems similarly create competitive advantages by fostering developer communities and users around their technologies.

In addition to ecosystem-building, companies have released AI OSS or OMs under restrictive licenses, which has been criticised as an act of "open-washing" (White et al., 2024; Liesenfeld et al., 2023). Liesenfeld and Dingemanse (2024) argue that companies engage in open-washing to reap the benefits of open source (e.g. reputation rewards and adoption), whilst not actually complying with open source standards or norms (e.g. via restrictive licenses). For example, Meta released Llama 2 with much fanfare, claiming that the "open source" model would benefit research and innovation, its distribution under a novel license with restrictive commercial terms (i.e. any company with greater than 700 million monthly active users in the preceding month must request a license that Meta may grant in its sole discretion) received backlash from the open source community (White et al., 2024; Widder et al., 2023; Maffulli, 2023). For example, Stefano Maffulli, the Executive Director of the Open Source Initiative, commented, "Unfortunately, the tech giant has created the misunderstanding that Llama 2 is 'open source' – it is not. Meta is confusing 'open source' with 'resources available to some users under some conditions,' [which are] two very different things" (Maffulli, 2023).

What is more, the democratisation of OSS and OMs must be situated within the wider supply chain of AI (Valdivia, 2024); that is, the interconnected commodity chains of natural resources, hardware, data centres, cloud compute platforms, and developer platforms required to develop and deploy AI systems. While companies may "democratise" software or models, industry concentrations persist in critical infrastructure and resources needed for AI development (Widder, Whittaker, & West, 2024; Varoquaux et al., 2024), including the production of hardware like AI accelerators (Sastry et al., 2024), the ownership of data centres (Lehdonvirta et al., 2024), as well as the ownership of digital platforms (Srnicek, 2017), from cloud computing services to developer platforms. This perspective acknowledges that no matter how much investment goes into OSS or OMs, AI democratisation efforts will do little to fundamentally reconfigure the distribution of power and resources in the wider AI industry. As it was put by Srnicek (2017, p.97): "Open-sourcing all software or capital investments are not enough to overturn monopolies; access to data, network effects, and path dependency place even higher hurdles in the way of overcoming a monopoly like Google."

This prior work provides insights into the OSAI strategies of industry giants and concerns related to open-washing. However, we still have significant gaps in our understanding of why and how different types of companies, beyond industry giants, employ open source as a means to "democratise" AI. To address this gap, in the next section, I draw on the extensive literature on the political economy of OSS, which provides a comprehensive theoretical foundation for understanding commercial incentives for AI democratisation efforts that are facilitated by open source technologies.

5.2.2 Commercial Incentives in OSS Development

Companies have participated in the development of OSS in a myriad of ways since the late 1990s (Broca, 2021; X. Li et al., 2024), including by deploying developers to contribute to projects as part of their job responsibilities or corporate social responsibility initiatives (Dahlander & Magnusson, 2005; Dahlander & Wallin, 2006; Lee & Herstatt, 2015), funding projects (O'Brien, 2019a; Osborne, 2024a), or joining project steering committees (Butler et al., 2018; Wagstrom, 2009), among others. These are common strategies through which companies seek to influence projects that develop maintain OSS that they use (Dahlander & Wallin, 2006).

It is also common for companies to spin-out proprietary software as company-hosted OSS, where the host company controls the intellectual property of the project (e.g. by requiring contributors to sign a contributor license agreement) and employs the maintainers of the project (M. Zhou et al., 2016; Yue & Nagle, 2024). This is a proven strategy to increase adoption of their software, benefit from external contributions, win market share, or reduce a competitor's market share (West & O'Mahony, 2005). In some cases, a handful of companies share control of a project; for example,

in 2017, Facebook and Microsoft jointly released ONNX to enable interoperability between various DL frameworks like Meta's PyTorch and Google's TensorFlow (Candela, 2017). One may speculate that Meta and Microsoft built ONNX to mitigate lock-in within the Google's TensorFlow ecosystem, which at the time was the dominant DL framework.

An extensive literature discusses the diverse incentives for the adoption and development of OSS at both the individual level (see Table 5.1) and the organisation level (see Table 5.2). Bonaccorsi and Rossi's (2006) taxonomy of social, economic, and technological incentives at the developer and organisation levels provides an enduring framework for categorising these diverse incentives. Individuals are mostly driven by social and technological incentives, such as their personal interest (Benkler, 2006; Raymond, 2001a; von Krogh et al., 2012), values (Kelty, 2008; K. Lakhani & von Hippel, 2003; Shah, 2006), or needs (Franke & Hippel, 2003; Hars & Ou, 2002; J. Roberts, Hann, & Slaughter, 2006). However, the incentives of individuals vary based on factors such as whether they are volunteers or paid (K. R. Lakhani & Wolf, 2003; Dahlander & Wallin, 2006), the governance of the OSS project (Shah, 2006), and their geography or cultural norms (Subramanyam & Xia, 2008; Takhteyev, 2012), among others. While individual developers, including volunteers, do not share the primary incentives of for-profit companies, they tend to accept commercial participation in OSS projects on condition that they comply with community norms (Bonaccorsi & Rossi, 2006). Furthermore, commercial participation in OSS projects can also attract volunteers, who see their presence as a signal of the complexity of the project (O'Mahony & Bechky, 2008; Santos et al., 2013).

By contrast, companies are chiefly motivated by economic and technological factors, such as influencing industry standards (Lerner & Tirole, 2005; Lindman et al., 2009), reducing development costs (Birkinbine, 2020; Chesbrough, 2023), and recruiting external developers (Ågerfalk & Fitzgerald, 2008; Fink, 2003). Moreover, commercial participation in OSS development can help companies to improve their market position by undercutting the product of a market rival (Fink, 2003; West & O'Mahony, 2005) as well as to enhance their reputation as an OSS patron among developers (Feller & Fitzgerald, 2002; Bonaccorsi & Rossi, 2006; Pitt, Watson, Berthon, Wynn, & Zinkhan, 2006), which in turn can help to recruit software developers (Ågerfalk & Fitzgerald, 2008; Lindman et al., 2009).

In particular, the distributed production model of OSS development, involving many more skilled developers beyond those within the organisational boundaries of any single company, is viewed as means to decrease in-house R&D costs (Lindman et al., 2009; Ågerfalk & Fitzgerald, 2008). Birkinbine argues that the greatest value of OSS for companies stems from the peer production model that expands the labour force that can test and develop the software. Specifically, he contends that the value for companies stems from the *processes*, not the *products*, of OSS development (Birkinbine, 2020). The extent of volunteer activity in OSS development, from bug-spotting to code contributions

(Feller & Fitzgerald, 2002; M. Zhou et al., 2016), raises ethical questions about the exploitation of volunteer work (H. Li et al., 2022) and the failure of most companies to adequately reciprocate to support the sustainability of OSS projects (Champion & Hill, 2021; M. Hoffmann, Nagle, & Zhou, 2024). For example, while Linus' Law—i.e. that "Given enough eyeballs, all bugs are shallow" (Raymond, 2001a)—is typically quoted to argue that the OSS development model offers security advantages over proprietary software development, one can extend it to convey the value of distributed bug-spotting and improvements that no single company must pay for by themselves.

It is often the case that companies specifically seek to collaborate with other companies, including market rivals, as a means to jointly share R&D costs (Teixeira et al., 2016; Nguyen-Duc et al., 2019) and shape industry standards (Lerner & Tirole, 2001, 2002; West & O'Mahony, 2005). While inter-company collaborations do not necessarily exclude volunteers, it is not uncommon for companies to engage in strategic or contractual collaborations, involving private collaboration methods, which volunteers cannot participate in (Osborne, Daneshyan, et al., 2024a). The prevalence of inter-company collaborations has turned many OSS communities "from networks of individuals into networks of companies" (Ågerfalk & Fitzgerald, 2008, p.396), resulting in a tangle of cooperation and competition between companies known as "open source co-opetition" (Teixeira & Lin, 2014).

The diversity of incentives of various stakeholders underlines the critical role of governance in OSS projects (O'Mahony & Bechky, 2008). Non-profit foundations have emerged as key mediators—or "boundary organisations"—whose vendor-neutrality and open governance have proven to be effective structural enablers of collaboration between "unexpected allies" (O'Mahony & Bechky, 2008). For instance, the LF is reputed to facilitate "communities of competitors," where "market rivals...intentionally coordinate activities for mutual benefit in precise, market-focused, non-differentiating engagements" (Germonprez et al., 2013). Foundations are reputed to limit the dominance of any single company in OSS projects, which increases adoption (M. Zhou et al., 2016) and attracts new contributors to projects (Di Giacomo, Kudzmanaite, Deveny, Dussutour, & Shaikh, 2020; West & O'Mahony, 2005; Link & Germonprez, 2016), especially volunteers who do not want to perform free work for a dominant company (Y. Zhang et al., 2018, 2021; M. Zhou et al., 2016).

However, foundations do not always prevent commercial dominance in OSS projects (Wagstrom, 2009; Y. Zhang et al., 2018, 2022). For instance, around 10% of companies account for 80% of commits to projects in the OpenStack ecosystem (Y. Zhang et al., 2021). Moreover, governance changes resulting from donations do not guarantee activity increases. For example, a study of PyTorch's governance transition from Meta to the LF revealed no net increase in project activity. Specifically, it found that contributions from Meta decreased significantly, contributions from users (e.g. app developers and cloud providers) remained unchanged, but those from complementers (e.g. chip manufactur-

Table 5.1: Incentives for Individuals to Participate in OSS Development

Category	Incentive	References
Social	Fun & geek culture	Torvalds (2001); Raymond (2001a); Gerosa et al. (2021); K. R. Lakhani and Wolf (2003); K. Lakhani and von Hippel (2003); Benkler (2002, 2006); Hertel, Niedner, and Herrmann (2003); Luthiger and Jungwirth (2007); J. Roberts et al. (2006); Shah (2006); von Hippel (2003)
	Reputation & peer recognition	Raymond (2001a); Gerosa et al. (2021); Bezroukov (1999); Ghosh, Glott, Krieger, and Robles (2002); K. R. Lakhani and Wolf (2003); K. Lakhani and von Hippel (2003); Lerner and Tirole (2005); Fershtman and Gandal (2004); Ghosh (2007, 1998); Hars and Ou (2002); Hertel et al. (2003); Lattemann and Stieglitz (2005); Lerner and Tirole (2002); Okoli and Oh (2007); Oreg and Nov (2008); J. Roberts et al. (2006); Spaeth, Haefliger, von Krogh, and Renzl (2008); Stewart and Gosain (2006); B. Xu, Jones, and Shao (2009); Yu, Jiang, and Chan (2007); Hemetsberger (2002); Loebbecke and Angehrn (2003); Krishnamurthy (2006); Howison and Herbsleb (2013)
	Altruism & reciprocity	Raymond (2001a); Gerosa et al. (2021); Raymond (2001b); Bergquist and Ljungberg (2001); Bitzer, Schrettl, and Schröder (2007); Crowston and Scozzi (2002); Hars and Ou (2002); Osterloh et al. (2003); P. David and Shapiro (2008); Ghosh and Schmidt (2006); Haruvy, Prasad, and Sethi (2003); Ke and Zhang (2008); K. R. Lakhani and Wolf (2003); Oreg and Nov (2008); Shah (2006); Stewart and Gosain (2006); C.-G. Wu, Gerlach, and Young (2007); Yu et al. (2007); Nagle, Wheeler, Lifshitz-Assaf, Ham, and Hoffman (2020); Ding et al. (2023)
	Kinship with OSS community	Pfaff and David (1998); Gerosa et al. (2021); Raymond (2001b); Hars and Ou (2002); P. David and Shapiro (2008); Hertel et al. (2003); K. Lakhani and von Hippel (2003); Schofield and Cooper (2006); Zeitlyn (2003); Bagozzi and Dholakia (2006); Hemetsberger (2002)
	Ideology & values	Stallman (1984); Raymond (2001a); Gerosa et al. (2021); Kelty (2008); Takhteyev (2012); Schoonmaker (2018); P. David and Shapiro (2008); P. David, Waterman, and Arora (2003); Ghosh and Schmidt (2006); K. Lakhani and von Hippel (2003); Shah (2006); Stewart and Gosain (2006); J. Xu, Christley, and Madey (2006); Yu et al. (2007); Akiki et al. (2022)
Economic	Payment	Lerner and Tirole (2005); Feller and Fitzgerald (2002); Hertel et al. (2003); Hars and Ou (2002); Ghosh and Schmidt (2006); K. R. Lakhani and Wolf (2003); Lattemann and Stieglitz (2005); Luthiger and Jungwirth (2007); J. Roberts et al. (2006); Howison and Herbsleb (2013)
	Career benefits	Lerner and Tirole (2001, 2002); K. R. Lakhani, Wolf, Bates, and DiBona (2002); K. R. Lakhani and Wolf (2003); Gerosa et al. (2021); Ghosh et al. (2002); Subramanyam and Xia (2008); Hossain (2021); Hann, Roberts, Slaughter, and Fielding (2002); Hars and Ou (2002); Hertel et al. (2003); Riehle (2007); J. Roberts et al. (2006); Shah (2006); Yu et al. (2007)
	No or low cost	Bonaccorsi and Rossi (2003); Kollock (1999); K. R. Lakhani and Wolf (2003)
Technological	Learning & skill development	Bonaccorsi and Rossi (2006); K. R. Lakhani and Wolf (2003); Ghosh et al. (2002); P. David and Shapiro (2008); P. David et al. (2003); Hars and Ou (2002); Gerosa et al. (2021); J. Roberts et al. (2006); Oreg and Nov (2008); Shah (2006); Stewart and Gosain (2006); Spaeth et al. (2008); von Hippel (2003); C.-G. Wu et al. (2007); Ye and Kishida (2003); Ding et al. (2023)
	Use cutting-edge software	Pfaff and David (1998); Pavlicek (2000); Green (2000); Franke and Hippel (2003); Feller and Fitzgerald (2002); Ding et al. (2023)
	Personal need	Raymond (2001a); von Hippel (2003, 2001); Hars and Ou (2002); Osterloh et al. (2003); Bitzer and Schröder (2006); Ghosh et al. (2002); Hertel et al. (2003); Lattemann and Stieglitz (2005); Schofield and Cooper (2006); Shah (2006); P. David and Shapiro (2008); Ghosh and Schmidt (2006)

N.B. This table was adapted from Bonaccorsi and Rossi (2006) and updated with more recent publications.

Table 5.2: Incentives for Companies to Participate in OSS Development

Category	Sub-category	References
Social	Reciprocity to OSS community	Feller and Fitzgerald (2002); Franck and Jungwirth (2002); Osterloh et al. (2003); Lerner and Tirole (2002); Germonprez et al. (2013)
	Build reputation as OSS patron	Osterloh et al. (2003); Lerner and Tirole (2002); Pitt et al. (2006)
	OSS philosophy (as a strategy to undercut market rivals)	Feller and Fitzgerald (2002); Widder, Whittaker, and West (2024)
Economic	Reduce software development costs	Ågerfalk and Fitzgerald (2008); Birkinbine (2020, 2018); Feller and Fitzgerald (2002); Hawkins (2004); Lindman et al. (2009); Markus et al. (2000); Marsan, Paré, and Beaudry (2012); Woods and Guliani (2005); Kendall, Kendall, and Germonprez (2016); Germonprez et al. (2013); K. R. Lakhani et al. (2002); Linåker et al. (2016); Loebbecke and Angehrn (2003); Nguyen Duc et al. (2017); Nguyen-Duc et al. (2019); Teixeira and Lin (2014); Teixeira et al. (2015); H. Li et al. (2022); Tapscott (2011)
	Independence from the price and license policies of software vendors	Chesbrough (2023); Lerner and Tirole (2002)
	Revenue from OSS services & products	Chesbrough (2023); Feller and Fitzgerald (2002); Lindman et al. (2009); West (2003); Wichmann (2002)
	Recruit talent	Ågerfalk and Fitzgerald (2008); Fink (2003); Lerner and Tirole (2002); Mugrage (2022); Wichmann (2002); Marlow and Dabbish (2013)
	Increase market competitiveness	Ahlawat et al. (2021); Chesbrough (2023); Hossain (2021); Lindman et al. (2009); Loebbecke and Angehrn (2003); Woods and Guliani (2005); Kendall et al. (2016); Widder, Whittaker, and West (2024)
Technological	Crowdsource innovation	Ågerfalk and Fitzgerald (2008); Birkinbine (2020, 2018); Chesbrough (2023); Feller and Fitzgerald (2002); Hawkins (2004); Lindman et al. (2009); Markus et al. (2000); Marsan et al. (2012); Woods and Guliani (2005); Kendall et al. (2016); Germonprez et al. (2013); K. R. Lakhani et al. (2002); Linåker et al. (2016); Teixeira (2014); H. Li et al. (2022)
	Enhance software security through distributed testing	Alrawashdeh, Elbes, Almomani, ElQirem, and Tamimi (2020); Bonaccorsi and Rossi (2006); Birkinbine (2020); Chesbrough (2023); Fink (2003); Widder et al. (2023); Hecker (1999); Hawkins (2004); Henkel (2006); von Hippel (2003); Lerner and Tirole (2002); Lindman et al. (2009); Krishnamurthy (2005b); Germonprez et al. (2013)
	Increase adoption of own software	Lerner and Tirole (2002); Wagstrom (2009); Widder et al. (2023); Widder, Whittaker, and West (2024)
	Promote standards & interoperability	Chesbrough (2023); Fink (2003); Lerner and Tirole (2002); Lindman et al. (2009); Wichmann (2002); West (2003)
	Influence roadmap of OSS projects	Dahlander and Wallin (2006); K. R. Lakhani and Wolf (2003); Lindman et al. (2009); von Krogh et al. (2012); Kendall et al. (2016)

N.B. This table was adapted from Bonaccorsi and Rossi (2006) and updated with more recent publications.

ers) increased (Yue & Nagle, 2024). While governance changes may address "hold-up" problems for certain companies, particularly for complementers whose value capture proposition depends on interoperability, they do not guarantee net increases in external contributions (Yue & Nagle, 2024).

This review has provided a theoretical foundation for examining commercial incentives behind open source AI democratisation efforts. The taxonomy of social, economic, and technological incentives at developer and organisation levels (Bonaccorsi & Rossi, 2006) offers a valuable framework for this study's exploratory aims. By applying this framework to AI OSS donations, this study aims to identify and categorise the commercial interests driving AI democratisation. The following section outlines the research aims and methodological approach in more detail.

5.3 Study Design

5.3.1 Research Aims

The objective of this study is to identify and categorise commercial interests for AI democratisation. Specifically, it examines the following RQ: Why do companies democratise AI? Given the various methods of AI democratisation (Seger, Ovadya, et al., 2023), it focuses on AI OSS donations to foundations—that is, the transfer of an OSS project from a company's ownership to a non-profit foundation (O'Mahony & Ferraro, 2007)—which in the AI industry are commonly presented as acts of AI democratisation. While this narrow scope enables an in-depth analysis of one method of AI democratisation, it limits the generalisability of the findings to other types of AI democratisation.

Within this scope, a mixed-methods approach was employed to identify and categorise the incentives for 43 OSS donations by a range of companies, from start-ups to multinational corporations, to the LF AI & Data Foundation and PyTorch Foundation, two foundations under the LF umbrella, between May 2018 and October 2022. The range of OSS projects and companies form a diverse sample (see Table 5.5), accounting for various project maturity levels, company sizes, sectors, and countries (Easterbrook et al., 2008). Furthermore, the mixed-methods analysis of multiple cases mitigates the uniqueness of single cases (Eisenhardt, 1989) and data sources (Yin, 2018).

5.3.2 Case Presentation

5.3.2.1 LF AI & Data Foundation

The LF AI & Data Foundation was founded in March 2018 as the LF DL Foundation and rebranded as the LF AI Foundation in May 2019, broadening its scope to encompass various AI sub-fields. In October 2020, it merged with the ODPi, an organisation promoting a big data software ecosystem.

The foundation subsequently rebranded as the LF AI & Data Foundation, acknowledging the vital role of data in AI R&D. At the point of data collection (October 2022), the foundation had 51 member companies from North America, Europe, and East Asia. It hosted 42 OSS projects that had been donated by diverse organisations, including start-ups (e.g. AI Squared), research institutes (e.g. the Beijing Academy of AI), consortia (e.g. ONNX), management consultancies (e.g. McKinsey & Co.), and technology giants (e.g. IBM, Samsung, and Tencent). When a company seeks to donate an OSS project to the LF AI & Data Foundation, they must be a member organisation of the LF or be endorsed by a member and submit their proposal for review by the technical advisory council (TAC). The TAC comprises representatives from the various projects at the foundation and premier member companies, who vote on the approval of donation proposals. The LF AI & Data Foundation segregates business and technical governance of hosted OSS projects, ensuring that developers retain technical control in their projects whilst the foundation assumes responsibility for funding, marketing, and license compliance, among others, and enforces open governance (Dolan, 2023).

5.3.2.2 PyTorch Foundation

The PyTorch Foundation was established in September 2022 to host Meta's PyTorch, the DL framework (PyTorch, 2023a). Its mission is to "driv[e] the adoption of AI tooling by fostering and sustaining an ecosystem of open source, vendor-neutral projects integrated with PyTorch" and "to democratise state-of-the-art tools, libraries, and other components to make these innovations accessible to everyone" (PyTorch, 2023a). It similarly separates business and technical governance for the PyTorch project and ecosystem, with PyTorch retaining its technical governance structure while the foundation assumed responsibility for funding, hosting expenses, and events, among others. The PyTorch Foundation manages the project's assets, including its website, GitHub repository, and social media accounts, and enforces open governance. Upon its launch, it formed a governing board comprising AMD, AWS, Google Cloud, HF, IBM, Intel, Meta, Microsoft Azure, and Nvidia (PyTorch, 2023a). The governing board members shape PyTorch's strategic direction, contribute to its technical roadmap, and benefit from early feature access and increased visibility in the PyTorch ecosystem, while they are expected to actively promote PyTorch's growth and adoption.

5.3.3 Data & Methods

5.3.3.1 Data Collection

This study comprises two sources of primary data and two sources of secondary data. First, two types of secondary data were collected from the Internet for 43 AI OSS donations to the LF (42 donations

to the LF AI & Data Foundation and 1 donation to the PyTorch Foundation). First, pre-donation technical pitches to the TAC were collected from the LF AI & Data Foundation wiki page (LFAI&Data, 2023). Second, post-donation blog posts by the LF and respective companies were collected from the LF AI & Data Foundation website and through Google search queries in the format of "[company name] + [project name] + [LF AI & Data Foundation]." This yielded 40 presentations (95%) and 37 (88%) blog posts for the 42 projects donated to the LF AI & Data Foundation. It was not possible to collect a pre-donation technical pitch for the PyTorch donation because, as an inaugural project of its namesake foundation, it did not follow this process. Blog posts were accessed via the PyTorch Foundation website and Google search queries using the aforementioned format. A full list of the 43 OSS projects, donors, and respective document links is provided in Table 5.5.

Subsequently, primary data was collected through questionnaires and 12 semi-structured interviews with ten project maintainers who had donated the project and two foundation employees. First, a questionnaire was distributed to the maintainers to gather information on the donation process, incentives, and outcomes, as well as to recruit interviewees. It was distributed via the LF AI & Data Foundation's mailing list and to the PyTorch maintainers via the Executive Director of the PyTorch Foundation, resulting in 16 responses from the maintainers of 16 projects at the LF AI & Data Foundation and 0 responses from the maintainers of PyTorch (in total, 37% of 43 projects). Ten donors were recruited for interviews through the questionnaire, who worked for 9 companies, diverse by geography, size, and sector (see Table 5.3). In addition, the Executive Director of the foundations (N.B. same person) and a LF AI & Data Foundation project coordinator were interviewed. The interviews lasted between 30 and 60 minutes and were semi-structured, combining standardised questions about the donation process with tailored questions based on their questionnaire responses, adding depth to the quantitative findings in Figure 5.1. The interviews were conducted digitally, and were recorded to aid analysis and to enhance the validity of the research findings (Yin, 2018).

5.3.3.2 Data Analysis

Thematic analysis was performed by the author as a systematic approach to identify commonalities, patterns, and relationships in the qualitative data (Cruzes & Dybå, 2011). A six-step procedure was followed to enhance the reliability of this analysis (Braun & Clarke, 2006). The initial coding procedure involved an integrated approach, combining inductive coding (Charmaz, 2006) and deductive coding informed by prior work on incentives that exist at the levels of individual developers (see Table 5.1) and organisations (see Table 5.2). This approach allowed for the identification of both commonalities with prior work and novel findings concerning AI democratisation efforts. The coding was conducted by the author until reaching saturation (Charmaz, 2006), then merged codes into 25

Table 5.3: List of Respondents and Affiliations

ID	Project	Role	Org Size	Sector
A	BeyondML	Maintainer	Small	Information technology
B	Elyra	Maintainer	Large	Information technology
C	Elyra	Maintainer	Large	Information technology
D	Kedro	Maintainer	Large	Professional services
E	Kedro	Maintainer	Large	Professional services
F	KServe	Maintainer	Large	Information technology
G	Ludwig	Maintainer	Small	Information technology
H	NNstreamer	Maintainer	Large	Information technology
I	ONNX-MLIR	Maintainer	Large	Information technology
J	ONNX-MLIR	Maintainer	Medium	Information technology
K	LF AI & Data	Project coordinator	Medium	Non-profit foundation
L	LF AI & Data, PyTorch Foundation	Executive director	Medium	Non-profit foundation

distinct themes (i.e. social, economic, and technological incentives at the developer and organisation levels, see Table 5.4). To address the limitation of single-author analysis, each step was rigorously documented, and the results were member-checked with the interviewees and discussed with two academic advisers (Edwards & Holland, 2013; King, 2009). Furthermore, the interviewees were invited to review the quotes attributed to their anonymised IDs and to state their attribution preference, ensuring consent for the inclusion of statements. Only three interviewees proposed revisions (e.g. to enhance specificity), indicating the resonance of the findings with the practitioners.

5.3.3.3 Reflexivity

In social science research, it is critical that one engages in critical self-evaluation of the one's positionality and disciplinary conventions, and how they may influence one's research, from its design to the reporting of of findings (Finlay, 2002). The exercise of reflexivity was particularly important for the credibility of this study, given my affiliation with the LF. I employed a social identity map as a tool to encourage reflection on my positionality and address potential biases in two areas (Jacobson & Mustafa, 2019). First, at the outset, it was used to consider the effects of my affiliation on data access and potential threats to reproducibility, recognising that it likely increased the willingness of foundation staff and OSS donors to participate in the study. To enhance reproducibility, the data collection strategy relied primarily on publicly available information (e.g. mailing lists). Second, to minimise social desirability bias in interview responses, the research's independent purpose and funding were mentioned in the information sheet and explained in the interviews. While these actions enhanced the credibility and reliability of the research process, I acknowledge that it is difficult, if not impossible, to perform bias-free research and, therefore, it should be understood as a best-effort attempt to control for potential biases in the research process (Jacobson & Mustafa, 2019).

5.4 Results

The findings reveal social, economic, and technological incentives at both developer and organisation levels for the AI OSS donations to the LF (see Table 5.4). This section presents these findings, which contribute a more nuanced understanding of commercial incentives for AI democratisation.

5.4.1 Developer-level Incentives

Key findings

Individual developers play a salient role in many OSS donations through bottom-up championing of donations within their organisations. The questionnaire found that 38% of donations were initiated by developers, while 13% came from individuals with dual developer-manager roles. Key incentives include reciprocating to the OSS community, ensuring project sustainability, and gaining access to better development tools and foundation support services. Personal reputation building and career advancement also emerge as significant incentives.

5.4.1.1 Social Incentives

At the individual level, social incentives are as a significant driver for OSS donations. Two key themes stood out: reciprocity and reputation. Many developers expressed their firm belief in "giving back" to the community. For instance, Respondent A (BeyondML) stated:

> *The vast majority of proprietary models and software in data science and ML are built on open source, so being part of that and contributing to that is really important to me personally and to our company.*

This sentiment was echoed by Respondent H (NNStreamer), who noted that their donation was "just for our own satisfaction" as OSS users and developers. These responses highlight the personal investment many developers have in the OSS ecosystem and their desire to contribute to its growth and sustainability. Personal reputation is also a significant social incentive, with successful donations to major foundations representing an opportunity for individual developers to enhance their standing in the OSS community. For example, Respondent J (ONNX-MLIR) explained that having a project accepted by a major foundation provides credibility among peers and brings developers like himself closer to their "dream of having your big open source project with 1000s of stars." Respondent I (ONNX-MLIR) added that improving one's reputation also leads to career benefits, as one can be hired based on one's reputation, and that individuals' aspirations align with company's goals to improve their corporate reputation as an OSS-friendly workplace.

5.4.1.2 Economic Incentives

Two primary economic incentives at the individual level are career benefits and access to foundation support services. The reputational gains from OSS contributions often translate into tangible career benefits. For example, Respondent I (ONNX-MLIR) noted that achievements in OSS projects make developers more competitive in the AI job market, as their expertise at the intersection of software engineering and AI becomes both known and knowable in the wider OSAI ecosystem. This enhanced visibility and credibility in the job market creates a powerful incentive for individuals to support OSS donations. As mentioned above, Respondent I (ONNX-MLIR) pointed out that these incentives of individual developers also align with organisational goals of fostering a skilled workforce, attracting and recruiting competitive AI talent, and improving their corporate reputation among OSS developers.

The desire to harness foundation support services is another important economic incentive. Many developers viewed foundation support as a means to address challenges faced by maintainers in managing projects alongside their full-time employment. 87% of respondents claimed this support was important for them (see Figure 5.1). Respondent F (KServe) elaborated on this point:

> *We, as developers, don't have a lot of time for [outreach and marketing]. We sought to benefit from support services, such as outreach to new contributors and marketing.*

Similarly, Respondent H (NNStreamer) sought marketing support to increase project visibility and to attract external developers. These examples illustrate how foundations are perceived to provide resources that individual maintainers often lack or cannot secure within their own companies, which in turn helps to enhance the growth of their OSS projects.

5.4.1.3 Technological Incentives

Technological incentives at the individual level include ensuring project sustainability and enabling the use of collaboration tools. Project sustainability is a significant concern for maintainers. Respondent G (Ludwig) described how he donated his project to ensure its survival following organisational restructuring and his personal departure from the company. He viewed the transition to a foundation as an effective strategy to ensure the project's continuation. This case demonstrates how personal attachment to projects, which he described as his "baby," can drive individuals to seek sustainable governance models for their OSS projects, especially when their affiliation with the company comes to an end. It also highlights the use of OSS donations as a mechanism to preserve source code that might otherwise be lost due to corporate changes or neglect.

The ability to use preferred collaboration tools is another technological incentive. Respondents D and E (Kedro) explained that transitioning the project to the LF AI & Data Foundation made it easier

to use tools like Slack and Discord, which were either forbidden or difficult to get approval for at their company. As Respondent E (Kedro) explained, "It untied our hands from our own bureaucracy." They elaborated that this freedom from corporate constraints concerning collaboration tools not only enhances developer productivity and satisfaction but also aligns with broader organisational goals of fostering innovation and efficiency, thus representing a win-win scenario for them.

5.4.2 Company-level Incentives

Key findings

While individual advocacy matters, company-level strategic objectives drive most OSS donation decisions. The questionnaire found that 44% of donations were initiated by managers, while 13% came from individuals with dual developer-manager roles. The questionnaire identified the adoption of open governance (81%), attraction of external contributors (100%), and enhanced corporate reputation (75%) as key incentives. Meanwhile the interviews highlighted economic incentives, such as reduced development costs and diversified project funding, and technological incentives, such as ecosystem integration, software improvements, and influence on standards, as key company-level incentives for OSS donations.

5.4.2.1 Social Incentives

At the organisation level, three primary social incentives emerged: adopting open governance, reciprocating to the OSS ecosystem, and building the company's reputation. The adoption of open governance upon donating a project to a foundation is a salient incentive, with 81% of respondents reporting it as important. This change in governance model is viewed as a structural enabler for downstream goals. For example, Detakin's press release highlighted this incentive:

> *The LF AI & Data provides a vendor-neutral governance structure that can help the project grow broad industry collaboration. Even more importantly, becoming a LF AI & Data project ensures that OpenLineage can never belong to a company.*

Similarly, Lyft emphasised the importance of a "neutral holding ground" when donating both Amundsen and Flyte. These statements underscore the perceived value of neutral governance in fostering collaboration and ensuring the independence of projects.

Reciprocity was mentioned as a key social incentive at the organisation level. Several respondents cited the critical importance of OSS dependencies in their companies' proprietary products and services, and perceived the donation of their OSS project as one way to "give back," as explained by

Respondent A (BeyondML). In a similar vein, Respondent C (Elyra) underscored the impact of OSS for advances in AI as a reason for why he champions giving back to the ecosystem:

> *The democratisation of AI software is really what is helping industry advance. If you look back like 10-20 years ago, it was very hard and you needed to have a specific set of skills to be able to even build a very simple model. Today with all the tools and stuff, it's much easier.*

Respondent C (Elyra) explained that their team's decision to donate Elyra stemmed in part from their desire to play their part in advancing industry. This sentiment was echoed in several companies' post-donation press releases, framing their donations as acts of AI democratisation. For instance, Uber stated that by donating Pyro, it hoped "to facilitate greater opportunities for researchers worldwide and [to] make DL and Bayesian modelling more accessible."

Building a corporate reputation as an AI leader was mentioned as another significant social incentive. Many respondents explained they hoped that the reputation of the LF AI & Data would enhance their company's credibility in the OSAI ecosystem, with 75% of respondents reporting it as important. This incentive was particularly strong for start-ups and companies without an established reputation in AI. Respondent A (BeyondML) explained:

> *One of the things that we hoped to get from the LF is its name recognition, obviously just about every developer in the world knows about the LF or knows the term Linux. So having that kind of badge, if you will, immediately gives you a level of credibility with your project.*

Respondents D and E (Kedro) echoed this sentiment, stating in their pre-donation pitch that they "would like to leverage the initial marketing announcements to build credibility in their technical and product-related capability." Then, in the post-donation press release, their company stated:

> *It's a substantial step forward for our organisation on our open source journey. The consultancy sector has traditionally been highly protective of intellectual property, but it's clear that open, collaborative innovation will help power the next phase for analytics technology.*

These findings highlight three key social incentives driving organisational decisions to donate AI OSS projects. The adoption of open governance is a key structural enabler for fostering collaboration and ensuring project independence. Reciprocity to the OSS ecosystem reflects companies' recognition of their dependence on open source technologies and their desire to contribute to industry advancement. Additionally, the opportunity to enhance corporate reputation, particularly for less established companies, serves as a salient motivator. These social incentives are strategically important for companies seeking to position themselves favourably within the OSAI ecosystem.

5.4.2.2 Economic Incentives

Several economic incentives inform the decision to donate OSS projects, including attracting external contributors, reducing development costs, diversifying project funding, and harnessing foundation support services. 100% of questionnaire respondents stated that attracting new contributors to their project was an important incentive. Respondent J (ONNX-MLIR) described OSS donations as a strategic trade-off, where companies exchange full control of their OSS project for the aspired for benefits of distributed development involving a community of contributors. He noted the self-interested logic:

> *We continue doing the same work as we would if it wasn't open source, but there's this expectation that we're going to benefit from a community helping us achieve our own goals.*

Respondent G (Ludwig) explained that above all it is beneficial for attracting contributors from other companies, who "prefer not to contribute to projects that are started by companies that could be competitors. They don't trust it as fully open source if it was started by Uber, Google, Facebook, or whatever company." Meanwhile, Respondent B (Elyra) provided a more nuanced view:

> *In reality, in all the projects that I've seen, they are still driven by the main inventors. Open governance just means that the feedback comes from additional contributors ... It's more like a community project.*

However, respondents cautioned that changing the governance model does not guarantee more or useful external contributions. Respondent D (Kedro) explained she had rejected pull requests "because they were not up to scratch," while Respondent G (Ludwig) discussed the role of mentorship in training external developers to become effective contributors, which he highlighted requires the company to invest time and money. Furthermore, diversification of funding is a significant economic incentive for companies. For example, Respondent E (Kedro) highlighted the relevance in their case of client concerns about the financial dependence of the project on the host company:

> *Clients would ask, 'What will happen if your team does not exist tomorrow?' They were afraid that if we left the code, they wouldn't be able to get new versions and then it would become unmaintained.*

Respondent D (Kedro) speculated that the only reason they were able to convince their senior management to approve the open-sourcing of Kedro was due to this client pressure. In addition, some companies seek to attract or increase external investment in their project. Furthermore, harnessing foundation support services was cited as an important incentive by 87% of respondents. For example,

Respondent A (BeyondML) described the resources and infrastructure provided by foundations as "stability offerings" that could help scale and sustain his project beyond the means of his start-up, thus making the donation an attractive option for sustainable project growth.

5.4.2.3 Technological Incentives

Technological incentives at the organisation level include ecosystem integration and adoption, software improvements, faster innovation, and influencing industry standards. Respondents highlighted that joining a foundation offered ecosystem benefits. For example, Respondent C (Elyra) noted:

> *Together with [open governance], we thought being in an ecosystem of other ML and AI projects would foster collaboration and integration, exposing Elyra to more use cases.*

This perspective illustrates how companies view foundation ecosystems as platforms for enhanced collaboration and project visibility, potentially leading to more adoption and more diverse applications of their software. This was confirmed by 88% of respondents who reported increasing adoption as an important incentive. Software improvements and faster innovation were also mentioned as key technological incentives. 69% of respondents cited speeding up development as important, while respondent B (Elyra) commented that, "That's something which is one strength of open source; it's much more properly tested than [proprietary] software." This underlines the aforementioned perspective of Respondent J (ONNX-MLIR) that donations are in large part self-interested actions, which seek to increase adoption, speed up innovation, and enhance software quality.

Finally, influencing industry standards was highlighted as important technological incentive. While 63% of questionnaire respondents cited it as important, Respondent L (LF) explained reflecting on his experience of overseeing over forty OSS donations to both foundations:

> *Every company may have a different set of incentives but what's common across all of them is the desire to make sure that the project becomes successful in the long term and becomes the de facto project for its given functionality.*

Various respondents agreed with this argument. Three respondents cited PyTorch as a standard-setting donation, speculating that Meta sought to make PyTorch the standard for DL frameworks, bringing an end to the age-old rivalry with TensorFlow. For example, Respondent E (Kedro) commented, "I think that move was actually made to destroy TensorFlow because TensorFlow [does not have] open governance." Respondent D (Kedro) suggested that Meta was seeking to beat TensorFlow out of the market since Google would struggle to compete with the strategic alliances of industry giants that had formed under the PyTorch Foundation's governing board (which even included Google). Respondent B (Elyra) was more direct, describing it as "a death knell to TensorFlow."

Figure 5.1: Commercial Incentives for AI OSS Donations to the Linux Foundation
Sample: 16/43 (37%) OSS projects hosted at LF AI & Data Foundation and PyTorch Foundation in October 2022.

5.5 Discussion

5.5.1 Implications for Research and Practice

5.5.1.1 A Taxonomy of Commercial Incentives for OSS Donations

Table 5.4: Commercial Incentives for AI Democratisation via OSS Donations

Category	Individual Incentives	Organisational Incentives
Social	• Personal interests in AI and OSS • Reciprocate to OSS ecosystem • Build personal reputation	• Transition to open governance • Reciprocate to community/ecosystem • Build company's reputation • Join community of companies
Economic	• Professional and career benefits • Harness foundation services	• Recruit new contributors to project • Reduce development costs • Diversify project funding • Harness foundation services • Recruit talent to join company • Increase competitive advantage
Technological	• Ensure project sustainability • Increase visibility and adoption • Contribute to OSAI stack • Access to new collaboration tools	• Increase visibility and adoption • Accelerate innovation • Improve software quality • Foster technical interoperability • Shape industry standards

These findings reveal social, economic, and technological incentives for AI OSS donations at both the developer and organisation levels (see Table 5.4). These insights contribute to a more nuanced understanding of the considerations that shape commercial contributions to the OSAI ecosystem.

The findings corroborate prior work on the salience of economic and technological incentives for companies (Bonaccorsi & Rossi, 2006), with the democratisation of governance in large part used as a means to various economic and technological ends, such as recruiting external developers (Ågerfalk & Fitzgerald, 2008; Fink, 2003), reducing development costs (Birkinbine, 2020; Chesbrough, 2023), and influencing industry standards (Lerner & Tirole, 2005; Lindman et al., 2009). According to the questionnaire, the most important incentive for companies was to recruit new contributors to their project, who could help to improve the quality of the software, ultimately serving the interests of the donor company. As Respondent J (ONNX-MLIR) noted, "there's an expectation that we're going to benefit from a community helping us achieve our own goals." These findings corroborate statements by Big Tech companies about their OSAI strategies, who want to "own the ecosystem and let open source work for us" (Patel & Ahmad, 2023) and for "everyone to be using [Llama] because the more people who are using it, the more the flywheel will spin" (South Park Commons, 2024).

However, the findings also add nuance to this perspective. While the promise of being able to

attract new contributors to their project is indeed important for many companies, it is not universally paramount. For some companies, particularly those with substantial resources or an existing community, other incentives such as standard-setting take precedence (Lerner & Tirole, 2005; Lindman et al., 2009; Widder et al., 2023). The case of PyTorch, which already had a large contributor community, exemplifies this scenario. Several interview respondents suggested that Meta sought to make PyTorch the de facto standard for DL frameworks, bringing to an end a neck-to-neck rivalry with Google's TensorFlow. For example, the donation of PyTorch resulted in the formation of the PyToch Foundation governing board, whose initial members comprised a number of powerful industry giants with significant resources and influence in the AI supply chain, from AI accelerator manufactures (e.g. AMD, Intel, Nvidia) to cloud hyperscalers (e.g. AWS, Google Cloud, Microsoft Azure). Thus, a result of the donation for Meta was the formation of a strategic alliance among industry giants to jointly develop and maintain the DL framework that underpins its own AI R&D. Crucially, it a strategic alliance that Google's TensorFlow will struggle to compete with. As put by Respondent B (Elyra), the donation of PyTorch may represent "a death knell to TensorFlow."

Even for companies primarily seeking to attract new contributors, the findings show that these benefits are far from guaranteed. The interviews identified two key challenges. First, the mere act of a donation does not ensure that a project will attract new contributors. Second, even if a project can recruit new contributors, the quality of their contributions may not meet the project's standards. Companies reject external contributions that "are not up to scratch" and need to invest in mentoring and training resources for external contributors. Companies also need to invest considerable time and resources into community building and contributor development, challenging the notion of open source as a straightforward cost-saving solution.

5.5.1.2 From Intentions to Outcomes: Evaluating the Outcomes of OSS Donations

The findings capture the stated incentives of companies at a particular point in time. However, the actual realisation of these incentives and goals may vary and would require longitudinal study to verify. For instance, the study by Yue and Nagle (2024) on PyTorch's governance transition provides a compelling case study that both supports and challenges some of the incentives identified in this research. Their finding that the governance change led to increased participation from complementors (e.g. chip manufacturers) aligns with the incentive of recruiting new contributors identified in this study. However, the decrease in contributions from Meta following the transition suggests that the realisation of incentives may be more complex than initially anticipated. This complexity is further underscored by the fact that users (e.g. app developers and cloud providers) did not change their rate of participation, indicating that different stakeholders may respond differently to governance

changes. These insights call for more longitudinal studies that track the outcomes of OSS donations over time. Such research could test whether long-term economic and technological benefits are indeed realised. Furthermore, it could shed light on how the balance of contributions shifts between the donor company and external contributors following a governance change.

5.5.1.3 AI Democratisation Incentives Beyond OSS Donations

The literature review and taxonomy and provide a valuable framework for examining commercial incentives behind other AI democratisation efforts, particularly the release of OMs.

Many of the identified incentives, such as building corporate reputation, attracting contributors, and shaping industry standards, likely extend to model releases. However, the specific manifestation of these incentives may differ due to the unique characteristics of OMs compared to OSS. On the one hand, incentives such as building a company's reputation, increasing adoption, recruiting new contributors, reducing development costs, and shaping industry standards likely apply to AI model releases. On the other hand, the adoption of open governance is, at least to this day, specific to the context of OSS donations to foundations. To date, while companies have released models, they have not transferred them to a vendor-neutral foundation for open governance. This may in part be due to the lack of established practices for openly governing OMs and their constituent components (White et al., 2024), or in part because the incentives may be more similar to those for spinning-out software, such as increasing adoption (West & O'Mahony, 2005). At this stage, we can only speculate and, therefore, the commercial strategies for OM releases warrant further investigation.

While this taxonomy provides a starting point for understanding commercial incentives for AI democratisation, it should be used as a flexible framework rather than a definitive list. Future research should build on it by testing its applicability to different approaches to AI democratisation, such as OMs (Seger, Dreksler, et al., 2023) and datasets releases (Don-Yehiya et al., 2024).

5.5.2 Threats to Validity

The validity of the findings are evaluated according to guidelines for qualitative research in software engineering research (Runeson & Höst, 2008; Yin, 2018; Easterbrook et al., 2008).

5.5.2.1 Credibility

Credibility refers to the believability of the findings (Easterbrook et al., 2008). The primary threat stems from the author's affiliation with the LF as a research contractor, creating potential conflicts of interest and research biases, despite the independent purpose and funding of this research study. A

social identity map was used to identify and address potential biases in three areas: data access, collection, and analysis (Jacobson & Mustafa, 2019). As previously stated, while a number of step were taken to enhance the credibility of the research process, it is acknowledged that there is nonetheless a risk of biases that were not controlled for or mitigated and as such it should be understood as an imperfect but best-effort attempt by the author. Another challenge concerned gaining access to data about commercial strategies. The working solution involved drawing on four sources—pre-donation technical pitches, post-donation press releases, questionnaires, and interviews—which offered a triangulated account of incentives (Yin, 2018). While this approach provided four perspectives, it was inevitably constrained by the limited participation of company-affiliated developers, who were willing or allowed to participate in a research interview, and by the fact that some strategic incentives may never be shared publicly. In addition, the willingness of maintainers to complete the questionnaire and/or participate in an interview may have been affected by company policies (e.g. NDAs).

5.5.2.2 Robustness

Robustness concerns strength, reliability, and soundness of the study's design, methods, and findings. There was a risk of social desirability bias or response bias in the interviews, owing to the author's affiliation with the LF. Steps were taken to minimise the risk of these biases, such as proactively communicating the independent purpose and funding of this study in the interview invitations and at the beginning of every interview. Another threat stems from the thematic analysis. The author sought to maximise the robustness of this analysis by following best-practice guidelines (Braun & Clarke, 2006) and integrating approaches to coding qualitative data (Cruzes & Dybå, 2011). In addition, the structure of the taxonomy poses risks to the validity of the findings. In some cases, it was difficult to demarcate incentives at the two units of analysis, which may have led to a misclassification or oversimplification. That being said, efforts were made to develop the categories by thoroughly reviewing prior work and member-checking findings with interviewees to evaluate their accuracy and resonance with practitioners (Lincoln & Guba, 1985).

5.5.2.3 Transferability

Transferability concerns the generalisability of qualitative research findings. There are two key threats to generalisability. First, the narrow focus on OSS donations as a method of AI democratisation limits the generalisability of the taxonomy, which both includes aspects specific to OSS donations (e.g. democratising governance or foundation support) and excludes aspects that may be typical of other methods of AI democratisation (e.g. OM releases). Future research directions were recommended in Section 5.5.1 to advance our understanding of the commercial incentives for different

approaches to AI democratisation. Second, this study is limited by the specific characteristics of the sample of companies, OSS projects, and the foundations. The sample was drawn from the LF AI & Data Foundation and PyTorch Foundation, two AI OSS foundations under the LF, which may not represent incentives for donations to other foundations. In addition, companies and individuals who were willing to participate in this study may systematically differ from those who declined or were unreachable, which may skew the findings in, for example, positive experiences of OSS donations.

5.5.2.4 Reliability

Reliability refers to the consistency and replicability of the research procedure. A comprehensive list of secondary documents collected for each OSS project is provided in Table 5.5. These documents were triangulated by questionnaires and interviews with project donors and staff from the foundations. With their consent, all interviews were recorded and transcribed to aid the analysis (Yin, 2018). With regards to the data analysis, guidelines were followed for the systematic analysis of qualitative data (Braun & Clarke, 2006; Cruzes & Dybå, 2011). Additionally, the accuracy of the analysis was validated through member-checking findings with interviewees, ensuring the resonance of the findings with developers who have donated OSS projects themselves.

5.6 Conclusion

Companies increasingly "democratise" AI through various open source approaches, which are often celebrated for advancing science and innovation. However, the strategic incentives driving these efforts warrant careful examination, as they shape which technologies are available, built, and how they are governed. This study investigated commercial incentives for AI OSS donations through mixed-methods analysis of 43 donations to the LF. The findings reveal an interplay of social, economic, and technological incentives at both developer and organisation levels. Companies primarily treat governance democratisation as a means to economic and technological ends, such as attracting contributors, reducing costs, and influencing industry standards, among others. Beyond OSS donations, the taxonomy of commercial incentives provides a theoretical foundation and practical toolkit for understanding other AI democratisation efforts like OM releases. While some incentives are unique to OSS donations, many extend to OM releases, such as increasing adoption, building an ecosystem, and crowdsourcing contributions. As the number of OM releases continues to grow at a rapid pace, it is timely for researchers to examine the commercial strategies driving these releases, as well as their effects on the development practices and norms in the OSAI ecosystem.

5.7 Appendix for Chapter 5 (RP2)

5.7.1 OSS Projects at the LF AI & Data Foundation and PyTorch Foundation

Table 5.5: OSS Donations to the LF AI & Data Foundation and PyTorch Foundation

Project	Company	Date	TAC Proposal	Press Release
Acumos	AT&T, Tech Mahdra	2018-05	https://drive.google.com/file/d/1L6fFhZnFqeR3mwy8Ya/KVoRGzCUjdnyrM	https://www.acumos.org/news/2018/11/14/lf-deep-learning-delivers-first-acumos-ai-release-making-it-easier-to-deploy-and-share-artificial-intelligence-models/
Angel	Tencent	2018-08	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/18481216/GMT20191121-140452_LF-AI-Foun_1920x1080.mp4	https://www.linuxfoundation.org/press/press-release/lf-deep-learning-adds-two-new-framework-projects-to-expand-community-and-ecosystem
Egeria	IBM, ING	2018-08	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30408948/September%2024%2C%202020_LF%20AI%20TAC%20Deck%202.pptx	https://www.linuxfoundation.org/press/press-release/new-ai-data-foundation-combines-industrys-fastest-growing-open-source-developments-in-artificial-intelligence-and-open-data
Elastic Deep Learning	Baidu	2018-08	N/A	https://www.linuxfoundation.org/press/press-release/lf-deep-learning-adds-two-new-framework-projects-to-expand-community-and-ecosystem

Project	Company	Date	TAC Proposal	Press Release
Horovod	Uber	2018-12	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30408797/August%2013%2C%202020_LF%20AI%20TAC%20Deck%20-%20compressed2.pptx	https://www.uber.com/en-GB/blog/horovod-deep-learning-foundation/
Pyro	Uber	2019-01	https://drive.google.com/file/d/1B0ZkJUKVZoJxsaUkge/02kGKddiZX8\acrshort{dl}j/view	https://www.uber.com/en-GB/blog/pyro-lf-deep-learning-foundation/
Adlik	ZTE	2019-09	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=%2F7733341%2F12091694%2FTAC+recording+08152019.mp4	https://lfaidata.foundation/blog/2019/10/21/lf-ai-welcomes-adlik-as-newest-incubation-project/
ONNX	ONNX Community	2019-11	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/18481160/TAC%2010-24-2019.mp4	https://cloudblogs.microsoft.com/opensource/2019/11/14/onnx-joins-linux-foundation/
Marquez	WeWork	2019-12	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/18481417/TAC-12192019.mp4	N/A

Project	Company	Date	TAC Proposal	Press Release
sparklyr	RStudio	2019-12	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/18481269/TAC-recording-12052019.mp4	https://lfaidata.foundation/blog/2020/01/29/sparklyr-joins-lf-ai-as-its-newest-incubation-project-scaling-data-science-and-machine-learning-workflows-using-apache-spark-and-r/
Milvus	Zilliz	2020-01	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/22249639/January%2016%2C%202020_LF%20AI%20TAC%20Deck.pdf	https://www.prnewswire.com/news-releases/milvus-the-ai-search-engine-originally-developed-by-zilliz-joins-lf-ai-as-new-incubation-project-301038716.html
OpenDS4All	IBM, ODPi, UPenn	2020-02	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30408948/September%2024%2C%202020_LF%20AI%20TAC%20Deck%202.pptx	https://community.ibm.com/community/user/ai-datascience/blogs/ana-echeverri1/2020/02/28/opends4all-is-live
NNStreamer	Samsung	2020-03	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/24281106/March%2012%2C%202020_LF%20AI%20TAC%20Deck_v2.pdf	https://research.samsung.com/news/LF-AI-Foundation-Announces-NNStreamer-as-Its-Newest-Incubation-Project

Project	Company	Date	TAC Proposal	Press Release
ForestFlow	DreamWorks Animation	2020-04	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/24281142/March%2026%2C%202020_LF%20AI%20TAC%20Deck.pdf	https://research.dreamworks.com/dreamworks-animation-releases-forestflow-machine-learning-model-server-to-the-open-source-community/
Ludwig	Uber	2020-05	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/24281544/Ludwig%20LFAI.pdf	https://www.uber.com/blog/introducing-ludwig/
Adversarial Robustness Toolbox	IBM	2020-06	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30409001/October%208%2C%202020_LF%20AI%20TAC%20Deck.pdf	https://developer.ibm.com/open/centers/codait/trusted-ai/
AI Explainability 360	IBM	2020-06	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30409001/October%208%2C%202020_LF%20AI%20TAC%20Deck.pdf	https://developer.ibm.com/open/centers/codait/trusted-ai/
AI Fairness 360	IBM	2020-06	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30409001/October%208%2C%202020_LF%20AI%20TAC%20Deck.pdf	https://developer.ibm.com/open/centers/codait/trusted-ai/

Project	Company	Date	TAC Proposal	Press Release
Amundsen	Lyft	2020-07	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341	https://eng.lyft.com/amundsen-1-year-later-7b60bf28602
DELTA	DiDi	2020-09	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30408797/August%2013%2C%202020_LF%20AI%20TAC%20Deck%20-%20compressed2.pptx	https://lfaidata.foundation/blog/2021/06/17/delta-joins-lf-ai-data-as-new-incubation-project/
Feast	Gojek	2020-09	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/30408948/September%2024%2C%202020_LF%20AI%20TAC%20Deck%202.pptx	https://feast.dev/blog/feast-joins-the-linux-foundation-for-ai-data/
SOAJS	Herron Tech	2020-09	https://wiki.lfaidata.foundation/download/attachments/7733341/September%2010%2C%202020_LF%20AI%20TAC%20Deck%20-%20updated.pdf?version=1&modificationDate=1599741898000&api=v2	N/A
DataPractices	DataPractices Org	2020-12	https://wiki.lfaidata.foundation/download/attachments/7733341/October%205%2C%202020_LF%20AI%20TAC%20Deck%282%29.pdf?version=1&modificationDate=1604497702000&api=v2	N/A

Project	Company	Date	TAC Proposal	Press Release
JanusGraph	JanusGraph Community	2021-01	https://wiki.lfaidata.foundation/download/attachments/7733341/December%203%2C%202020_LF%20AI%20TAC%20Deck.pdf?version=2&modificationDate=1606864208000&api=v2	https://lfaidata.foundation/blog/2021/01/12/janusgraph-joins-lf-ai-data-as-new-incubation-project/
Flyte	Lyft	2021-02	https://wiki.lfaidata.foundation/download/attachments/7733341/February%2025%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1614021188000&api=v2	https://eng.lyft.com/flyte-joins-lf-ai-data-48c9b4b60eec
Datashim	IBM	2021-03	https://wiki.lfaidata.foundation/download/attachments/7733341/January%2014%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1610631576000&api=v2	https://lfaidata.foundation/blog/2021/03/23/datashim-joins-lf-ai-data-as-new-incubation-project/
RosaeNLG	BNP Paribas	2021-03	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/39092353/March%2011%2C%202021_LF%20AI%20TAC%20Deck(2).pdf	https://lfaidata.foundation/blog/2021/04/28/rosaenlg-joins-lf-ai-data-as-new-sandbox-project/

Project	Company	Date	TAC Proposal	Press Release
Substra	OWKIN	2021-03	https://wiki.lfaidata.foundation/download/attachments/7733341/March%2025%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1616590543000&api=v2	https://lfaidata.foundation/blog/2022/11/28/owkin-launches-open-science-push-by-open-sourcing-ai-software-substra-and-releasing-two-open-source-ai-innovations-at-neurips/
Kompute	The Institute for Ethical AI	2021-05	https://wiki.lfaidata.foundation/download/attachments/7733341/May%206%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=2&modificationDate=1620316765000&api=v2	https://lfaidata.foundation/blog/2021/08/26/kompute-joins-lf-ai-data-as-new-sandbox-project/
OpenLineage	Datakin, IBM	2021-07	https://wiki.lfaidata.foundation/download/attachments/7733341/December%2015%202022_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1673523471000&api=v2	https://openlineage.io/blog/joining-lfai/
TonY	LinkedIn	2018-09	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/43287136/July%2015%2C%202021_LF%20AI%20TAC%20Deck.pdf	https://engineering.linkedin.com/blog/2018/09/open-sourcing-tony--native-support-of-tensorflow-on-hadoop

Project	Company	Date	TAC Proposal	Press Release
Kedro	McKinsey QuantumBlack	2021-08	https://wiki.lfaidata.foundation/download/attachments/7733341/August%2026%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1630089002000&api=v2	https://medium.com/quantumblack/kedro-joins-the-linux-foundation-to-become-an-open-standard-for-machine-learning-engineering-b0061152ff73
KServe	KServe Community	2021-11	https://wiki.lfaidata.foundation/download/attachments/7733341/October%2018%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1637936838000&api=v2	https://lfaidata.foundation/blog/2022/02/24/kserve-joins-lf-ai-data-as-new-incubation-project/
OpenBytes	Graviti	2021-11	https://wiki.lfaidata.foundation/download/attachments/7733341/October%2021%2C%202021_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1635325659000&api=v2	https://www.linuxfoundation.org/press/press-release/linux-foundation-and-graviti-announce-project-openbytes-to-make-open-data-more-accessible-to-all
Artigraph	Replica	2022-01	https://wiki.lfaidata.foundation/download/attachments/7733341/January%2027%2C%202022_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1643715756000&api=v2	https://lfaidata.foundation/uncategorized/2022/04/13/lf-ai-data-announces-artigraph-as-new-sandbox-project/

Project	Company	Date	TAC Proposal	Press Release
1chipML	Ericsson	2022-04	https://wiki.lfaidata.foundation/download/attachments/7733341/April%207%2C%202022_LF%20AI%20TAC%20Deck%20%281%29.pdf?version=1&modificationDate=1650500517000&api=v2	https://lfaidata.foundation/blog/2022/07/21/lf-ai-data-announces-three-new-sandbox-projects/
BeyondML	Squared AI	2022-06	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/61964452/June%2016%2C%202022_LF%20AI%20TAC%20Deck.pdf	https://lfaidata.foundation/blog/2022/07/21/lf-ai-data-announces-three-new-sandbox-projects/
FlagAI	BAAI	2022-06	https://wiki.lfaidata.foundation/download/attachments/7733341/June%2030%2C%202022_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1657325720000&api=v2	N/A
FedLCM	VMWare	2022-08	https://wiki.lfaidata.foundation/download/attachments/7733341/July%2028%2C%202022_LF%20AI%20TAC%20Deck%20%281%29.pdf?version=1&modificationDate=1660155848000&api=v2	https://blogs.vmware.com/opensource/2022/10/27/open-source-project-fedlcm-to-the-lf-ai-data/

Project	Company	Date	TAC Proposal	Press Release
FATE	LinkedIn	2022-08	https://wiki.lfaidata.foundation/download/attachments/7733341/August%2025%2C%202022_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1662390660000&api=v2	https://cloudblogs.microsoft.com/opensource/2022/09/12/feathr-feature-store-joins-lf-ai-data-foundation/
OpenDataology	OpenDataology Community	2022-08	https://wiki.lfaidata.foundation/download/attachments/7733341/August%2011%2C%202022_LF%20AI%20TAC%20Deck.pdf?version=1&modificationDate=1661373254000&api=v2	N/A
PyTorch	Meta	2022-09	N/A	https://pytorch.org/blog/PyTorchfoundation/
Elyra	IBM	2022-10	https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=7733341&preview=/7733341/39092353/March%2011%2C%202021_LF%20AI%20TAC%20Deck(2).pdf	https://developer.ibm.com/blogs/open-source-elyra-ai-toolkit-simplifies-data-model-development/

6. The AI Community Building the Future? A Quantitative Analysis of Development Activity on Hugging Face Hub

Peer review status: *This RP was accepted for presentation at the International Conference on Computational Social Science (IC2S2) and published in the Journal of Computational Social Science.*

Co-authorship statement: *I am the first-author of this RP alongside Jennifer Ding and Hannah Kirk. I document our relative contributions in Section 6.7.4. N.B. I use "we" rather than "I" due to co-authorship.*

Abstract

The developers of OMs have emerged as key actors in the political economy of AI, but we still have a limited understanding of development or collaboration practices in the OM ecosystem. This paper responds to this gap with a three-part quantitative analysis of development activity on HF Hub, a popular platform for building, sharing, and demonstrating OMs and open datasets. This analysis contributes three key findings. First, various types of activity, from likes to commits, across 348,181 model, 65,761 dataset, and 156,642 space repositories exhibit right-skewed distributions. Activity is extremely imbalanced between repositories; for example, over 70% of models have 0 downloads, while 1% account for 99% of downloads. Furthermore, licenses matter: we find that there are statistically significant differences in collaboration patterns in model repositories with permissive, restrictive, and no licenses. Second, the social network structure of collaboration in model repositories has a core-periphery structure, with a core of prolific developers and a majority of isolate developers (89%). Ignoring the isolates, collaboration is characterised by high reciprocity regardless of developers' network positions. Third, we examine model adoption through the lens of model adoption in spaces, finding that a minority of models, developed by a handful of industry leaders, are widely used on HF Hub. The findings provide an empirical baseline for further study on OM development and adoption, whilst challenging prevalent AI democratisation narratives. We conclude with recommendations for researchers and practitioners to advance our understanding of OM development.

6.1 Introduction

Open source developers have become central actors in the political economy of AI. The rise of OSAI, specifically the practices of releasing and fine-tuning publicly available pre-trained models (OMs), has extended open science practices crucial to AI advances (Langenkamp & Yue, 2022; White et al., 2024), including OSS and open access to research papers (arXiv, 2024) and datasets (Kaggle, 2024; CommonCrawl, 2024; ImageNet, 2024). OSAI has attracted attention as a potential challenger to the dominance of start-ups and Big Tech in AI R&D (Ahmed et al., 2023; Tarkowski, 2023). Grassroots initiatives like EleutherAI (EleutherAI, 2021), BigScience (Akiki et al., 2022), and BigCode (HuggingFace, 2024a) have shown the feasibility of OM development (Ding et al., 2023), while HF Hub has emerged as a popular platform used by millions to host, download, and collaborate on a growing number of models, datasets, and spaces[1] (HuggingFace, 2024d).

While the benefits and risks of OSAI have been widely debated (e.g. Law & Krier, 2023; Solaiman, 2023; Kapoor et al., 2024; Seger, Dreksler, et al., 2023; Eiras et al., 2024), the collaborative practices in OM development have received relatively little attention. To date, only a handful of scholars have explored some aspects of OM development, including user contributions to grassroots initiatives (Akiki et al., 2022; Ding et al., 2023), commercial participation in model development (Ding et al., 2023; Widder et al., 2023), model maintenance (Castaño et al., 2024), and the processes and tools used by open data communities (Heltweg & Riehle, 2023).

We contribute to this nascent research agenda with a three-part analysis of development activity on HF Hub. First, we investigate activity patterns in 348,181 model, 65,761 dataset, and 156,642 space repositories on HF Hub (**RQ1**). Subsequently, we examine the social network structure of the developer community as well as collaboration practices among developers through the lens of code contributions to model repositories (**RQ2**). We replicate this analysis for models with natural language processing (NLP), computer vision (CV), and multimodal (MM) tags for a comparative analysis of the social network structures of collaboration in these respective sub-fields. Finally, we quantify model adoption through the lens of model usage in spaces on HF Hub (**RQ3**), providing insights into the widespread use of a minority of models on HF Hub and the actors driving their development.

Overall, our analysis reveals that various aspects of development activity on HF Hub—e.g. interactions in model, dataset, and space repositories; collaboration in model repositories; and model adoption in spaces—exhibit right-skewed, Pareto distributions, which is a well-documented pattern in OSS development (Goeminne & Mens, 2011; Mockus, Fielding, & Herbsleb, 2002; Szymański &

[1] On HF Hub, spaces are web applications to demonstrate and try out models

Ochodek, 2023; J. Xu et al., 2006; Y. Zhang et al., 2021). While the OM development life-cycle involves unique practices which differ from OSS development (Castaño et al., 2024), such as model training and fine-tuning, the observed similarities in the overall patterns of activity suggests that future research on OSAI can benefit from drawing on the extensive, multidisciplinary literature on the social dynamics of OSS development. Based on our findings, we propose a number of recommendations to facilitate research and evidence-based policy discussions on OSAI.

The paper has the following structure. First, it provides an overview of prior work on both OSAI and OSS development in order to draw comparisons between OSS and OM development practices (Section 6.2). Second, we present the RQs and research design (Section 6.3). Third, we introduce the main findings from the three-part analysis (Section 6.4). Fourth, we discuss the contributions of the findings to research and practice, and make recommendations for research and practice (Section 6.5). We conclude with a discussion of what further clarification of the practices in OM development can offer for OSAI researchers, developers, policymakers, and platform providers (Section 6.6).

6.2 Related Work

6.2.1 "We Have No Moat": The Emergence of OSAI

Open science practices, from the development of OSS to the provision of open access research papers—e.g. via arXiv (arXiv, 2024)— and datasets—e.g. via Kaggle (Kaggle, 2024), ImageNet (ImageNet, 2024), or Common Crawl (CommonCrawl, 2024)—have been integral to advances in AI R&D (White et al., 2024). The culture of openness in AI has evolved significantly in the last 15 years (Gururaja et al., 2023). In 2007, 16 researchers lamented the lack of OSS that standardised the implementation of ML algorithms, highlighting this as a major obstacle to advances in AI research (Sonnenburg et al., 2007). Yet today AI R&D is simply unimaginable without OSS (Langenkamp & Yue, 2022), drawing on a growing commons of over 300 AI OSS libraries (Haddad, 2022), over a million AI OSS repositories (GitHub, 2024a), and over a million OMs (HuggingFace, 2024h).

Following years of debate about the safety of openly releasing AI models (Solaiman et al., 2019; Kapoor et al., 2024), in recent years we have seen the emergence and proliferation of OMs, which individuals and organisations share on an open access basis on platforms such as HF Hub (White et al., 2024). Prior to this, AI models, in particular LLMs, were principally developed and maintained behind closed doors. The start of this trend is attributed to EleutherAI, a grassroots research group, which formed on a Discord server with the intention of developing an open source variant to OpenAI's GPT, resulting in The Pile in December 2020 (Gao et al., 2020), a library of datasets for training LLMs, and GPT-Neo in March 2021 (Black et al., 2022). Subsequently, OMs gained more visibility

with the release of other state-of-the-art AI models (Tarkowski, 2023), including BLOOM by the BigScience workshop in July 2022 (Scao et al., 2023), Stable Diffusion by Stability AI in August 2022 (Stability.AI, 2022), and Llama 2 by Meta in July 2023 (Meta, 2023), among others.

The proliferation of OMs, especially open FM, has ignited heated debate about their potential benefits and risks (Kapoor et al., 2024; Law & Krier, 2023; Solaiman, 2023; Bdeir & François, 2024; Seger, Dreksler, et al., 2023; Eiras et al., 2024). On the one hand, OMs are said to promise benefits for research, innovation, and competition by lowering entry barriers and widening access to state-of-the-art AI technologies (Cihon, 2024). Furthermore, drawing on Linus' Law from OSS development that "given enough eyeballs, all bugs are shallow" (Raymond, 2001a), proponents argue that the distributed development and auditing of OMs offer safety advantages (Wladawsky-Berger, 2023). In addition, access to OMs lowers the barriers for adaptability and customisation for diverse language contexts (Pipatanakul et al., 2023; Kapoor et al., 2024). On the other hand, OMs can pose risks of harm by both well-intended and malicious actors, including the creation of deepfakes (Nguyen et al., 2022; Lakatos, 2023; Thiel et al., 2023), disinformation (Goldstein et al., 2023; Musser, 2023), and malware (Tsamados et al., 2023; C. David & Paul, 2023). A study by 25 experts concluded that OMs have five distinctive properties that present *both* benefits and risks: broader access, greater customisability, local adaptation and inference ability, the inability to rescind model access, and the inability to monitor or moderate model usage (Kapoor et al., 2024).

The development of OMs has been described as a potential challenge to the dominance of Big Tech companies in AI R&D (Ahmed et al., 2023; Gulson & Webb, 2021). This was underlined by a leaked Google memo that claimed there is "no moat" around closed-source AI development and "open source solutions will out-compete companies like Google or OpenAI" (Patel & Ahmad, 2023). There is much excitement about the potential benefits of OSAI. Venture capitalists have bullishly invested in OSAI start-ups (Wiggers, 2023; Abboud et al., 2024), and world leaders like President Macron of France have pledged public funds to support OSAI (Chatterjee & Volpicelli, 2023). In addition, the Mozilla Foundation has launched mozilla.ai with $30 million in investment to build a trustworthy, independent, and OSAI ecosystem "outside of Big Tech and academia" (Mozilla, 2023). While proponents champion OMs as good news for innovation and competition, others temper this optimism by pointing to market concentrations at several layers of the AI stack, from chips to cloud compute infrastructure, which remain unchallenged by innovations stemming from OSAI communities (Lehdonvirta et al., 2023; Srnicek, 2022; Widder et al., 2023).

A myriad of meanings are attached to "open source AI." Often this term is understood as making pre-trained models, parameters (or weights), and documentation available on platforms like HF Hub. In some cases, it refers to open collaboration on the development of models (Ding et al., 2023). The

description of OMs as “open source” has been fiercely contested for failing to meet OSS standards as defined by the OSI (OSI, 2007; White et al., 2024; Maffulli, 2023; Nolan, 2023). For example, when Meta imposed limits on use of Llama 2, Stefano Maffulli from the OSI commented, “Unfortunately, the tech giant has created the misunderstanding that Llama 2 is ‘open source’ – it is not. Meta is confusing ’open source’ with ‘resources available to some users under some conditions,’ [which are] two very different things” (Maffulli, 2023). Companies have also been criticised for “open-washing” by promoting their models as “open source,” when they are typically “open weight” models, as a commercial strategy to present themselves as patrons of openness, whilst disguising their intent to set standards and benefit from crowdsourced innovation (Widder et al., 2023; Srnicek, 2022). Furthermore, a review of open LLMs found that, “[W]hile there is a fast-growing list of projects billing themselves as ‘open source’, many inherit undocumented data of dubious legality, few share the all-important instruction-tuning (a key site where human annotation labour is involved), and careful scientific documentation is exceedingly rare” (Liesenfeld et al., 2023).

It remains an open question whether one can or should classify open AI models as open source. The OSI has defined OSAI systems as AI systems that are made available under terms that grant the freedoms to use, study, modify, and share the system (OSI, 2024c). Countering this binary approach, Solaiman (2023) makes the case that AI systems are not either fully open or fully closed; rather, the openness of AI systems can be plotted along a gradient with six degrees of openness. Each grade of openness involves trade-offs between concentrating power and mitigating risks (Solaiman, 2023).

6.2.2 The Nascent Research Agenda on OSAI Development

While the benefits and risks of OMs have been widely discussed, we still have a limited understanding of the collaborative practices involved in their development both prior to and following their public release. In this section, we review prior work on OM development and motivate our empirical research to address this research gap.

HF Hub has emerged as a popular platform used by individuals and organisations to share, download, and collaborate on models, datasets, and spaces (HuggingFace, 2024e; Ait et al., 2023b). HF Hub is a “model marketplace,” which is “a new form of user-generated content platform, where users can upload AI systems and AI-related datasets, which in turn can be downloaded, and depending on the business model, queried, tweaked, or built upon by other users” (Gorwa & Veale, 2024). Much of the activity among the emerging developer community on this platform involves individuals fine-tuning pre-trained models that were released by industry leaders for downstream use in research and applications (Widder et al., 2023). In addition, a few grassroots initiatives have embraced open collaboration methods to develop OMs. For example, the development of BLOOM, a 176B parameter

multilingual LLM, and its training dataset, ROOTS, was the largest "open source" AI collaboration to date, involving over 1,000 volunteers from over 70 countries and over 250 institutions (Akiki et al., 2022). Such initiatives have demonstrated alternative pathways for AI development beyond the handful of companies that dominate AI R&D (Ahmed et al., 2023; Ding et al., 2023). Prior work has also highlighted the leadership role of companies, such as HF, in organising "values-driven initiative[s]," such as the BigScience workshop, and attracting contributors who have diverse motivations, from developing new skills and working on new problems to publishing research and giving back to the ecosystem (Ding et al., 2023; Akiki et al., 2022).

Due to the growing popularity of HF Hub, scholars have examined the suitability of HF Hub for empirical research on OM development (Ait et al., 2023b, 2023a).[2] Castaño et al. (Castaño et al., 2024) provide most comprehensive empirical insights into maintenance practices in model repositories on HF Hub.[3] They find that commit activity follows a right-skewed distribution, with a few models receiving extensive activity while the majority of repositories receive limited activity (Castaño et al., 2024). While the majority of models are developed by singular developers (1.18 mean, 1.0 median), some model repositories, such as $bigscience/bloom$ or $bigcode/santacoder$, are co-developed and co-maintained by up to 20 developers (Castaño et al., 2024). They also find that developers tend to prioritise "perfective tasks" to enhance model performance and align with technological advances, unlike OSS maintenance that focuses on bug fixes and feature additions (Castaño et al., 2024). The authors contend this "reveals the need for methods and tools specifically designed for the unique demands of ML model maintenance. Such tools may include advanced version control systems optimized for data and model tracking, as well as automated monitoring tools capable of detecting model drift or degradation" (Castaño et al., 2024).

Our research builds on this prior work. As one of the first studies to investigate OM development practices, in the next section we draw on prior work on OSS development in order to be able to compare our findings to prior research and to lay the groundwork for a more comprehensive understanding of OM development in the future.

6.2.3 Learning from Prior Work on OSS development

Prior work on OSS development provides a baseline for understanding the social dynamics of OM development. In the early 2000s, a number of metaphors were used to describe the social structure of "the OSS community." For example, the Linux community was described as a "bazaar" that vibrated

[2]The authors define "suitability for empirical research" as "the amount and adequacy of the features to enable software development practices and the sufficient quantity of data to enable the conduction of empirical studies about such practices" (Ait et al., 2023b).

[3]Model maintenance is defined as "a higher number of commits, regular commit frequency, shorter intervals between commits, fewer days without commits, and a slightly higher number of authors" (Castaño et al., 2024).

with the activity of geeks, hackers, and hobbyists, who performed various tasks from bug-spotting to writing code to "serving the hacker culture itself" (Raymond, 2001a). However, prior work illustrates that OSS communities have diverse social structures (Eghbal, 2020; M. Zhou et al., 2016), from "caves" with singular developers (Krishnamurthy, 2005b) to "core-periphery" networks, akin to "layered onions" (Crowston, Annabi, Howison, & Masango, 2005), with uneven activity distributions ranging from core contributors, such as project initiators, to users, such as bug-spotters (Bird et al., 2006; Crowston & Howison, 2006; Long & Siau, 2007; Orucevic-Alagic & Host, 2014).

Numerous studies highlight that various types of activity in OSS development, such as discussions in mailing lists, bug-spotting in issue trackers, and commit activity, exhibit right-skewed, Pareto distributions (Goeminne & Mens, 2011; Szymański & Ochodek, 2023; Y. Zhang et al., 2021). Indeed, it is well-documented observation that OSS development is typically characterised by the Pareto principle, commonly known as the 80/20 rule or the law of the vital few, which states that approximately 80% of effects come from 20% of causes (M. Newman, 2005). These findings are congruent with a wide range of Internet phenomena, which similarly exhibit right-skewed distributions (Faloutsos, Faloutsos, & Faloutsos, 1999; Mahanti, Carlsson, Mahanti, Arlitt, & Williamson, 2013).

However, there are exceptions to the rule; for example, a study of 2,496 projects on GitHub found that the Pareto principle does not always characterise development activity in OSS repositories, thus highlighting the need to be cautious about generalising the Pareto principle as an incontestable law of OSS development (Yamashita, McIntosh, Kamei, Hassan, & Ubayashi, 2015). Furthermore, many activities, such as mentorship and hackathons, take place outside of the repository (Geiger et al., 2021; Hossain, 2021; Osborne, 2024a; Takhteyev, 2012) and are, therefore, invisible to quantitative scholars who focus on observable and quantifiable OSS development activity.

The various social structures of OSS communities are shaped, among others, by the diverse incentives of individuals and companies that participate in them (Feller & Fitzgerald, 2002; Bonaccorsi & Rossi, 2006; X. Li et al., 2024). Individuals are typically motivated by factors such as personal values, altruism, enjoyment, reputation-building, and career benefits (von Krogh et al., 2012; Shah, 2006; K. R. Lakhani & Wolf, 2003; Ghosh et al., 2002). However, there are also major barriers to participation, including gender disparities (Brooke, 2021; Vasilescu et al., 2014) and geographic inequalities (Hossain, 2021; Takhteyev, 2012). Activity tends to be concentrated in the Global North (Braesemann et al., 2019) and the English *lingua franca* is a barrier for many developers (Takhteyev, 2012; Williams, 2023). Furthermore, the incentives of OSS developers vary by geography: while developers in the USA show a relatively strong interest in "geek culture," developers in India and China tend to be motivated primarily by career benefits (Subramanyam & Xia, 2008). Thus, "researchers studying open source should be mindful of geographic variation in what motivates participation and

what forms participation may take, particularly outside of the code repository" (Hossain, 2021).

Meanwhile companies primarily participate in OSS development for strategic reasons, such as recruiting developers (Ågerfalk & Fitzgerald, 2008; Birkinbine, 2020; West & Gallagher, 2006), reducing costs (Chesbrough, 2023; Lindman et al., 2009; Birkinbine, 2020), influencing OSS projects (Dahlander & Wallin, 2006; Lindman et al., 2009), promoting open standards (Chesbrough, 2023; Lerner & Tirole, 2002), and building a reputation as an OSS patron (Bonaccorsi & Rossi, 2006; Pitt et al., 2006; Osborne, 2024a). Commercial participation has mixed effects on the social structure of OSS communities. Typically, one company or a few companies emerge as dominant contributors in projects (Y. Zhang et al., 2021; Nguyen-Duc et al., 2019). The dominance of a company is negatively associated with the participation of volunteers, while it is positively associated with the productivity of contributors and the quality of issue reports (Y. Zhang et al., 2022; M. Zhou et al., 2016). It is also common for companies, including market rivals, to collaborate (Germonprez et al., 2013; Nguyen-Duc et al., 2019; Teixeira & Lin, 2014; Y. Zhang et al., 2020), turning many OSS communities "from networks of individuals into networks of companies" (Ågerfalk & Fitzgerald, 2008, p.396).

Building on this prior work, this study aims to provide novel insights into the collaborative dynamics in OM development. Specifically, we investigate patterns of activity across model, dataset, and space repositories HF Hub (**RQ1**), the social network structure of the platform's developer community (**RQ2**), as well as the adoption of and actors who develop the most widely-adopted models (**RQ3**). The findings contribute to a more comprehensive understanding of and provide an empirical foundation for future research on OM development.

6.3 Study Design

6.3.1 Research Aims & Research Questions

This study extends the nascent research agenda on OSAI development with a quantitative analysis of development activity on HF Hub. We adopted a quantitative approach to explore large-scale patterns and trends in development activity on HF Hub, which is a suitable approach when one seeks to generate baseline insights on a new phenomenon (Easterbrook et al., 2008). In particular, we examine different aspects of development activity on HF Hub via the following RQs:

- **RQ1**: What are typical patterns of development activity across HF Hub?
- **RQ2**: What is the social network structure of the HF developer community?
- **RQ3**: What is the distribution of model development and adoption activity on HF Hub, and which actors develop the most adopted models?

These RQs examine different aspects of development activity on HF Hub. **RQ1** focuses on identifying common patterns across various types of activity, such as likes, discussions, commits, and downloads, in the repositories of models, datasets, and spaces, expanding prior work that focuses on commit activity in model repositories (Castaño et al., 2024). **RQ2** concerns the social network structure of the developer community on HF Hub. In particular, we analyse a snapshot of collaboration interactions in model repositories among around 100,000 developers, building on prior descriptions of OM collaborations (Ding et al., 2023; Akiki et al., 2022) and maintenance (Castaño et al., 2024). Lastly, **RQ3** empirically tests a prior observation of uneven model adoption and the influence of Big Tech companies (Widder et al., 2023) by examining the distribution of model use in spaces and identifying the developers of the most used models. In addition, we examine model co-usage patterns to provide insights into the interdependencies and ecosystems surrounding popular OMs.

6.3.2 HF Hub: A New Platform & Source of Research Data

HF Hub was launched in 2021 by HF, a start-up whose mission is to "democratise AI" (HuggingFace, 2024e). HF Hub is a Git-based social coding platform, widely used by researchers, developers, and hobbyists to share, discover, discuss, and collaborate on OMs (HuggingFace, 2024h), datasets (HuggingFace, 2024b), and spaces (HuggingFace, 2024j). Spaces are interactive web applications that facilitate the creation of demonstrations and make models hosted on the platform more accessible to end-users. The platform provides a number of tools for OM development, such as version control for collaboration and tracking (HuggingFace, 2024h), and evaluation and benchmarking of model performance (HuggingFace, 2024c). HF Hub API allows programmatic access to platform resources as well as metadata from repositories hosted on the platform (HuggingFace, 2024d). In light of its features and data availability, prior work underlines the platform's suitability for empirical studies on OM development (Ait et al., 2023b; Castaño et al., 2024). Building on this prior work, this paper aims to advance the research community's understanding of the development practices in OM development as well as methodological considerations regarding HF Hub.

When using data from HF Hub, it is important to consider the ethical implications and adhere to the platform's terms of service. In the study, we only collected publicly available data through the official HF Hub API, respecting the privacy settings of users and repositories. For example, we did not attempt to access or include data from private repositories in the analysis. Additionally, we anonymised the collected data by focusing on aggregate measures and avoiding the disclosure of personally identifiable information in the findings. Ethical clearance for this study was obtained from the CUREC institutional review board at the University of Oxford.

6.3.3 Data Collection

We collected data via HF Hub's API in October 2023 (HuggingFace, 2024d), using Python scripts that are available on GitHub (Osborne, Daneshyan, et al., 2024b). For **RQ1**, we collected and processed metadata for a number of activities from the public repositories of 348,181 models, 65,761 datasets, and 156,642 spaces, using the `list_models()`, `list_datasets()`, and `list_spaces()` API endpoints. These included: likes (`n_likes`), downloads (`n_downloads`)[4], discussions (`n_discussions`), commits (`n_commits`), unique developers who have contributed commits (`n_commiters`), unique developers who started discussions (`n_disc_starters`),[5] and the repository's community size (`n_community`), calculated as the cardinality of the set union of `n_disc_starters` and `n_commiters`. As per prior work (Lin et al., 2017; Y. Zhang et al., 2020; Robles & Gonzalez-Barahona, 2005), we removed bots and merged multiple developer identities before enumerating `n_disc_starters`, `n_commiters`, and `n_community`. As a result, `n_community` is recorded as 0 if no user has made a commit or started a discussion in the repository, which ignores the creator of the repository. We acknowledge that alternatively such repositories could have the value 1.

For **RQ2**, we operationalised collaboration on models as instances where a pair of developers contributed commits to the same model repository, with direct edges recorded between developers that were weighted by the number of times a developer contributed a commit to the same repository as the other developer (Lopez-Fernandez, 2004). We operationalised commit activity as acts of collaboration because commits are easily measurable, represent "validated" contributions, and represent an accurate audit trail of collaboration (Orucevic-Alagic & Host, 2014; Y. Zhang et al., 2020). However, we acknowledge that the fact that two developers commit to the same repository does not necessarily imply direct interaction; for example, it would have been more accurate to focus on developers' contributions to the same file in a repository, as we discuss in Section 6.5.3. Formally, we modelled collaboration as a network $N = (D, E, W)$, where D is the set of developers, $E = \{(i, j, w_{ij}) \mid i, j \in D, w_{ij} \in \mathbb{N}\}$ is the set of directed edges denoting the relationships between developers, and $W = \{w_{ij} \mid (i, j, w_{ij}) \in E\}$ represents the weights associated with each directed edge. For a developer pair i and j, we denote the directed relationship as (i, j, w_{ij}), where w_{ij} signifies the number of times developer i has committed to the same repository as developer j.

To collect data for the analysis of **RQ2**, we collected commit data from public model repositories via HF Hub API. We started by retrieving a list of all available model IDs using the `list_models()` endpoint. Then, for each model repository, we used the `list_repo_commits()` endpoint to retrieve

[4]N.B. We do not report data for downloads of spaces because spaces cannot be downloaded.

[5]N.B. As per the API, data collection for participation in discussions was limited to users that had started discussions. It was not possible to collect data about users that had made comments in discussion threads.

the commit data, including the authors associated with each commit. For each commit, we recorded an edge between the the developer who made the commit (`source_node`) and all other developers who had contributed to the repository (`target_node`). In cases where a repository had only one contributor, we created self-loop edges to capture the isolate contributor's activity. We did not take temporal dynamics of commit activity into account, which we discuss as a threat to construct validity under Section 6.5.3. We collected data for collaboration in NLP, CV, and MM model repositories by filtering repositories based on the tags, which developers add to their repositories to aid discoverability on HF Hub. We used the list of tags per sub-field provided by HF Hub, including `computer-vision` and `image-classification` for CV models; `translation` and `summarization` for NLP models; and `image-to-text` and `image-to-video` for MM models.

For **RQ3**, we collected data on model usage in spaces using the `list_models()` and `model_info()` API endpoints. We modelled model usage in spaces as a bipartite network, akin to the representation of software dependency networks (Savić, Ivanović, & Jain, 2019). The bipartite model usage network is denoted as $D = (M, S, E)$, where M is the set of models, S is the set of spaces, and $E = \{(m, s) \mid m \in M, s \in S\}$ is the set of undirected edges signifying that "space" s uses model m. The edges are unweighted, representing the model usage relationship between a "space" and a model. From the bipartite network D, we derived an undirected model co-usage network $C = (M, E, W)$. In this network, M is the set of models, $E = \{(m_i, m_j) \mid m_i, m_j \in M\}$ is the set of undirected edges connecting models based on their co-usage in a "space," and $W = \{w_{ij} \mid (m_i, m_j) \in E\}$ is the set of weights assigned to the edges, reflecting the frequency of co-usage of models m_i and m_j across spaces. This analysis complements the former analysis of model usage with insights into the interdependencies and ecosystems surrounding widely used models on HF Hub.

6.3.4 Username Merging

Following prior work, before the analysis, we undertook data preprocessing to merge multiple developer identities per unique developer, which can be caused by how Git records usernames based on users' local repository credentials (Bird et al., 2006; Goeminne & Mens, 2013; Kouters et al., 2012; Robles & Gonzalez-Barahona, 2005; Y. Zhang et al., 2021). We assumed this might be an issue on HF Hub, too. To ensure the accuracy of the dataset of 101,144 developers, we applied a three-pronged approach. First, we classified username string similarity (threshold=90%) between pairs of developers who contributed to the same repository, accepting 126 out of 180 (70.00%) pairs based on manual username searches on HF Hub. Second, in light of the presence of potential real names (i.e. usernames with spaces like "Jessica Smith"), we examined string similarity (threshold=90%) between 1,979 potential real names and the remaining 99,041 usernames, accepting 358 out of 403

(87.75%) username pairs after manual searches on HF Hub. Finally, we inspected the usernames of 700 developers with a network degree of 10 or higher, who represented 0.7% of developers but accounted for 44.78% of edges, via manual searches on HF Hub. This resulted in the identification of 212 username pairs. In total, we merged 546 usernames after removing duplicates.

6.3.5 Data Analysis

To investigate activity on HF Hub (**RQ1**), we conducted a descriptive analysis of various types of activity in 348,181 model repositories, 65,761 dataset repositories, and 156,642 space repositories. Pearson correlation coefficients were calculated to assess the pairwise relationships between the activity variables. In addition, we employed the Mann-Whitney *U* test to compare activity levels across repositories with different licenses. The Mann-Whitney *U* test is a non-parametric test that examines whether two independent samples come from the same distribution, which does not require the data to be normally distributed or to meet the assumption of homogeneity of variance (McKnight & Najab, 2010). Given the large sample sizes, the *U* values are expected to be large, and the salient test statistic is the p-value which indicates the statistical significance of observed differences. Due to capacity constraints in labelling licenses, we limited this analysis to repositories with licenses used in at least 100 repositories (n =339,502, 98% of all repositories). Subsequently, we analysed a snapshot of the social network structure of collaboration on HF Hub (**RQ2**), using techniques defined in Table 6.1 in Appendix 6.7.1. This analysis provides insights into collaboration patterns in model repositories at this point in time. Furthermore, we analysed collaboration patterns in the three AI sub-fields (NLP, CV, and MM) to enable comparisons. Lastly, we examined model adoption on HF Hub (**RQ3**) by calculating the ranked degree of models in the bipartite model usage networks and ranked degree of models in the model co-usage networks to identify the most used models in spaces and their respective developers. These two complementary approaches quantified model popularity (i.e. which models are most frequently used in spaces) and model co-popularity (i.e. which models are most commonly used in conjunction with other models). We replicated this analysis for spaces with NLP, CV, and MM tags for comparative analysis of the three AI sub-fields.

6.4 Results

We first report results for activity in the 348,181 model, 65,761 dataset, and 156,642 space repositories in Section 6.4.1, relying on the metrics described in Section 6.3.3. We then report results on the structure and dynamics of collaboration in Section 6.4.2, based on the analysis of collaboration interactions between around 100,000 developers in model repositories. Finally, we present the re-

sults of our analysis of model adoption in spaces in Section 6.4.3, where we examine the distribution of model usage in spaces on HF Hub and identify the developers of the most used models.

6.4.1 Development Activity on HF Hub

Key findings

We present three key findings: right-skewed distributions across different types of activity (Section 6.4.1.1), strong correlations between development activities (Section 6.4.1.2), and a significant lack of licenses in model and dataset repositories (Section 6.4.1.3).

6.4.1.1 Right-Skewed Distributions in Development Activity

Activity per repository is extremely imbalanced, with right-skewed distributions of `n_likes`, `n_commits`, `n_discussions`, and `n_downloads` across model, dataset, and space repositories (see Figure 6.1). For example, while the maximum number of likes among models is over 9,000, the average model only receives 1.14 likes (see Tables 6.2a-6.2c). The majority of repositories get minimal engagement: 91% of models and 88% of datasets have 0 likes; 84% of models, 91% of datasets, and 96% of spaces have 0 discussions; and 71% of models and 70% of datasets have 0 downloads. Most activity is concentrated in a small number of repositories: $< 1\%$ of models account for 80% of likes, 10% for 80% discussions, 30% for 80% commits, and $< 1\%$ for 80% downloads. Upon increasing the threshold, 8% of models account for 99% likes, 15% for 99% discussions, and 1% for 99% downloads.

Most repositories have a community size of 1; for example, 87% of model repositories have 1 contributor and the 75$^{\text{th}}$ quartile value of `n_committers` is 1 across repository types (see Table 6.2a). The respective maximum values of `n_committers` are 18, 100, and 282 across repository types, and the respective maximum values of `n_community` are 246, 110, and 4,685. The differences between `n_committers` and `n_community` are due to large `n_disc_starters` values, indicating a division of roles in repositories, where many developers participate in discussions but few are involved in model maintenance. The model repositories with the most `n_committers` are `bigscience/bloom` (n=18), `bigcode/santacoder` (n=16), and `deepset/roberta-base-squad2` (n=15).

6.4.1.2 Correlation between Community Size and Engagement

We correlate frequency counts over the different types of activity described in Section 6.3.3 (see Figure 6.2). In model repositories, we find a strong positive correlations between `n_community` and `n_likes` ($\rho = 0.75$, $p < 0.001$). In space repositories, we find strong correlations between various activities, especially `n_likes`) and `n_discussions` ($\rho = 0.74$, $p < 0.001$), `n_disc_starters`

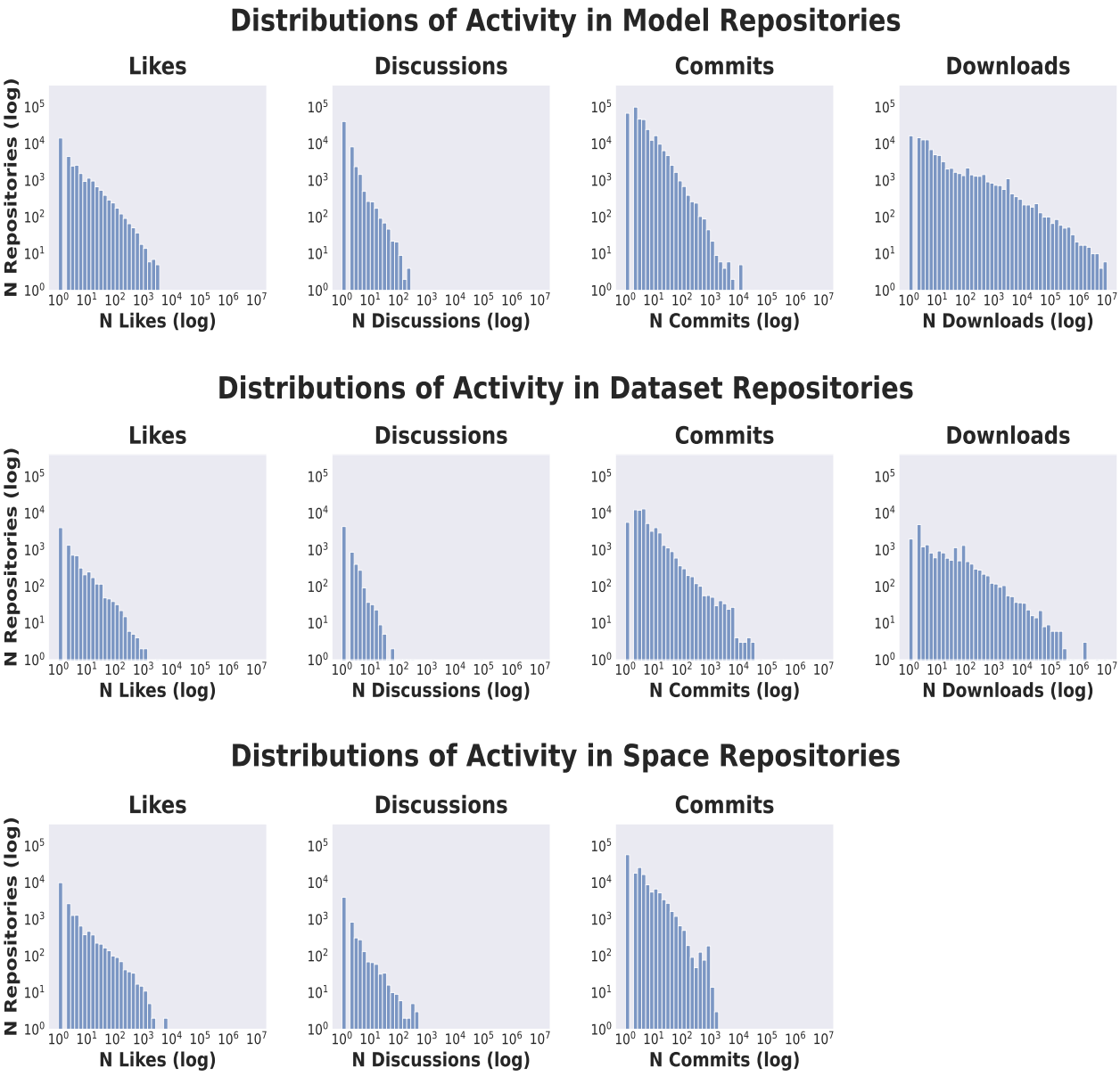


Figure 6.1: Distributions of Development Activity in HF Hub Repositories

($\rho = 0.76$, $p < 0.001$), and `n_community` ($\rho = 0.76$, $p < 0.001$). However, in general, we observe weak correlations between most activities in model and dataset repositories. Furthermore, we do not find a strong correlation between commit activity (`n_commits`) and other types of activity, indicating that commit activity is not strongly linked to community engagement.

6.4.1.3 Impact of Licenses on Collaboration

A significant proportion of model and dataset repositories lack licenses, which can create uncertainty and potential legal issues for users and developers. Specifying a license is not the norm: the majority of model repositories (65%) and datasets (72%) do not have a license. Among the licensed models, the most commonly used licenses are Apache v2.0 (37%), MIT (17%), OpenRAIL (14%), and CreativeML OpenRAIL-M (10%). The most used licenses for datasets are MIT (28%), Apache v2.0 (15%), OpenRAIL (9%), and licenses from the family of Creative Commons v4.0 (7%).

The choice of license matters: there is a moderate to strong correlation between the use of a license and level of activity in model repositories (see Figure 6.3). Furthermore, the Mann-Whitney *U* tests provide strong evidence of statistically significant differences between collaboration dynamics in model repositories with different types of licenses (all tests have $p < 0.001$). Specifically, model repositories with permissive licenses consistently have the highest levels of activity compared to model repositories with no license and those with restrictive licenses (see Table 6.3). However, repositories with restrictive licenses also exhibit significantly higher activity than those with no license. This pattern holds across all activity metrics measured, suggesting that while permissive licenses foster the highest engagement, restrictive licenses also promote more collaboration compared to model repositories that do not have a license. However, we acknowledge that other factors beyond license use and license type may affect collaboration patterns (see Section 6.5.3).

6.4.2 Social Network Structure and Dynamics of Collaboration

Key findings

We present three key findings: a highly uneven distribution of collaboration, with 89% of developers working in isolation (Section 6.4.2.1); a core-periphery network structure characterised by high reciprocity but low elitism among the most prolific developers (Section 6.4.2.1); and similar collaboration patterns across AI sub-fields (Section 6.4.2.2).

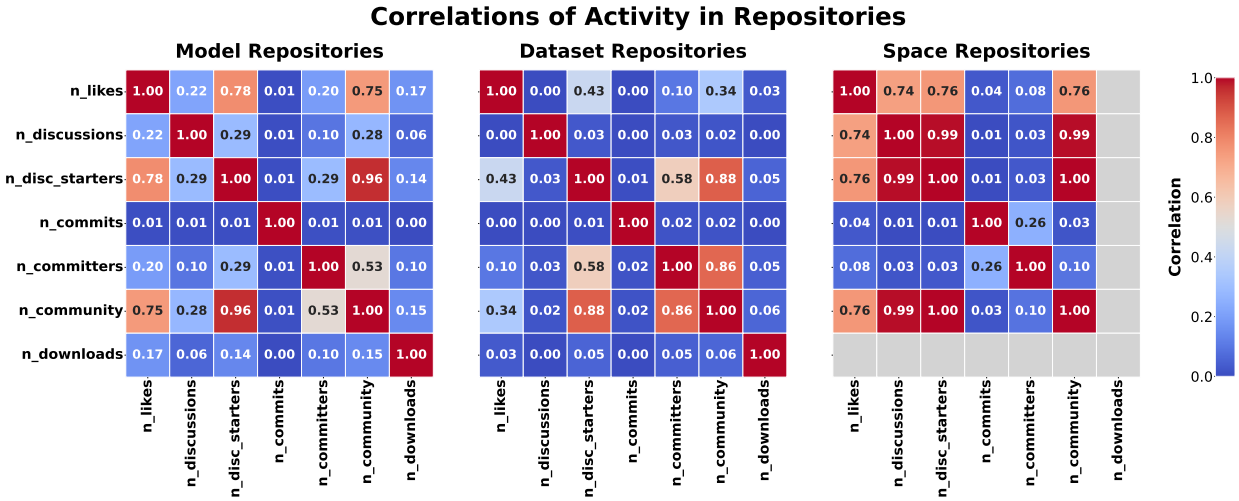


Figure 6.2: Correlations of Activity in Model, Dataset, and Space Repositories

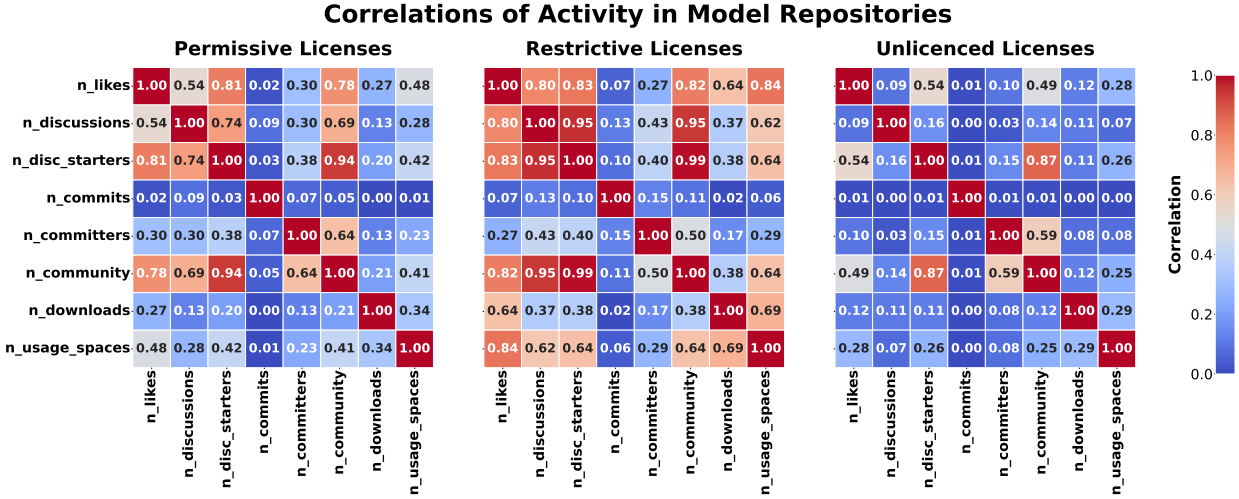


Figure 6.3: Correlations of Activity in Model Repositories with Different License Types

6.4.2.1 Collaboration in Model Repositories on HF Hub

The majority of developers (89%) have not collaborated with others. Excluding these isolate developers, the remaining 10,524 developers have an average degree of 4.10 (SD: 32.63) and node degrees range from 1 to 3,140. The HF Hub collaboration network exhibits right-skewed distributions of degree and PageRank centrality (see Figure 6.4) suggest that a small group of influential developers plays a central role in driving collaboration on OMs on HF Hub. Specifically, the degree centrality distribution has a mean of 4 and a median of 2, with a maximum of 3,140 and a standard deviation of 33, while the PageRank centrality distribution has a mean and median of 0.0001, a maximum of 0.04, and a standard deviation of 0.0005.

HF Hub developer community exhibits a core-periphery structure, with a tightly interconnected core of prolific developers. The k-core decomposition analysis reveals that as the k-core value increases, the number of distinct communities decreases, ultimately converging into a single densely interconnected core at k=26 (see Table 6.4a). The high modularity (0.81) at k=1 suggests that the whole network consists of loosely connected groups of developers. As the k-core value increases, the modularity decreases to 0.00 at k=26, indicating a transition from a compartmentalised community structure with distinct clusters to an integrated core characterised by high cohesion and a lack of discernible sub-groups (i.e. cliques of developers).

Collaboration is characterised by high reciprocity values, ranging from 0.81 to 1.00 across all k-core levels (see Table 6.4a), indicating the prevalence of mutual relationships among developers. The low assortativity values, ranging from -0.49 to 0.08, suggest that developers collaborate regardless of their network centrality, implying that other factors, such as shared interests, skills, or project roles, may be more significant in driving collaboration than their network centrality. Furthermore, the low average rich club coefficients, ranging from 0.04 to 0.41, indicate that highly central developers do not primarily collaborate with each other and a lack of elitism among power developers.

6.4.2.2 Collaboration in Model Repositories in AI Sub-Fields

Collaborations on models in the NLP, CV, and MM sub-fields, despite the different sizes of the respective communities, are similarly characterised by core-periphery structures ith high modularity and low density (see Tables 6.4b,6.4c, and 6.4d). At k=1, all networks are highly modular (CV: 0.80, NLP: 0.82, MM: 0.71) and have very low density (CV: 0.01, NLP: 0.00, MM: 0.00), implying that collaborations in the respective AI sub-fields are clustered into distinct communities of collaborators. As the k threshold increases, the networks undergo a similar transformation process, with modularity decreasing to 0.00 and the number of communities reducing to a single cohesive community at

the maximal k values (CV: 10, NLP: 25, MM: 26). Concurrently, density increases, reaching 1.00 for CV and MM and 0.97 for NLP at their respective maximal k values.

Collaboration in sub-fields is also characterised by reciprocity and connectivity in the core. At k=1, reciprocity values range from 0.84 to 0.93 and increase to 1.00 at the maximal k for CV and MM, while NLP maintains has a reciprocity value of 0.98 at its maximal k. The average degree of all networks increase with increases in k. This suggests that as we move towards the core of the collaboration networks, developers become more interconnected and collaborate with a larger number of peers. However, the low average clustering coefficients and low average rich club coefficients across all networks indicate that the more prolific developers in the respective sub-fields tend to collaborate with a diverse set of individuals rather than forming tightly-knit groups.

6.4.3 Model Adoption in Spaces on HF Hub

Key findings

We present two key findings: model adoption in spaces is characterised by a right-skewed distribution (Section 6.4.3.1), and the most used models in HF Hub spaces are built by a small number of developers (in particular, Big Tech companies) (Section 6.4.3.2).

6.4.3.1 Right-Skewed Distribution of Model Adoption

The bipartite model usage network displays a disparity in model adoption in spaces. The degree distribution of the bipartite network is right-skewed, as shown in Figure 6.5. Only three models are used in 1,000 or more spaces, including `runwayml/stable-diffusion-v1-5` (n=1747), `skytnt/anime-seg` (n=1162), and `gpt2` (n=1002). The mean degree (6.68) is significantly higher than the median (1.00), and the large standard deviation (34.75) confirms the high variability in model usage. The majority of models have a low degree of usage, with at least 50% being used in only one space, while a small number of highly popular models dominate the usage, with the maximum degree reaching 1,747. This suggests that a few key models are widely adopted in AI applications, while many other models have limited use cases. The model co-usage network provides an additional perspective on the uneven interdependencies of models in spaces, complementing insights gained from examining model downloads or individual model usage in spaces. Specifically, the degree distribution of this network exhibits a multi-modal pattern, with five distinct clusters, each exhibiting a right-skewed shape (see Figure 6.5). A small cluster at the far-right tail of the distribution represents a few highly interconnected models with significantly higher co-usage degrees compared to the other clusters.

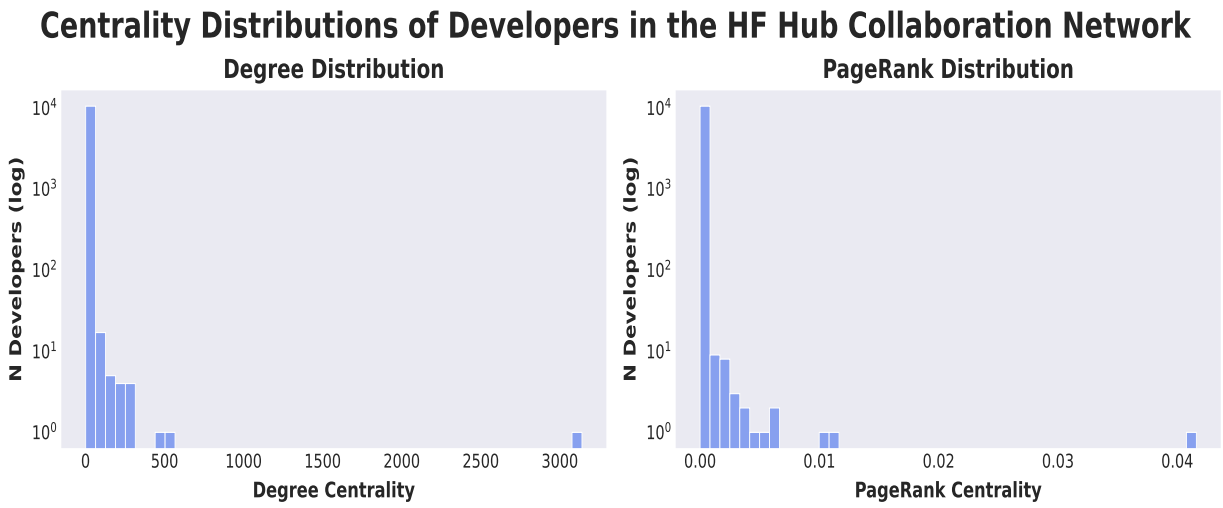


Figure 6.4: PageRank and Degree Distributions of Developers on HF Hub

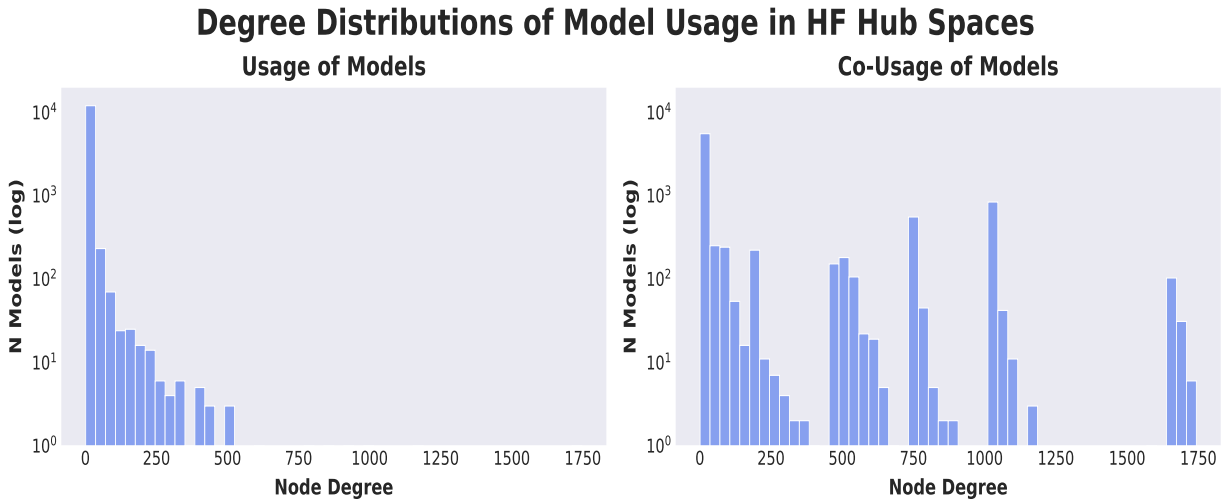


Figure 6.5: Degree Distribution of Model Adoption in Spaces on HF Hub

6.4.3.2 Dominance of a Few Models by a Few Developers

When we rank the models by their usage in spaces, we observe that major organisations, rather than individual developers or grassroots initiatives, have developed the most used models. among the 100 most used models in spaces, the following organisations have developed the most models: Meta (n=8), Google (n=7), StabilityAI (n=5), OpenAI (n=4), Microsoft (n=4), and Fudan University (n=4). These five organisations account for 33% of the 100 most used models in spaces. We note that the individual user nitrosocke (n=5), an employee at StabilityAI, ranked highly among these organisations. With regards to the model co-usage network, the key developers of the 100 most co-used models in all spaces are: EleutherAI (n=15), Meta (n=12), h20ai (n=11), BigScience (n=9), and lmsys (n=9). These five organisations account for 56% of the 100 most co-used models in spaces.

The model usage networks in the sub-fields similarly exhibit right-skewed degree distributions, highlighting the dominance of a minority of models in each sub-field. The most used models in spaces with NLP tags (n=3,995) are `gpt2` (n=1,001), `bertbaseuncased` (n=621), and `gpt2medium` (n=445). The organisations that developed the most models among the 100 most used models are Google (n=9), Meta (n=5), and Fudan University (n=5). For comparison, in the NLP model co-usage network, EleutherAI ranks first (n=16), followed by h20ai (n=12) and Meta (n=11). The most used models in spaces with CV tags (n=416) are `saltacc/anime-ai-detect` (n=500), `openai/clip-vit-large-patch14` (n=454), and `openai/clip-vit-base-patch32` (n=277). The most prolific developer of models in spaces with CV tags is the user lllyasviel (n=20), followed by Meta (n=8) and the user DucHaiten (n=7). For comparison, in the CV model co-usage network, LAION AI ranks as the developer of the most models among the top 100 (n=17). Finally, the most used models in spaces with MM tags (n=2,394) are `runwayml/stable-diffusion-v1-5` (n=1748), `CompVis/stable-diffusion-v1-4` (n=925), and `stabilityai/stable-diffusion-2-1` (n=854). among the developers of the 100 most used models, Stability AI ranks first, with 15 of the 100 most used models and 22 of the top-ranked co-used models in such spaces. These findings highlight the key models and players in the NLP, CV, and MM communities.

6.4.3.3 Correlations between Model Likes and Model Usage

We observe a strong positive correlation between `n_likes` and `n_usage_spaces` ($\rho = 0.66$, $p < 0.001$), and a weak positive correlation between `n_downloads` and `n_usage_spaces` ($\rho = 0.29, p < .001$). These findings suggest that the number of likes is more strongly associated with the usage of models in spaces compared to the number of downloads, and that likes in model repositories are a good indicator of its adoption in applications on HF Hub. However, as mentioned in Section 6.3.3,

we note that download counts are limited and therefore may only provide a snapshot of correlations between downloads and likes or usage, which may not generalise in all time periods.

6.5 Discussion

In this section, we discuss the implications of our findings for research and practice. We highlight the study's contributions to the literature in Section 6.5.1.1. We reflect on the methodological considerations of using HF Hub as a data source for research on OSAI in Section 6.5.1.2. We make five recommendations for future research to advance the research agenda on OSAI in Section 6.5.1.3. Finally, we discuss the implications for practice and make recommendations in Section 6.5.2.

6.5.1 Implications for Research

6.5.1.1 Contributions to Academic Literature

Uneven influence in HF Hub developer community: We extend prior findings of right-skewed distributions of commit activity in model repositories (Castaño et al., 2024) with observations of right-skewed distributions of various development activities on HF Hub, including interactions in model, dataset, and space repositories; code collaborations between developers; and model usage in spaces. Activity distributions follow power law patterns, with a small fraction of repositories accounting for most interactions (e.g. $< 1\%$ for 80% of likes, 10% for 80% discussions, 30% for 80% commits, <1% for 80% downloads). Similarly, the collaboration networks exhibit right-skewed centrality distributions, indicating that influence is concentrated among few developers, congruent with prior observations that OSS development patterns generally follow Pareto distributions (Goeminne & Mens, 2011; Mockus et al., 2002; Szymański & Ochodek, 2023; J. Xu et al., 2006; Y. Zhang et al., 2021). Influence also flows across HF Hub, with likes per model having strong correlations with their usage in spaces ($\rho = 0.66$, $p < 0.001$).

Impact of license on collaboration: The Mann-Whitney *U* tests show that license choice significantly impacts the level of activity and engagement in repositories, with permissive licenses exhibiting the highest activity levels, followed by repositories with restrictive licenses, and finally ones with no license. Furthermore, the Pearson correlations indicate that the use of a license (permissive or restrictive) is associated with stronger correlations between various types of activity compared to repositories without licenses. A possible explanation for this difference is the legal liability that a license provides for the use of a model compared to an unlicensed model. Furthermore, the license terms can affect development behaviour; for example, a we may expect models with a permissive license to receive greater usage by or contributions from companies which adopt the models for

commercial purposes. Overall, these findings highlight the important role of licensing decisions in influencing the collaborative dynamics in OM development as well as open new avenues for future research thereon.

Core-periphery structure of the HF developer community: To the best of our knowledge, only one prior study has investigated model development practices in the HF developer community, showing that most models only have one contributor and that model maintenance chiefly involves "perfective tasks" to enhance model performance (Castaño et al., 2024). We extend this finding with three insights. First, we corroborate the findings that most developers (89%) are islands, who have not collaborated with other developers in model repositories on HF Hub. This is not unique to HF Hub: the majority of OSS projects are developed by individuals (Krishnamurthy, 2005b). However, what may be specific about the small community sizes in model development is the nature of the model development life-cycle ("code once, train often"). Second, the social network structure of collaboration in model repositories is characterised by a core-periphery structure, with a dense core of highly active developers, akin to the "layered onion" structure common in OSS (Crowston et al., 2005). Third, collaborations have high reciprocity and low assortativity, signifying the prevalence of mutual relationships among developers, regardless of their social positions in the community.

Uneven model adoption in spaces: By examining model adoption in spaces, we empirically tested prior claims about the disproportionate influence of industry-leading companies in the OSAI ecosystem (Widder et al., 2023). We identified the popularity of a relatively small number of models used in spaces as well as the influential role of a few organisations, including Meta, Google, Stability AI, OpenAI, Microsoft, and EleutherAI, who have developed the most widely used models. Some critics of the open-source model of AI development fear that too many unknown actors will introduce distributed safety issues, while advocates of the development model tout democratisation of power as a core benefit. Our findings show that a few organisations possess majority influence in this ecosystem, which challenges both of these narratives. In many cases, the most influential actors in the OSAI ecosystem are one in the same as those in closed-source AI (Widder et al., 2023).

6.5.1.2 HF Hub: A New Source of Research Data about OSAI Development

This paper contributes to the research effort to use HF Hub as a data source for empirical studies on OM development (Ait et al., 2023b, 2023a; Castaño et al., 2024). We share two reflections on methodological considerations. First, informed by prior work that underlines the importance of merging usernames for unique developers, we anticipated that this might be an issue on HF Hub (Bird et al., 2006; Goeminne & Mens, 2013; Kouters et al., 2012; Robles & Gonzalez-Barahona, 2005; Y. Zhang et al., 2021). While our three-pronged approach strikes a balance between the

impracticality of manually inspecting over 100,000 developers versus the risk of misclassification through a fully automated approach, it is still imperfect. Future research may consider more sophisticated approaches to this problem. Second, the API is not optimised for research purposes, which makes data collection time-consuming (e.g. one must make a unique API call to retrieve commit histories of each model and handle rate limits) and limited (e.g. user metadata is not available). The lack of user metadata hinders the ability to study the characteristics and behaviours of individual developers, such as their expertise and affiliations, as well as automated approaches to username merging that incorporate user metadata. To overcome these limitations, researchers may explore alternative approaches and tools, such as the HFCOMMUNITY database developed by Ait et al. to facilitate empirical studies of activity on the platform (Ait et al., 2023a).

6.5.1.3 Recommendations for Future Research

We recommend five research directions that can advance the OSAI research agenda.

1. Implications of concentrations in HF Hub developer community: We confirm prior observations that the models of a handful of companies are dominant among HF Hub developer community (Widder et al., 2023). We encourage future research to investigate what these concentrations mean in practice, such as the potential benefits that these companies accrue from their ecosystems, including increased visibility, crowdsourced contributions (e.g. via commits and discussions), and access to diverse fine-tuned versions shared by other developers on HF Hub. Furthermore, there is a concern that dominant companies benefit from developers being locked-in to their ecosystems, potentially limiting competition and entrenching their dominance. Future research could investigate the factors contributing to such concentrations, such as the reputation of the companies developing the models, their access to resources, or the performance and versatility of their models, as well as the implications of these concentrations for the broader AI community, including the impact on research, innovation, and the distribution of benefits and resources.

2. Incentives and modes of participation: Future research could investigate the incentives of individual developers and companies. A number of companies have released OMs on HF Hub, such as Meta's Llama models (HuggingFace, 2024f), Mistral AI's Mixtral models (HuggingFace, 2024g), and OpenAI's Whisper models (HuggingFace, 2024i). Often these releases are presented as acts of AI democratisation (Seger, Ovadya, et al., 2023). Future research could critically examine the commercial incentives behind these releases. In addition, future research could examine commercial approaches to model governance and maintenance—e.g. if and how companies welcome or engage with community contributions—as well as if and how companies collaborate with each other on OM development, as they do in OSS development (Nguyen-Duc et al., 2019; Germonprez et al., 2013;

Linåker et al., 2016; Teixeira & Lin, 2014; Y. Zhang et al., 2020).

3. Collaboration dynamics in active repository communities: We know that model maintenance focuses on model performance improvements (Castaño et al., 2024); and in the minority of repositories that have active communities, most developers contribute to discussions rather than commits (see Tables 6.2a-6.2c). Going further, we encourage researchers to examine collaboration dynamics in repositories with active communities from multiple angles. Given the sizeable differences in `n_committers` and `n_disc_starters`, future research could investigate the division of roles between discussion and code contributors, typical topics of discussion (e.g. model performance, new ideas, etc.), how discussions inform model maintenance if at all, and the journeys of developers from discussion contributors to committers, among others. In addition, future research could examine the governance approaches (e.g. contribution policies) that repository owners use to encourage collaboration. Future analyses could also take into account temporal dynamics, providing insights into evolving patterns, social structures, and trends of the developer community on HF Hub.

4. Impact of model size on collaboration Future research should examine the impact of model size (i.e. parameters) on the nature of collaboration in repositories on HF Hub. For instance, it could examine how resource constraints (e.g. computational power or data availability) influence collaboration for various stakeholders (e.g. individual developers or developers from industry labs) on models of different sizes. By shedding light on facilitators and barriers for collaboration on OMs, such research could guide efforts to foster inclusive and diverse communities.

5. Collaboration beyond HF Hub: While this analysis provides insights into the developer community on HF Hub, we have a limited understanding of the development of the various components involved in the development of models (White et al., 2024), which largely takes place in proprietary settings or on other platforms like GitHub (Ding et al., 2023). We encourage future research to examine how HF Hub is used in the wider ecosystem of platforms and offline venues for the collaborative development of OMs and datasets. This research direction would enable comparisons of the collaboration patterns among model developers and model fine-tuners. In addition, researchers could undertake a multi-sited analysis, examining collaboration on the same project across platforms.

6.5.2 Implications for Practice

6.5.2.1 Recommendations for Open Source Practitioners

Beyond our research suggestions, we encourage open source researchers and practitioners to develop standardised metrics for studying OM development. Groups like the Linux Foundation's Community Health Analytics in Open Source Software project (CHAOSS, 2024), which has created metrics to assess the health and sustainability of OSS developer communities, are well-positioned to lead this

effort. The lack of empirical data on OM development hinders evidence-based decision-making in this rapidly evolving field, and by working together to establish appropriate metrics, open source practitioners can help to address the data gap in OSAI research.

6.5.2.2 Recommendations for Open Model Hosting Platforms

We make two recommendations to OM hosting platforms like HF Hub. First, HF could work with researchers to identify features and API improvements that would aid research efforts concerning OM development on its platform, building on efforts by members of the HF community, such as `Weyaxi/huggingfaceleaderboard`. This collaboration could include collecting and publishing data on OM development patterns and collaboration, which would help fill the current "data gap" in this area. HF may take inspiration from GitHub's Innovation Graph (GitHub, 2024b) or its annual Octoverse reports (Daigle, 2023), which provide access to data and insights on activity on its platform.

Second, a large proportion of models (64.67%) and datasets (72.13%) lack licenses, which may be due to uncertainty about how or whether they should be licensed (Hardy, 2023; OSI, 2023). For comparison, the number of unlicensed repositories on GitHub is lower at 46% (Weaver, 2020). In the interest of promoting responsible development, HF should consider developing educational resources on licenses, such as guides or tutorial videos, or developing features, such as a license drop-down menu, which can inform developers of the options available as well as their merits and drawbacks. Such a feature could be considered in addition to previous recommendations for moderating models that are hosted on HF Hub, such as hiring AI safety researchers and proactively red-teaming unsafe models (Tsamados et al., 2023; Gorwa & Veale, 2024).

6.5.2.3 Recommendations for Policymakers

As the use and impact of OMs increases over time, policymakers need empirical data to inform discussions about the benefits, risks, and governance of OMs. Our analysis provides one empirical lens on the extent of model proliferation and adoption, which can help ground policy decisions. For example, it is illuminating to observe that most models (70.99%) have not been downloaded once or that 1% of models account for 99% of downloads. This is a reminder that the availability of a model does not mean it will be (widely) used. Furthermore, while download counts were limited to the past 30 days, the fact that only 86 models had over one million downloads indicates that the number of widely used models is not excessively large and governable. What is more, we observe that models developed by a number of non-profit, grassroots initiatives like EleutherAI and BigScience, gained traction. Following the charge of the French government to fund the OSAI commons (Chatterjee & Volpicelli, 2023), policymakers may use such data to identify non-commercial projects that could

be supported. Overall, the data points reported in Section 6.4.1 could help policymakers assess the real-world impact of OMs and develop appropriate governance frameworks to maximise their benefits while mitigating potential risks of OSAI.

6.5.3 Threats to Validity

We evaluate the validity of our findings by following guidance for empirical software engineering research (Easterbrook et al., 2008; Runeson & Höst, 2008).

6.5.3.1 Construct validity

Construct validity concerns the extent to which a measurement accurately assesses the theoretical construct it intends to measure. Our study aimed to measure typical patterns of development activity on HF Hub, but we acknowledge several threats to construct validity. First, our analysis is limited to activity in public repositories and does not account for collaboration in private repositories. Second, download counts have a few limitations: they are limited to the past 30 days, download counts may be incorrectly reported (e.g. if the repository lacks a configuration file or if the model is used on-device versus in continuous integration), and dataset downloads are limited to the count of `load_dataset()` calls (HuggingFace, 2024h, 2024b). Third, our operationalisation of collaboration relies on commits to model repositories, assuming that the co-occurrence of commits indicates collaboration. However, this assumption may not always hold true, especially in large repositories where developers may work on independent tasks. Future research could operationalise collaboration on specific files and quantify the relative contribution of developers to specific files (Orucevic-Alagic & Host, 2014). Furthermore, this analysis is limited to snapshot of HF Hub developer community in October 2023, which does not capture the dynamics of collaboration and activity over time, which should be considered in future research, as discussed in Section 6.5.1.3.

6.5.3.2 Internal validity

Internal validity concerns the extent to which a study can confidently attribute the observed results to the investigated variables, minimising the influence of confounding factors or alternative explanations. As explained in section 6.3.4, there may be a slight inaccuracy in the enumeration of community size per repository and the number of developers included in the collaboration networks due to discrepancies in username data, such as multiple accounts or usernames per developer. This is a common problem in OSS research, and there is no perfect solution to username merging (Bird et al., 2006; Kouters et al., 2012; Robles & Gonzalez-Barahona, 2005). API limitations prevent the use of methods that incorporate user metadata for username merging (Amreen, Mockus, Zaretzki,

Bogart, & Zhang, 2020; Y. Zhang et al., 2021). For example, we rejected 34 username pairs due to insufficient evidence to confirm the match with confidence.

6.5.3.3 External validity

External validity concerns the generalisability of the findings. While HF Hub has gained significant popularity, it is important to acknowledge that there may be other platforms where OM development takes place and that our findings may not generalise to those platforms. Future research could explore collaboration practices across different platforms to provide a more comprehensive view of the OSAI ecosystem. That being said, we observe that development activity on HF Hub is characterised by the Pareto principle, conforming with OSS development patterns on platforms like GitHub (e.g. Goeminne & Mens, 2011; Szymański & Ochodek, 2023; Y. Zhang et al., 2021). Another threat to the external validity of the findings concerns the analysis of model usage. While there were as many as 156,642 spaces at the time of data collection, they do not represent the use of OMs beyond HF Hub platform, thus limiting the generalisability of our claims, with the exception of finding a strong positive correlation between likes of model repositories and their usage in spaces ($\rho = 0.66$, $p < 0.001$). Future research could address this limitation by exploring other sources of data on model adoption, such as academic publications, industry reports, or user surveys, to triangulate the findings.

6.5.3.4 Reliability

Reliability refers to the consistency and reproducibility of the study's results. To enhance the reliability of our study, we have uploaded the Python scripts used for data collection and processing to a public GitHub repository (Osborne, 2024b). Due to privacy and ethical considerations, we do not share the raw data (see Data Availability statement).

6.6 Conclusion

The proliferation of OMs has become a focal point of the political economy of AI, yet their development has received little study to date. In this study, we produced empirical insights on development practices in the burgeoning developer community on HF Hub. First, we found that various types of development activity, from likes and downloads to discussions and commits, across 348,181 model, 65,761 dataset, and 156,642 space repositories exhibit right-skewed distributions. In addition, activity and engagement is highly imbalanced between repositories; for example, over 70% of models have 0 downloads and 1% account for 99% of downloads. Second, we found that the developer community has a core-periphery structure, with a core of highly prolific developers and a majority

of isolate developers (89%) who do not collaborate with others. However, collaboration is characterised by high reciprocity and low levels of assortativity regardless of developers' social positions in the HF developer community. Third, we found that a minority of models are widely used in HF spaces, which have been developed by a handful of industry leaders, which signifies the concentrated influence of a handful of actors in HF Hub ecosystem. We concluded with a discussion of the implications of our findings and recommendations for researchers and practitioners, with the hope that the OM development and collaboration can be more deeply investigated in the future.

6.7 Appendix for Chapter 6 (RP3)

6.7.1 Definitions of Network Properties in Social Network Analysis

Table 6.1: Definition of Network Properties

Property	Definition
Nodes (N)	Number of nodes in the network (e.g. developers or models).
Edges (E)	Number of edges in the network (e.g. links between developers).
Degree centrality	Degree centrality is a measure of the importance of a node in a network based on the number of connections it has. It is calculated as the number of edges a node has with other nodes.
PageRank centrality	PageRank is a network centrality measure that assesses a node's importance based on the quantity and quality of incoming links, considering the recursive influence of nodes pointing to it (Page, Brin, Motwani, & Winograd, 1999). PageRank values range from 0 to 1, with higher values indicating greater global influence and importance in the network.
k-core	*k*-core decomposition identifies the maximal subgraph in which every node is connected to at least *k* other nodes, helping to reveal the network's core structure (Batagelj & Zaversnik, 2003).
Modularity (MOD)	Modularity measures the extent to which a network can be divided into distinct and densely interconnected communities. The modularity value ranges from -1 to 1, with positive values indicating a high degree of community structure and negative values implying the absence of community structure (M. E. J. Newman, 2006). We used the Clauset-Newman-Moore greedy modularity maximization algorithm to calculate the modularity of networks (Clauset, Newman, & Moore, 2004).
Communities (COM)	Community detection identifies cohesive groups in a network by optimizing a measure of modular structure. The goal is to find a partition of the network that maximizes the density of connections in communities while minimizing the connections between them. We used the Clauset-Newman-Moore greedy modularity maximization algorithm (Clauset et al., 2004) to find the community partition with the highest modularity value and to determine the number of communities in the network (COM).
Density (DENS)	Density is a measure of how connected a network is, calculated as the ratio of the number of edges present in the network to the maximum possible number of edges in the network (NetworkX, 2023b).
Reciprocity (RECIP)	The reciprocity of a directed graph is the ratio of the number of edges pointing in both directions to the total number of edges in the graph.(NetworkX, 2023e). A value of 1 means that all edges are bidirectional, while a value of 0 means there are no mutual connections.
Average Rich Club Coefficient (ARCC)	The average rich club coefficient measures the extent to which high-degree nodes tend to be more connected than expected by chance (S. Zhou & Mondragon, 2004). Detecting the rich-club phenomenon reveals high-level semantic insights about the network (McAuley, Costa, & Caetano, 2007), a key property of power-law networks (S. Zhou & Mondragon, 2004). The coefficient ranges from 0 to 1, where 0 indicates no preferential connection and 1 indicates a fully connected high-degree subgraph. We implemented this calculation on directed networks (Smilkov & Kocarev, 2010). For each network, we set the threshold *k* as the degree value corresponding to the minimum degree among the top 10% of highest-degree nodes.
Assortativity (ASS)	Assortativity measures the tendency of nodes to be connected to nodes with similar degrees, with values ranging from [-1,1] (M. E. J. Newman, 2002).
Average Degree (AD)	The average number of edges per node.
Average Clustering Coefficient (ACC)	The average clustering coefficient [0, 1] measures the extent to which nodes in a network tend to cluster together. It quantifies the level of triadic closure in the network, which is the tendency for a node's neighbours to be connected to each other (Saramäki, Kivelä, Onnela, Kaski, & Kertész, 2007).

6.7.2 Summary Statistics of Development Activity in Repositories

Table 6.2: Summary Statistics of Development Activity on HF Hub

(a) Development Activity in Model Repositories

	n_likes	n_discussions	n_disc_starters	n_commits	n_committers	n_community	n_downloads
mean	1.14	0.28	0.10	7.28	1.06	1.13	1,693.93
std	30.56	5.42	1.17	167.14	0.41	1.26	158,207.09
min	0	0	0	0	0	0	0
25%	0	0	0	2	1	1	0
50%	0	0	0	3	1	1	0
75%	0	0	0	5	1	1	0
max	9,314	3,006	240	75,653	18	246	65,729,394

N.B. `n_community` *equals 0 if no user has made a commit or started a discussion in the repository, except for its creator.*

(b) Development Activity in Dataset Repositories

	n_likes	n_discussions	n_disc_starters	n_commits	n_committers	n_community	n_downloads
mean	0.92	1.41	0.12	30.04	1.13	1.20	476.70
std	18.14	309.14	0.61	1298.11	0.67	0.87	43,786.43
min	0	0	0	0	0	0	0
25%	0	0	0	2	1	1	0
50%	0	0	0	4	1	1	0
75%	0	0	0	7	1	1	2
max	3,493	79,256	64	314,813	100	110	9,651,261

N.B. `n_community` *equals 0 if no user has made a commit or started a discussion in the repository, except for its creator.*

(c) Development Activity in Space Repositories

	n_likes	n_discussions	n_disc_starters	n_commits	n_committers	n_community	n_downloads
mean	1.33	0.34	0.15	8.22	1.27	1.40	N/A
std	37.35	54.58	15.36	40.36	0.99	15.39	N/A
min	0	0	0	0	0	0	N/A
25%	0	0	0	1	1	1	N/A
50%	0	0	0	3	1	1	N/A
75%	0	0	0	5	1	2	N/A
max	9,124	18,061	4,684	2,150	282	4,685	N/A

N.B. `n_community` *equals 0 if no user has made a commit or started a discussion in the repository, except for its creator.*

6.7.3 Mann-Whitney U Tests for Activity in Model Repositories

Table 6.3: Mann-Whitney *U* Test Results

Activity	**Permissive vs. Restrictive**	**Permissive vs. No license**	**Restrictive vs. No license**
`n_likes`	1,464,356,547 ($p < 0.001$)	8,712,552,137 ($p < 0.001$)	5,558,887,829 ($p < 0.001$)
`n_disc_starters`	1,501,884,124 ($p < 0.001$)	8,323,015,321 ($p < 0.001$)	5,200,616,824 ($p < 0.001$)
`n_discussions`	1,727,651,315 ($p < 0.001$)	9,180,111,459 ($p < 0.001$)	4,983,923,782 ($p < 0.001$)
`n_committers`	1,573,347,885 ($p < 0.001$)	8,744,559,271 ($p < 0.001$)	5,238,059,429 ($p < 0.001$)
`n_commits`	2,013,736,368 ($p < 0.001$)	11,333,305,055 ($p < 0.001$)	5,451,995,655 ($p < 0.001$)
`n_downloads`	1,884,325,690 ($p < 0.001$)	10,136,574,180 ($p < 0.001$)	4,983,179,068 ($p < 0.001$)
`n_community`	1,563,484,573 ($p < 0.001$)	8,856,297,978 ($p < 0.001$)	5,333,406,861 ($p < 0.001$)
`n_usage_spaces`	1,509,785,617 ($p < 0.001$)	8,276,280,946 ($p < 0.001$)	5,148,816,091 ($p < 0.001$)

6.7.4 Social Network Structure of Collaboration on HF Hub

Table 6.4: Network Structure of Collaboration in Model Repositories on HF Hub

(a) All Model Repositories

k-core	N	E	MOD	COM	DENS	RECIP	ARCC	ASS	AD	ACC
1	10,524	21,598	0.81	2,894	0.00	0.83	0.04	0.08	4.10	0.00
5	1,109	7,849	0.60	68	0.01	0.81	0.08	0.01	14.16	0.00
10	330	3,769	0.52	18	0.03	0.85	0.11	0.07	22.84	0.00
26 (max)	14	182	0.00	1	1.00	1.00	0.15	NaN	26.00	0.04

(b) Computer Vision Model Repositories

k-core	N	E	MOD	COM	DENS	RECIP	ARCC	ASS	AD	ACC
1	371	697	0.80	112	0.01	0.90	0.26	0.26	3.76	0.01
5	41	203	0.52	4	0.12	0.93	0.67	-0.00	9.90	0.04
10 (max)	6	30	0.00	1	1.00	1.00	0.40	NaN	10.00	0.16

(c) Natural Language Processing Model Repositories

k-core	N	E	MOD	COM	DENS	RECIP	ARCC	ASS	AD	ACC
1	4,606	10,010	0.82	1,366	0.00	0.93	0.07	0.33	4.35	0.00
5	565	3,980	0.62	41	0.01	0.91	0.11	0.14	14.09	0.00
10	182	1,980	0.52	13	0.06	0.90	0.15	0.07	21.76	0.00
25 (max)	15	204	0.00	1	0.97	0.98	0.25	NaN	27.20	0.01

(d) Multimodal Model Repositories

k-core	N	E	MOD	COM	DENS	RECIP	ARCC	ASS	AD	ACC
1	1,546	3,661	0.71	321	0.00	0.84	0.10	0.19	4.74	0.00
5	218	1,480	0.57	18	0.03	0.89	0.16	-0.00	13.58	0.00
10	82	805	0.50	6	0.12	0.91	0.37	-0.33	19.63	0.02
26 (max)	14	182	0.00	1	1.00	1.00	0.15	NaN	26.00	0.06

Co-authorship Statement

I lead-authored this paper alongside two co-authors. As the first author, I carried out the majority of tasks, namely: research conception and design; literature review; data collection, processing, and analysis; paper writing; feedback collection; and submission. Jennifer Ding provided feedback on three drafts of the manuscript and contributed to the discussion section, and Hannah Kirk provided detailed feedback on three drafts of the manuscript.

For their feedback on previous versions of this manuscript, I would also like to thank Loubna Ben Allal, Daniel van Strien, Peter Cihon, Mer Joyce, Stefano Maffulli, Matt White, Seb Elmes, Alek Tarkowski, Johan Linåker, Sean P. Goggins, as well as the reviewers at the International Conference on Computational Social Science and the Journal of Computational Social Science.

7. OSS Developers' Views on Public and Private Funding: The Case of scikit-learn

Peer review status: *This paper was presented at the 2023 OpenForum Academy Symposium at Technische Universität Berlin in November 2023, and published by the Proceedings of the ACM on Human-Computer Interaction (Track: Computer-Supported Cooperative Work and Social Computing).*

Abstract

The dominance of industry in the OSAI commons underlines the importance of mobilising levers like funding that can support community-led projects as well as public interest alternatives to commercial offerings. The debate about OSS funding has gained traction in recent years. Governments are increasingly getting involved in funding OSS development due to the growing recognition of OSS as digital public goods and digital infrastructure, and have introduced funding programmes to enhance the security of software supply chains and national competitiveness in science and innovation, among others. However, little is known about how OSS developers view the relative benefits and drawbacks of public funding and how it compares to other funding sources. This study explores this question through a case study on scikit-learn, a Python library for ML, which has been funded by public research grants, commercial sponsorship, micro-donations, and a €32 million grant announced in France's national AI strategy in 2021. Through 25 interviews with scikit-learn's maintainers and funders, this study makes two key contributions. First, it contributes empirical findings about the benefits and drawbacks of public and private funding for community-led OSS projects, and the governance protocols employed by the maintainers to balance the diverse interests of their community and funders. Second, it offers practical lessons on funding community-led OSS projects for OSS developers, governments, and companies based on the experience of scikit-learn. The paper concludes with a number of recommendations for OSS funders and future research directions.

7.1 Introduction

The dominance of industry in the OSAI commons underlines the importance of mobilising levers like funding that can support community-led projects as well as public interest alternatives to commercial offerings. The debate about OSS funding has gained traction in recent years. OSS funding has moved up the priority list of governments due to the growing recognition of OSS as digital public goods and digital infrastructure as well as the unsustainability of the status quo. In particular, governments have employed OSS funding as a policy instrument to enhance digital sovereignty (Osborne et al., 2023), the security of software supply chains (Herpig, 2023), the maintenance of digital infrastructure (Keller, 2022; Eghbal, 2016), and national competitiveness in science and innovation (Nagle, 2019), among others. While the increase in governmental interest and involvement in funding OSS development has been generally well-received by OSS developers, little is known about how OSS developers view the relative benefits and drawbacks of public funding and how it compares to other funding sources that have funded OSS developers to date.

This study explores this question through a case study on scikit-learn, a Python library for ML, which is developed by researchers based at the French Institute for Research in Computer Science and Automation (Inria) and a global community of 2,250 contributors. It is described as "the Swiss army knife of ML" due to its widespread use in AI R&D (Inria, 2020). Its sustainability as a community-led project in the industry-dominated field of AI (Ahmed et al., 2023) is partly attributed to its mixed funding model, including public research grants, commercial sponsorship, community donations, and the institutional backing from Inria. In November 2021, the French government published its AI strategy, which included a €32 million grant to support the expansion of scikit-learn into a wider data science commons (SGPI, 2021), marking a significant increase of public funding for scikit-learn.

Through 25 interviews with scikit-learn's maintainers and funders over two years, this study investigates the role of public and private funding in supporting this community-led OSS project. This study makes two key contributions to research and practice. First, it contributes empirical findings about the benefits and drawbacks of public and private funding in an OSS project. Furthermore, it sheds light on how the maintainers employed governance protocols to balance the diverse interests of their funders and to safeguard their community ethos. Second, it offers practical lessons on funding in community-led OSS projects for OSS developers, companies, and governments.

The paper has the following structure. First, it discusses prior work on the role and impact of funding in OSS development (Section 7.2). Second, it presents the study design and the scikit-learn project (Section 7.3). Then, it presents the results and discusses their implications (Section 7.4). Finally, the paper concludes with a discussion of recommendations for practitioners (Section 7.5).

7.2 Related Work

7.2.1 The Role of Funding in OSS Sustainability

The adoption of OSS is ubiquitous: it is estimated to be used in 96% of code bases (Synopsys, 2023) and to constitute up to 90% of software stacks (OpenSSF, 2022). As a result, OSS is considered a critical component of the digital infrastructure that underpins modern society and the digital economy (Eghbal, 2016; Scott, Brackett, Herr, & Hamin, 2023). Given the societal and economic importance of OSS, the health and sustainability of OSS projects is a concern for a diverse range of stakeholders (Goggins, Lumbard, & Germonprez, 2021; Lumbard, Germonprez, & Goggins, 2024). Project health in OSS refers to how well a project can maintain its viability and functionality over time (Linåker, Papatheocharous, & Olsson, 2022), depending on sustainable maintenance from both core maintainers and community contributors (Linåker, Link, & Lumbard, 2024).

A range of social, economic, and technological factors are relevant to an OSS project's health, including open governance (O'Mahony & Bechky, 2008; Trinkenreich, Guizani, Wiese, Sarma, & Steinmacher, 2020); maintainer best practices, such as timely responses (Salkever, 2023) and mentorship (Tan, Chen, Wu, Zhou, & Zhang, 2023); a steady inflow of new contributors (M. Zhou & Mockus, 2015); and diversity of contributors within a community (Bosu & Sultana, 2019). Among these factors, the financial health of a project plays a crucial, but hitherto understudied, role in the sustainability of OSS projects, determining the extent to which development, maintenance, and community animation are funded or reliant on volunteers (Eghbal, 2020).

Funding is increasingly understood as key to the sustainability of OSS projects. However, money has long been a contentious subject among OSS developers. On the one hand, many developers take pride in the origins of the OSS movement as a social movement (Broca, 2021), describing it as a "programmers' paradise" for geeks, hackers, and hobbyists, who do not contribute to OSS for money (Raymond, 2001a). Financial reward is the least important incentive for OSS developers (Gerosa et al., 2021), who often cite their political ideals (Kelty, 2008), altruism (Markus et al., 2000), or their passion for solving bugs (Loebbecke & Angehrn, 2003) as intrinsic incentives for contributing to OSS. Furthermore, many developers complain that "money ruins everything" (Eghbal, 2017). In addition, evidence from the Rust community illustrates that voluntary contributors have some prejudices against paid contributors to OSS projects, such as that they do "do boring work", "rarely care [for] documentation", and "lack personal attachment" (Y. Zhang et al., 2024).

However, concerns regarding maintainer burnout (Eghbal, 2020), security vulnerabilities (Vaughan-Nichols, 2021b), and commercial exploitation of volunteer labour (Birkinbine, 2020; H. Li et al., 2022) have shifted perspectives within OSS developer communities (Osborne, 2024a; Salkever,

2023). A plethora of OSS projects are unfunded and maintained by volunteers, who struggle to keep up with the workload of maintaing their projects (Eghbal, 2020; Geiger et al., 2021).

By many measures, OSS are digital public goods that contribute significantly to science and innovation, yet suffer from a stark imbalance between the resources (including labour and funding) that go into OSS development and the value that is extracted from OSS. According to one study, while supply-side value is around $4.15 billion (M. Hoffmann, Nagle, & Zhou, 2024), its demand-side value is around $8.8 trillion. In other words, it is estimated that it would cost $8.8 trillion to replace OSS that currently exists, a sum that far exceeds the resources that are invested in their development. Similarly, scholars have highlighted the imbalance between the supply of maintenance labour and usage demand of OSS, resulting in a risk of underproduction (Champion & Hill, 2021).

The consequences of systemic issues of free-riding and under-investment in OSS are highlighted when developers spot major vulnerabilities that pose significant threats to digital systems in governments, critical infrastructure, and companies' products and services. In particular, the discovery of vulnerabilities, such as the *log4j* vulnerability in December 2021, which was described as the "worst security problem in a generation" (Vaughan-Nichols, 2021a), shocked industry and governments alike. It highlighted the need for the public and private sectors to fund the developer communities, who build and maintain the open source digital infrastructure of their systems, products, and services (Keller, 2022). As Birkinbine (2020, p.119) has argued, there is an urgent need "not just [for] investment in institutions, organisations, technologies, or innovations, but long-term and sustainable investment in the true source of their value, which is to say, people."

7.2.2 Individual Funding for OSS Development

Individuals or organisations support OSS developers and project by making micro-donations to individual developers and OSS projects. For instance, the GitHub Sponsors scheme, launched in 2019, allows users to make one-off or recurring payments to individuals or organisations with sponsored profiles (GitHub, 2023). Common incentives for giving donations via GitHub Sponsors include expressing appreciation, recognising a developer's work, and encouraging future contributions (X. Zhang et al., 2021). By comparison, common incentives for maintainers to receive donations include receiving recognition for their work or providing a side-income; while maintainers who do not participate in GitHub Sponsors state that they do not need to be sponsored or that they do not contribute to OSS for money (X. Zhang et al., 2021). Furthermore, prior work shows that the amount that individuals donate is associated with the length of their use of an OSS (Krishnamurthy & Tripathi, 2009). The impact of micro-donations is still contested. On the one hand, they have been shown to shorten response times to issues (Nakasai, Hata, Onoue, & Matsumoto, 2017; Nakasai, Hata, & Matsumoto,

2019) and to increase maintenance-related activities (Medappa, Tunc, & Li, 2023). On the other hand, the impact appears trivial (Overney, Meinicke, Kästner, & Vasilescu, 2020) and donations tend to be ad-hoc, small, or legally complicated for individuals to receive (Eghbal, 2022).

7.2.3 Private Funding for OSS Development

The private sector has been the largest funder of OSS to date, funding OSS developers and projects in both direct and indirect ways. Companies make investments in OSS as a mode of open research and development. According to estimates, companies in the USA invested $37.8 billion in OSS in 2019 (Korkmaz et al., 2024), and €1 billion invested in OSS by EU companies yielded up to €95 billion for EU GDP in 2018 (Blind et al., 2021). Open source business models are also a common revenue source for OSS projects (Chesbrough, 2006; Krishnamurthy, 2005a; Birkinbine, 2020). However, as mentioned, there is an extremely stark imbalance between OSS development and use (M. Hoffmann, Nagle, & Zhou, 2024; Champion & Hill, 2021).

Direct funding methods include the sponsoring OSS developers or by allowing their employees to contribute to OSS projects at work (Butler et al., 2021; Xia et al., 2023). Sponsoring developers is a recognised strategy that companies employ to influence projects (Dahlander & Magnusson, 2008; Dahlander & Wallin, 2006) as well as to improve their reputation as OSS patrons and recruit developers (Bonaccorsi & Rossi, 2006; Pitt et al., 2006). In the mid-2000s, it was estimated that 40% of OSS contributors were paid (K. R. Lakhani & Wolf, 2003). We should expect this number to be much higher today, with companies like Google, Microsoft, and Amazon accounting for a rapidly growing number of OSS developers and contributions to OSS projects (Hale, 2022).

In addition, companies fund projects through donations and by joining project steering committees (Butler et al., 2018). Recently, companies have begun to establish FOSS Contributor Funds, which allow employees to nominate and vote on OSS projects that should receive funding (O'Brien, 2019b), and represent a more democratic way for companies to fund the OSS projects that they use (O'Brien, 2019a). A shortcoming is that these funds typically award up to $10,000 over a one-year time frame. Wright from Bloomberg's FOSS Contributor Fund (Wright, 2023) has acknowledged this shortcoming, arguing that "while financial support is important, real engagement through contribution and dialogue is as crucial to the ecosystem's survival as writing a check."

An indirect funding approach is the sponsorship of OSS foundations and consortia, which operate on a fee-paying membership model and facilitate OSS projects and their developer communities through hosting projects, organising events, and offering training courses for novice developers (O'Mahony & Ferraro, 2007). Foundations and consortia are known to play a critical role as "boundary organisations," which through their vendor-neutrality and open governance protocols facilitate

collaboration between diverse volunteers and companies (O'Mahony & Bechky, 2008). The institutional design and governance protocols of such organisations aim to limit the dominance of individual companies (Y. Zhang et al., 2022; Di Giacomo et al., 2020). However, a survey of 369 OSS developers found that only 12% of respondents found funding via a foundation, a consortium, or an independent legal entity either extremely or very useful to their development activity, while 5% of respondents reported that it was ineffective funding approach (Tidelift, 2020).

7.2.4 Public Funding for OSS Development

The public sector has indirectly or directly funded OSS since the origins of the World Wide Web (CERN, 2023) through research grants for OSS (During, 2006; Howison & Herbsleb, 2011) and bespoke funding bodies, such as the Open Technology Fund in the USA, the Next Generation Initiative by the European Commission, or the Sovereign Tech Fund in Germany (Keller, 2022). Furthermore, OSS has moved up the priority list of governments in light of concerns about the security of OSS (Vaughan-Nichols, 2021b; Herpig, 2023); its role as critical digital infrastructure (Eghbal, 2016; Panezi, Feldman, & Bernholz, 2020; Scott et al., 2023); and its importance for digital sovereignty (Osborne et al., 2023; Burwell & Propp, 2022) and national competitiveness in science and innovation (Nagle, 2019). In the EU, the concern for digital sovereignty is particularly salient. For example, the EU's (2020, p.2) OSS strategy argues that OSS can give "Europe a chance to create and maintain its own, independent digital approach and stay in control of its processes."

Public sector support for OSS is also a recognised strategy to stimulate competition in domestic software markets, which might otherwise be monopolised by industry giants (Jokonya, 2015). For example, it has been estimated that the *Circulaire 5608*, a French law requiring government agencies to favour OSS when procuring software, led to an increase of nearly 600,000 OSS contributions from French developers per year as well as yearly increases in the national competitiveness of the French IT market (Nagle, 2019). Beyond governmental adoption of OSS, government investments in OSS are often targeted at stimulating national competitiveness. For example, in June 2023, President Emmanuel Macron of France announced a €40 million fund to create a digital commons to support the development of open LLMs to support domestic start-ups and to challenge the dominance of Big Technology companies in the AI industry (Chatterjee & Volpicelli, 2023).

The discovery of the *Log4Shell* vulnerability in November 2021 marked a turning point for governmental interest and involvement in funding OSS development. In light of the magnitude of the security concerns (Vaughan-Nichols, 2021b), governments rapidly mobilised to address the threats posed by OSS vulnerabilities to their critical infrastructure and digital economies. For example, in January 2022, the White House convened cross-sector stakeholders, including government agencies,

Big Technology companies, and OSS foundations, to discuss the urgent need to improve the security of OSS, recognising its indispensable role as digital infrastructure whose maintenance largely depends on volunteers (House, 2022). Across the pond, in June 2022, the newly formed European Working Team on the Digital Commons, comprising representatives from 19 member states of the EU and the European Commission, published its inaugural report, highlighting the importance of funding OSS for bolstering the digital sovereignty of Europe (on Digital Commons, 2022). In October 2022, Germany established the Sovereign Tech Fund, allocating a budget of €11.5 million for its first year alone to fund OSS maintenance in the interest of digital sovereignty, software security, innovation, and digital democracy (STF, 2022). These developments underscore the growing recognition of the role and responsibility of the public sector in funding the development and maintenance of OSS as digital public goods (Keller, 2022; Osborne et al., 2023).

While there is a growing recognition of the need for public funding for OSS development, little is known about how OSS developers view the relative merits and drawbacks of public funding for OSS development and how it compares to other types of funding that have supported OSS development so far. This study addresses this gap by examining the case of scikit-learn via the following RQs:

- **RQ1**: What are the interests of the public and private funders of the scikit-learn project, and how have they aligned or conflicted with the interests of the maintainers?
- **RQ2**: How do the scikit-learn maintainers view the relative benefits and drawbacks of public and private funding?

By shedding light on these questions, this study extends our understanding of the role of funding in OSS sustainability and yields practical insights for both OSS researchers and practitioners.

7.3 Study Design

7.3.1 Case Presentation

scikit-learn is a Python library that implements ML algorithms for classification, regression, and clustering, as well as tools for data preprocessing, model evaluations, and data visualisation. Initially it was started as *scikits.learn* by David Courapeau during a Google Summer of Code project in 2007. After a dormant period, it was relaunched as scikit-learn by Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and Vincent Michel with the backing of Inria. Its first public release was in February 2010. It has since become one of the most popular OSS projects for ML, described as "the Swiss army knife of ML" (Inria, 2020). In November 2023, it had 2,752 contributors and 61,085

GitHub stars (Insight, 2023). The success of scikit-learn as a community-led OSS project is noteworthy, considering the dominance of industry giants in AI R&D (Whittaker, 2021; Ahmed et al., 2023) and the OSAI commons (Langenkamp & Yue, 2022; Widder et al., 2023).

The sustainability of scikit-learn is partly credited to its mixed funding model (scikit learn, 2023; Varoquaux, 2018). First, Inria has funded scikit-learn since 2010, including by paying salaries, administering research grants, and providing office space, among others. Second, community members have made micro-donations via NUMFOCUS and student projects have been sponsored by the Google Summer of Code programme. Third, the scikit-learn consortium, established under the Inria Foundation in 2018, offers three levels of annual memberships for companies: silver (€30,000), gold (€50,000), and platinum (€100,000), each providing varying degrees of involvement in the Technical and Advisory Committees. The Technical Committee develops the strategic technical roadmap for the project, while the Advisory Committee provides guidance on strategic matters. Since its founding, it has included Dataiku, Microsoft, Nvidia, Intel, AXA, Boston Consulting Group, BNP Paribas Cardif, HF, Nvidia, Fujitsu, and Chanel. Companies, including QuantSight Labs, Nvidia, and HF, have also sponsored maintainers. Fourth, in November 2021, the French AI strategy, "la Stratégie IA," announced a €32 million grant to support scikit-learn (SGPI, 2021).

7.3.2 Methods and Data

This study employed a qualitative research methodology, comprising 25 semi-structured interviews with scikit-learn maintainers, commercial sponsors in the consortium, and government officials between November 2021 and April 2023 (see Table 7.1). The sampling approach sought to ensure comprehensive coverage of the different stakeholder groups in the project's funding model and diversity within each group, assuming heterogeneity within the groups. Interviews were conducted with maintainers with various tenures in the project, including some who had been involved for over 10 years and recent university graduates. One maintainer worked for a company that was a consortium member but spoke on their own behalf. Interviews were conducted with the coordinator and the deputy coordinator of the French AI strategy. I had previously gained access to these stakeholders through their prior employment in AI policy at the UK government. In addition, interviews were conducted with three consortium representatives from companies with different sponsorship tiers, organisational sizes, and sectors. The community manager facilitated access to the consortium members. It was challenging to access to consortium representatives, despite the outreach efforts of the community manager and financial incentives (50€ vouchers). However, given the study's focus on the views of the maintainers, this did not pose a significant threat to the study.

The interviews were semi-structured (Kvale, 1996), combining general questions about scikit-

Table 7.1: List of Respondents and Affiliations

ID	Role	Affiliation
A	Community Manager	scikit-learn
B	Maintainer	scikit-learn
C	Maintainer	scikit-learn
D	Maintainer	scikit-learn
E	Maintainer	scikit-learn
F	Maintainer	scikit-learn
G	Maintainer	scikit-learn
H	AI Strategy Coordinator	French Government
I	Deputy AI Strategy Coordinator	French Government
J	Consortium Member	Anonymous Company
K	Consortium Member	Nvidia
L	Consortium Member	Dataiku

learn's funding model and tailored questions to each interviewee. To tailor questions for maintainers, their contribution histories were reviewed on GitHub's Contributor Insights Page; and for the funders, public records of their funding history were reviewed. I conducted and recorded every interview to aid the analysis and enhance the validity of the findings (Yin, 2018). The interview data was supplemented and triangulated through detailed notes from secondary documents and on-site visits (Yin, 2018). The field notes stem from two on-site visits to the scikit-learn office; two on-site visits to the offices of sponsors; and attendance at the 2021 Global Partnership on AI Summit, hosted by the French government and attended by the AI strategy coordinators and scikit-learn maintainers. The secondary documents include information from the scikit-learn website as well as private documents about scikit-learn's funding model provided by the community manager.

Throughout the 17 months of data collection, I employed an integrated approach to code and identify themes in the interview data, field notes, and secondary documents (Braun & Clarke, 2006). The integrated approach combined grounded theory analysis (Charmaz, 2006) and a deductive approach informed by prior literature about the different interests and types of involvement of individuals, companies, and governments in OSS development (Birkinbine, 2020; Bonaccorsi & Rossi, 2006; Jokonya, 2015; O'Mahony & Bechky, 2008). Thematic codes were continuously developed as data was collected, and in turn translated into themes until the point of saturation (Cruzes & Dybå, 2011). To address the limitations of single author analysis, the resulting themes were member-checked with interviewees to ensure their accuracy (Lincoln & Guba, 1985). In addition, maintained a social identity map throughout the research process to exercise reflexivity and address the influence of personal biases stemming from my personal views on OSS funding and my evolving relationships with the scikit-learn team during the 17 months of data collection (Jacobson & Mustafa, 2019).

7.4 Results

The results are organised into three parts, sequentially responding to **RQ1** on how the funders' interests align with those of the scikit-learn maintainers and **RQ2** on the relative benefits and drawbacks of public and private funding from the perspective of the maintainers.

7.4.1 Stewards of the Community: Roles & Interests of the scikit-learn Maintainers

Key findings

The scikit-learn maintainers, who are partially paid by Inria and and by their respective companies, perform diverse technical and social roles. They are motivated by their scientific interest, their pride in the project's global impact, and a strong sense of stewardship for their community and the broader scientific Python ecosystem. They maintainers credit the contributions of their global contributor base, including both individual volunteers and major tech companies.

The *scikit-learn's* maintainers have various professional backgrounds. While some core team members work full-time on scikit-learn as employees of Inria, others balance their maintainer responsibilities with their day jobs as professors at universities or as ML engineers at companies. In addition, the core team includes a few company-sponsored developers, who make focused contributions to the project, such as the development of new features, which are agreed upon by the maintainers and the respective sponsor companies. For example, Nvidia sponsors a full-time maintainer, who is building a plug-in to enable alternative computational backends for scikit-learn to enable compatibility with Nvidia's GPU AI hardware accelerators.

The maintainers explained that they perform a range of technical and social tasks, including reviewing code, mentoring contributors, and organising community events. Respondent A explained, "Of course, software development itself is the major part of the workload—going through issues, pull requests, and responding to community feedback—but lots of other work goes into the project. Community animation is a huge part." Several maintainers expressed a strong sense of stewardship for the project's diverse community, comprising contributors (see Table 7.2a) and companies (see Table 7.2b) from across the world.[1] Respondent B highlighted the importance of community contributions, stating, "The major part of the labour [in the project] is based on benevolence of people working in their free time and not asking to get paid." Furthermore, Respondent C highlighted community members have an additional impact on the project by making micro-donations to the

[1]Tables 7.2a- 7.2b record the top 10 countries and companies for three categories of users: A `stargazer` is a user that stars a GitHub repository, akin to liking or following a repository; an `issue creator` is a user that creates an issue in a GitHub repository; and a `pull request creator` is a user that creates a pull request in a GitHub repository.

Table 7.2: Top Countries and Organisations by Community Engagement in the scikit-learn Project

(a) Top 10 Countries by Community Engagement in the scikit-learn Project

Rank	Stargazer	Issue Creator	Pull Request Creator
1	China (23.7%)	USA (33.7%)	USA (34.0%)
2	USA (21.8%)	Germany (7.5%)	India (8.9%)
3	India (7.9%)	UK (7.0%)	Germany (8.5%)
4	Germany (4.4%)	India (6.8%)	France (6.9%)
5	Brazil (3.8%)	France (5.5%)	UK (6.4%)
6	UK (3.7%)	China (5.4%)	Canada (4.5%)
7	Canada (3.3%)	Canada (4.5%)	China (2.6%)
8	France (3.2%)	Netherlands (2.5%)	Japan (2.5%)
9	Japan (2.5%)	Switzerland (2.2%)	Switzerland (2.3%)
10	Russia (1.3%)	Australia (2.0%)	Netherlands (2.2%)

Source: `https://ossinsight.io/analyze/scikit-learn/scikit-learn#overview` (2023)

(b) Top 10 Organisations by Community Engagement in the scikit-learn Project

Rank	Stargazer	Issue Creator	Pull Request Creator
1	Google (1.5%)	Google (2.5%)	Google (3.7%)
2	Microsoft (1.5%)	Microsoft (2.0%)	Microsoft (2.5%)
3	Tencent (0.9%)	Amazon (0.7%)	Inria (0.9%)
4	Alibaba (0.8%)	DeepMind (0.7%)	Meta (0.8%)
5	Amazon (0.8%)	IBM (0.6%)	Amazon (0.8%)
6	ByteDance (0.8%)	Johns Hopkins (0.6%)	Dataiku (0.7%)
7	Baidu (0.6%)	Tencent (0.5%)	UC Berkeley (0.6%)
8	Meta (0.5%)	ETH Zurich (0.5%)	Shopify (0.6%)
9	AWS (0.4%)	Uni. of Washington (0.4%)	MIT (0.6%)
10	Tsinghua University (0.4%)	NeuroData (0.4%)	AWS (0.6%)

Source: `https://ossinsight.io/analyze/scikit-learn/scikit-learn#overview` (2023)

project. They explained that while "the major stuff is not funded through NUMFOCUS donations," the donations have funded marketing, events, and an internship for underrepresented groups.

In addition, several maintainers expressed their scientific interest in ML as well as the project's scientific culture as important incentives for maintaining scikit-learn. Many maintainers had academic backgrounds, worked in research labs, or taught at universities. For example, Respondent B said that they enjoyed working at the intersection of statistical theory and software engineering, which was lacking in their previous industry job. In addition, Respondent E explained that while they may not become a millionaire, they had a comfortable life and enjoyed the intellectual challenge of the job. Furthermore, for many, being a scikit-learn maintainer was a source of pride due to its global reputation and usage among researchers and data scientists across the world.

Beyond maintaining scikit-learn, several maintainers discussed the importance of contributing to their core dependencies, such as *numpy* and *scipy*, and the wider scientific Python ecosystem. Respondent A explained that they collaborate with other Python projects by applying for grants together. However, their ability to support their dependency is constrained by their limited capacity.

For example, Respondent B explained that while projects like scikit-learn, *numpy*, and *scipy* are very important for ML researchers and engineers, "There are many which are not visible to the public, for example OpenBLAS which is fundamental and only maintained by two people, who have very specific expertise. This is tremendous software work, but it is invisible to the public. Projects like this are cornerstones for many projects. They are the foundations; they are infrastructure." This ethos resonated with other maintainers, for whom supporting their dependencies was an unquestionable priority to ensure the sustainability of both scikit-learn and the broader scientific Python ecosystem.

7.4.2 Public Funding: From Research Grants at Inria to la Stratégie IA

Key findings

While public funding has supported scikit-learn since 2010 through Inria (€1.5 million), France's AI strategy grant (€32 million) marks a significant expansion aimed at enhancing national competitiveness in AI R&D, increasing economic adoption of AI, and supporting digital sovereignty. While the grant was welcomed by maintainers, they faced several challenges in aligning the government's policy goals with their expertise and needs as OSS developers.

Public funding has supported the development of scikit-learn since its beginning in 2010. Over 13 years since its first public release, Inria has provided an estimated €1.5 million through staff salaries, research grants, office space, computing resources, and event sponsorship. This led Respondent G to explain that, "[Public funding] is not new with the AI strategy." However, the €32 million grant announced in France's AI strategy marked a significant expansion of public funding, supporting the expansion of scikit-learn and the development of a commons for data science. The size of the grant is underpinned by clear policy objectives. According to Respondent I, the government sought to enhance France's competitiveness in AI R&D, to facilitate AI adoption by businesses in France, and to support the digital sovereignty of France and Europe. Respondent H underscored the importance of supporting alternatives that could reduce dependence on software sold by US industry giants.

The maintainers were explained that they were delighted to receive this funding and recognition from the French government. Respondent C highlighted the stability it provided for the project, enabling longer-term planning for hiring and the technical roadmap. Respondent B characterised it as "a good signal that the French state is getting involved in funding projects like this." Respondent A particularly commended the requirement to seek matching funds from both public and private sources across the EU, viewing this as an important safeguard for the project's independence, describing it as "dangerous for us to be exclusively funded by the private sector or the French government."

At the same time, challenges emerged in aligning the government's policy goals with the expertise

of the maintainers. The initial demand by the government officials to expand scikit-learn's functionality to DL in order to provide a "sovereign" alternative to the PyTorch and TensorFlow frameworks was resisted by the maintainers, who insisted on focusing on the project's specialism in ML and the need to allocate sufficient funding to maintenance. The coordinators of the AI strategy deferred to their expertise, with Respondent H acknowledging that it does not make sense for scikit-learn compete against TensorFlow and PyTorch. However, other demands have proven trickier. A requirement to simultaneously develop a commons of OSS for data science whilst not competing with domestic companies that offer data science tools and services left the maintainers in a bind, considering that the resulting tools would provide free alternatives to the proprietary offerings of companies. Maintainers also expressed frustration with the slow pace of disbursement, with funds still not arriving 17 months after the announcement at the time of the last interview. While scikit-learn's diversified funding model provided a buffer for the project, Respondent A worried that such sluggishness could threaten the sustainability of more precarious OSS projects.

Multi-stakeholder consultation played an important role in including the grant in the AI strategy. Respondent H explained that the decision to fund scikit-learn, as well as how to fund it, was informed by an ongoing multi-stakeholder process: "It was not just the government. It was a dialogue between several actors—with the AI strategy team, the 'Confiance IA' project,[2] Inria, with scikit-learn, and so on." He confessed they had faced hurdles within the government: "Financing open source means financing a product that cannot be valued on the market and *a priori* it is not clear what its success will be. So, if [we want] the government to finance it in a substantial way—here we are talking about €32 million—the government wants to know where the money is going and to achieve a measurable return on investment." Ultimately, it took months of coordination for the grant approved internally.

Two consortium members raised concerns that such a substantial grant could politicise scikit-learn, potentially influencing it to prioritise French policy goals over the interests of the community. For example, Respondent J underlined the need to balance the French government's concern about its digital sovereignty with the norms of *scikit-learn's* global contributor community, while Respondent K argued that imposing French policy priorities on the project would not be received well by the community. These concerns were acknowledged by the maintainers, who insisted that the community ethos of the project remained at the heart of their discussions with the government officials.

[2] "Confiance IA" is a government-sponsored programme that aims to develop and industrialise trustworthy AI technologies in France.

7.4.3 Commercial Sponsorship: The Role of the scikit-learn Consortium

Key findings

The scikit-learn consortium provides stable funding from companies seeking to support maintenance, provide input in the technical roadmap, and build goodwill with the community. While sponsors' interests generally align with maintainers' goals, there have been conflicts of interests and governance protocols have been implemented to limit sponsors' influence.

Companies have funded scikit-learn directly through membership in the consortium or indirectly through the sponsorship of maintainers. Companies in the consortium have pursued various social, economic, and technological goals. Respondent A explained that companies typically join the consortium to support the maintenance of scikit-learn because they use it in their products or services, as well as to build goodwill with the scikit-learn community, which can be useful for hiring ML researchers and developers. He commented, "A lot of people come from this open source background and if you see that a company is promoting such an open source project, you might say, 'Okay, maybe, I would be happy working with them.' They could be okay with me contributing to OSS at work." He suggested this incentive may be more important for companies that do not already have a strong reputation in ML or AI, making them more attractive to graduate students and job-seekers.

For their part, consortium members cited a mix of strategic and principled reasons for funding scikit-learn. On a basic level, it allows them to support the maintenance of a tool they use and rely on. For example, Respondent K shared that their concern about unmaintained dependencies "keeps [him] up at night," citing *numpy* as a critical library that had been under-maintained for some time. Thus, funding the project was a means to handle this concern. Respondent L expressed a similar sentiment, explaining, "We are very much aware that we rely on open tools like scikit-learn" and they felt that it was important to "give back to the community." Respondent J was even more direct: "Companies pay millions in commercial licenses, so why not support the OSS projects we use?" Maintainers noted their appreciation for private sponsors' attention to maintenance and financial support for the project. Respondent B commented that they appreciates the exchanges in the consortium meetings because "it shows culturally that those companies are not just using the software because it's free and it works; they are giving money to make the project sustainable."

Representatives of consortium members admitted that their sponsorship was not simply charity. In exchange for sponsorship, they prized gaining a voice in scikit-learn's development through membership in the Technical Committee and Advisory Board, which are venues for surfacing industry use cases and needs. Respondent L characterised the interactions in these committees as a "win-win" for the maintainers and the companies, allowing scikit-learn to benefit from "the perspective of people

in industry, which...is a very different type of feedback from the one they get from the open source community," whilst giving consortium members direct input into the project.

In practice, this dynamic has yielded synergies and tensions. In large part, companies' interests have been compatible with those of the maintainers. For example, Nvidia has sponsored a maintainer to build a plug-in enabling alternative computational backends for scikit-learn. Respondent K explained, "Nvidia is very interested in scikit-learn's effort to accelerate with GPUs. I would love to see scikit-learn accelerated natively with a nice user experience, like *PyTorch*, which you can install and accelerate without changing code." Respondent D described Nvidia's simultaneous consortium membership and maintainer sponsorship as "a win-win situation" for the project. However, in other cases, sponsors' attempts to influence the project have run afoul of the project's norms. Two maintainers explained that a sponsor had once requested modifications to the technical documentation to link to their website and tools, which amounted in their view to free advertising of the company, which was unacceptable to the maintainers. While disputes of this kind have been rare, a dispute with a consortium member had led to the termination of their sponsorship in the past.

To balance the diverse interests of their community and the member companies in the consortium, the scikit-learn maintainers have employed governance protocols that limit funders' influence and prioritise community interests. For instance, the Technical Committee cannot override decisions that achieve rough consensus among the maintainers. Respondent C explained, "We don't want the technical committee to have power over the core contributor team as long as [the maintainers] have a consensus. If there's two thirds majority in cast vote, for whatever vote we do, then the technical committee doesn't step in...because the power is given to the [maintainers] and that's by design." These controls are important to maintain the trust of the community and to mitigate perceptions the project has been bought out. Respondent A explained that as a community-led OSS project, they communicate to the consortium that their feedback is valued but they cannot guarantee that they will implement it. For example, the maintainers organised a workshop on MLOps because this was a topic that the consortium members had expressed an interest in. "MLOps will probably end up in the technical roadmap in some way...but it's a discussion and in the end the team will decide if it's worth it or not," they explained. Through these governance protocols, the maintainers aim to ensure that the project remains "a community with different interests."

Finally, even though relationships with commercial sponsors can be troublesome at times, several maintainers explained they were grateful for their sponsorship, unlike most companies which simply free-ride on scikit-learn. There are a number of companies that contribute to the project in kind (i.e. in code) but the maintainers explained that this kind of contribution creates additional work for the maintainers, who struggle with towering workloads. For example, during an on-site visit, Respondent

A showed me a dashboard, comprising world maps of contributors and charts of contributions from companies. Upon observing that the most active companies are Chinese and American giants (see Table 7.2b), I asked if these insights had influenced their opinion about whether they contribute sufficiently to the project. He explained that while these companies typically contribute useful issues and pull requests, at the end of the day their contributions create more work for the maintainers. In addition to code contributions, they explained, it would be helpful if these companies sponsored a maintainer or joined the consortium, which would enhance the maintainers' capacity to process the workload created by their contributions and ultimately implement them.

7.5 Discussion

7.5.1 Implications for Research

7.5.1.1 Benefits and Drawbacks of Public and Private Funding

This study contributes novel findings on how OSS developers view the relative benefits and drawbacks of public and private funding. In the case of scikit-learn, private funding is associated with a welcome understanding of the importance of maintenance and it has introduced the maintainers to industry use cases, which have resulted in new features. Yet it has also required careful management of sponsors' interests and demands, which in some cases can conflict with community norms. By contrast, the sizeable public grant from the French AI strategy has enabled the scikit-learn maintainers to undertake long-term planning for the future of scikit-learn, which they explained is rare for OSS developers, but it has involved challenges in aligning the government's policy goals with their needs and wants as OSS developers. These findings show that a diversified funding model can be an effective strategy for balancing the relative benefits and drawbacks of public and private funding. As Respondent A explained, "Overall, it's good to have a panaché."[3]

The study also contributes novel insights about the diverse interests of OSS funders, from the economic and technological goals of companies (Bonaccorsi & Rossi, 2006) to the policy objectives of governments like digital sovereignty (Osborne et al., 2023), national competitiveness (Nagle, 2019), and economic growth (Jokonya, 2015). While public and private funders pursue different goals, the case study illustrates that they are not necessarily in conflict, sharing the goal of sustaining scikit-learn as a public good. For example, while the maintainers and the companies do not necessarily share the government's policy goals, such as digital sovereignty, its long-term investment in scikit-learn aligns with the maintainers' and companies' interest in its sustainability for research and innovation purposes. Similarly, while the maintainers and the government officials do not necessarily

[3] "Panaché" is the French word for a drink that is usually half lemonade and half beer.

share the companies' business interests in scikit-learn, the project benefits from companies' feedback and funding for feature development and maintenance.

7.5.1.2 Future Research Directions on OSS Funding

There is still a dearth of research on the nature and impact of OSS funding. With new funding models emerging, from FOSS Contributor Funds to governmental funds like Germany's Sovereign Tech Fund, it is timely to empirically investigate emerging OSS funding models from multiple angles, including funders, incentives, governance models, and impacts. It is important to recognise that impacts can also be negative, given that the injection of money changes the social fabric of OSS developer communities. It would be valuable to conduct research on cases when funding has not been helpful in order to produce practical guidance for OSS developers about accepting—as well as when not to accept—funding. Furthermore, future research could investigate the design and outcomes of different OSS funding programmes, from top-down funding approaches to grassroots funding initiatives, in order to shed light on the most impactful ways for organisations, such as governments or public research agencies, to provide funding. By examining these aspects, researchers can contribute to the development of best practices and guidelines for both existing and new OSS funding programmes.

7.5.2 Implications for Practice

7.5.2.1 Recommendations for OSS Developers

The case of scikit-learn shows that a diversified funding model together with funding governance can be an effective strategy for safeguarding a project's independence whilst benefiting from both public and commercial sponsorship. Furthermore, public and private funding come with their unique benefits and drawbacks. On the one hand, private funders understand the importance of maintenance and share industry use cases that lead to the development of new features. However, their demands can conflict with community norms. On the other hand, public funding can provide sizeable grants, which support long-term planning but may prioritise innovation over maintenance or policy goals over the specialism or needs of the project. In both cases, maintainers must be prepared to advocate for and defend their project's needs and community values. Open governance is critical and a diverse funding model can provide stability and leverage in these discussions.

7.5.2.2 Recommendations for Companies

For companies, the findings are a timely reminder that sponsoring OSS projects can make a substantial difference to maintainers, who often struggle with limited capacity and towering workloads. As

shown in Table 7.2b, industry giants, such as Google, Tencent, Alibaba, Amazon, and Meta, account for the most `stargazers` (and presumably, users) of scikit-learn, but they have not funded the project. While these companies contribute useful issues and pull requests, these contributions create more work for the maintainers. Sponsoring a developer or a project represents a low-effort, high-impact way for such companies to support the OSS developers and projects that they use and depend on.

7.5.2.3 Recommendations for Governments

For governments, the case study highlights the importance of refraining from the novelty bias when funding OSS development; that is to say, they should ensure adequate funding for the maintenance of existing OSS in addition to new innovation (Strasser, Hertweck, Greenberg, Taraborelli, & Vu, 2022). Furthermore, a multi-stakeholder processes that involve the views of diverse stakeholders from the OSS community ensure that funding programmes are designed in a manner that is cognisant to the challenges and needs of OSS developers. Without this input, policymakers risk designing and implementing funding programmes that fail to support the actual needs of OSS developers.

7.5.3 Threats to Validity

The study's validity was assessed following guidance for qualitative case studies in software engineering research (Runeson & Höst, 2008; Yin, 2018; Easterbrook et al., 2008).

7.5.3.1 Credibility

Credibility refers to the believability of the findings (Easterbrook et al., 2008). I took three steps to enhance the credibility of the findings according to best practices (Lincoln & Guba, 1985; Yin, 2018). First, I combined multiple data sources—interviews, fieldwork, and secondary documents—to triangulate findings. Second, my prolonged engagement over 17 months built trust with the maintainers and provided in-depth insights, minimising the biases of snapshots at one moment in time. Third, I member-checked the findings with the respondents to enhance their accuracy.

7.5.3.2 Robustness

Robustness concerns the reliability and soundness of the study's design and findings. I took a number of steps to enhance the robustness. First, I used an iterative inductive-deductive analysis approach during data collection, guided both by theoretical insights from prior work and arguments that were emphasised by the interviewees (Cruzes & Dybå, 2011). Second, I maintained a reflexivity map to control for personal biases during the 17 months of data collection (Jacobson & Mustafa, 2019).

7.5.3.3 Transferability

Transferability concerns the generalisability of findings in qualitative research. The single case study threatens the generalisability of the findings (Eisenhardt, 1989) and raises the risk of case uniqueness (Herriott & Firestone, 1983). However, these threats can be tempered by considering that case studies aim for analytical generalisation, not statistical generalisation (Yin, 2018), and that the research aim was exploratory. Specifically, it sought to shed light on how OSS developers view the relative merits and drawbacks of public and private funding, and to offer practical recommendations based on their experience. With this in mind, generalisability is not a major concern for this study.

7.5.3.4 Dependability

Dependability refers to the consistency of the research process. To maximise the dependability, a systematic study protocol was employed for the collection, analysis, and storage of research data. Where consent was given, interviews were recorded and transcribed to aid analysis. When interview responses were ambiguous, follow-up discussions with the interviewees were initiated to obtain clarifications. In addition to interview data, detailed field notes from on-site visits and supplementary secondary documents facilitated triangulation. Finally, the case study underwent thorough review, benefiting from the feedback from domain experts and the peer review process.

7.6 Conclusion

This case study on the funding model of scikit-learn makes two key contributions to research and practice. First, it extends the literature on the role of funding in the political economy of OSS with novel findings on how OSS developers view the relative benefits and drawbacks of public and private funding, and how they balance different funding sources to support their project's goals. Second, it offers practical lessons for the funding of community-led OSS projects and makes practical recommendations to OSS developers, governments, and companies. Crucially, it highlights the effectiveness of diversified funding models, as well as carefully crafted funding governance protocols, as an approach to sustaining community-led OSS projects as well as simple ways in which both governments and companies can effectively co-operate to support the long-term sustainability of OSS projects.

8. Conclusion

This chapter concludes the thesis with a summary (Section 8.1), a discussion of the key contributions and future research directions for the political economy of OSAI (Section 8.2), as well as concluding remarks on the potential role of OSAI in the emerging Public AI policy agenda (Section 8.3).

8.1 Summary of Thesis

This thesis makes both theoretical and empirical contributions to the nascent research agenda on the political economy of OSAI. In Chapters 1-2, I extended Widder et al's (2023, 2024) theoretical foundation for the political economy of OSAI by connecting it with scholarship on the critical political economy of the digital commons and the supply chain capitalism of AI, which respectively provide theoretical frameworks to scrutinise commercial interests in OSAI and to locate OSAI development within the material dependencies and power structures of the AI supply chain. Then, through four RPs, I empirically examined commercial interests and involvement in the development of OSAI technologies and the implications thereof for development practices and norms in the OSAI ecosystem. While companies often frame their OSAI activity as acts of AI democratisation, these RPs collectively show that strategic business objectives are paramount. In particular, RP1-3 shine a light on how the OSAI ecosystem in large part operates as a "commons of capital," where companies collaborate and compete on the development of OSAI technologies. In RP1, we saw the role of commercial agendas in the governance choices and development practices in three critically important OSAI projects, from strategic partnerships to the normalisation of the ecosystem strategies of industry giants that seek to make their "flywheels" spin faster. In RP2, we saw that even when companies adopt open governance in their projects through donations to non-profit foundations, there is a commercial logic at play: be it to integrate in an ecosystem of companies, attract developers to their projects, or shape industry standards, among others. Then, RP3 provided empirical evidence that, rather than reconfigure power structures in AI development, the development of OMs is characterised by commercial dominance, with a small number of industry leaders enjoying unparalleled adoption of their OMs.

The implications of the prevalent commercial participation in OSAI development are manifold.

Commercial contributions have certainly played a major role in the adoption and development of state-of-the-art AI technologies beyond a handful of powerful companies. However, AI democratisation is not merely a question of access. We must challenge the terms and conditions of this access, in particular commercial ecosystem strategies and single-vendor governance models that seek to strengthen the hold of industry giants. What is more, the dominance of a handful of industry leaders in the OSAI ecosystem begs the question of what can be done—or rather, what must be done—to resist commercial dominance and mitigate the crowding out of public interest initiatives. RP4 provides a strong case for the role that public funding, alongside private funding, can play in supporting development and maintenance in community-led OSS projects. However, given the compute-intensive nature of AI development, funding for OSS alone will not be sufficient; rather, public investments will be necessary at each layer of the AI stack (Public.AI, 2024). At the same time, we must recognise that competing with industry giants at their own compute-intensive game will be a costly and perhaps futile endeavour. As Varoquaux et al. (2024) argue, it is becoming increasingly urgent that we reconsider this path and instead pursue approaches to AI development that are more affordable, environmentally sustainable, and conducive to democratic governance.

8.2 Discussion of Key Contributions and Future Directions

In this section, I discuss the key theoretical and empirical contributions as well as future research directions for the nascent research agenda on the political economy of OSAI.

8.2.1 Theoretical Foundations for the Political Economy of OSAI

In Chapters 1-2, I extended the theoretical framework for the nascent research agenda on the political economy of OSAI developed by Widder et al (2023, 2024). Their work was a timely, critical intervention in the debate about the promise and perils of OSAI, providing a critical lens on how OSAI may present an avenue for already powerful industry giants to extend their dominance in AI, rather than leading to the democratising or equalizing effects that are commonly cited in the "AI democratisation" discourse. In particular, they draw our attention to how powerful corporations in the AI industry may seek to assert their "dominance through—not in spite of—open-source" (Widder, Whittaker, & West, 2024). However, their work is marked by, at least, two limitations. First, they do not define political economy or power. It is important to define these concepts, as the definitions determine what one can "see" or question. Second, they do not adequately draw on relevant prior scholarship on the critical political economy of the digital commons and the political economy of AI, which respectively provide theoretical frameworks that can inform this emerging research agenda.

I addressed these limitations by, firstly, drawing on definitions from the political economy of communications (Mosco, 2009) and, secondly, by bridging these previously disconnected streams of scholarship. In particular, the critical political economy of digital commons offers a critical perspective on how commons-based peer production, including OSAI development, has become incorporated by digital capitalism (Birkinbine, 2020). It illustrates how companies seek to harness the collective labour power of developer communities, while using the rhetoric of openness to advance their strategic objectives. Above all, it focuses our attention on how companies do not merely seek to capitalise on the *products* of OSS developer communities, but more importantly on their *processes* of distributed development. This lens, in turn, helps us to critically examine similar strategies in OSAI development, adding a theoretical layer to Widder et al's (2024) comparative analysis of commercial strategies in the development of OSS and "open" AI systems. Moreover, this framework provides theoretical and practical tools to identify how OSAI developer communities may negotiate their boundaries with companies and safeguard their interests. It points us to the conditions that must be in place and the actions that must be taken to ensure the sustainability of community-led OSAI projects.

Then, I drew on the supply chain capitalism of AI (Valdivia, 2024), as well as prior work on industry concentrations along this supply chain, to extend Widder et al's analysis of "the materials—models, data, labour, frameworks, and computational power—frequently involved in creating and using large AI systems." (Widder, Whittaker, & West, 2024, p.827). Understanding the material dependencies of AI development is essential for evaluating the potential of OSAI to either challenge or reinforce the power of industry giants. There may be over a million AI OSS repositories on GitHub and over a million OMs on HF Hub, but their availability does not change the fact that there are monopolies and oligopolies at the various stages of the AI supply chain, which influence the development, distribution, and consumption of OSAI technologies. The empirical RPs contribute back to this body of scholarship, illustrating that strategic partnerships between AI companies, chip makers, and cloud providers drive the development of key OSS for AI R&D as well as that 1% of models on the HF Hub, developed by a handful of industry leaders, account for 99% of downloads.

In summary, by bridging these complementary streams of scholarship, this thesis builds on Widder et al and offers an extended theoretical framework for future research on the political economy of OSAI. Through this integration, I hope to encourage dialogue between scholars in these disciplines as well as new research directions on the political economy of OSAI. What is more, it provides a practical framework for practitioners and policymakers that seek to promote public interest OSAI initiatives in the industry-dominated ecosystem. The aforementioned findings are not a reason not to abandon OSAI that may serve the public interest; rather, they underline the need to create the conditions along the wider AI supply chain that can enable public interest OSAI initiatives to succeed.

8.2.2 Debunking the Myth of AI Democratisation: Commercial Interests in OSAI

The OSAI ecosystem has undergone a remarkable transformation since 2007, when researchers lamented the dearth of OSS as a critical impediment to AI R&D (Sonnenburg et al., 2007). Nowadays over a million AI OSS repositories AI and over a million OMs, including state-of-the-art DL frameworks and open LLMs, are available at the fingertips of researchers and developers across the globe. This growth has in large part been driven by commercial contributions, following a broader pattern in the software industry where open source has evolved from a fringe activity of hackers and hobbyists to a quotidian mode of software development (Broca, 2021). However, what distinguishes AI from other domains is the extraordinary scale of commercial investments in AI R&D, which make public funding pale in comparison. In particular, the AI democratisation narrative has become ubiquitous in the AI industry, heralding open source as an antidote to industry concentrations. However, it is a notoriously ambiguous term that in most cases simply means enabling the use or development of AI technologies—for example, through open-sourcing AI software or models—rather than democratising their governance (Seger, Ovadya, et al., 2023).

Against this backdrop, in this thesis, I examined why and how companies participate in the development of OSAI technologies and the implications thereof for development practices, governance norms, and the potential trajectories of the OSAI ecosystem. In the following sub-sections, I discuss these findings and future research directions for the political economy of OSAI.

8.2.2.1 Commercial Incentives for Collaboration on OSAI Development

Indeed, in the last two decades, OSS has evolved significantly from its humble beginnings in hacker communities and academic labs. Many OSS developer communities have evolved "from networks of individuals into networks of companies" (Ågerfalk & Fitzgerald, 2008, p.396), which include market rivals who collaborate on the development of base technologies in order to share R&D costs (Teixeira et al., 2016; Nguyen-Duc et al., 2019), shape industry standards (Lerner & Tirole, 2005; Lindman et al., 2009), and accelerate innovation (Chesbrough, 2023; Birkinbine, 2020), among others. Chapters 4-5 (RP1-2) demonstrate that OSAI is no exception to this trend.

Chapter 4 (RP1) contributes empirical insights into the structures, strategies, and practices of inter-company collaborations that underpin the development of critical OSS projects in AI R&D. In particular, it identifies three distinct types of collaborations in company-hosted OSS projects. Strategic collaborations, typically dyadic relationships between a host and an external company, are driven by competitive interests to ensure the compatibility and competitiveness of their interdependent products. Contractual collaborations involve the outsourcing of development tasks to third-party

companies, as exemplified by Meta's contracting of QuantSight to develop PyTorch. These relationships blur the lines between contributors who are affiliated with the host company and the contracted company. Meanwhile, non-strategic collaborations encompass various contributions that are motivated by personal interest, bug-fixing needs, and corporate OSS initiatives, which are performed by developers during working hours but are not driven by commercial objectives.

What is more, the strategic collaborations reflect competitive dynamics among dominant companies across the AI supply chain. At the hardware layer, AI accelerator manufacturers like Nvidia, Intel, and AMD collaborate with Google and Meta to ensure the optimal performance of their respective AI accelerator offerings with the DL frameworks. Google even contributes to Meta's PyTorch, not despite the fact but precisely because of the fact that it is also in the business of selling AI accelerators. At the model layer, AI companies like Meta and Stability AI collaborate with HF to ensure day-zero integrations of their models with HF's platform and tools. At the cloud layer, HF collaborates with hyperscalers like AWS, Google Cloud, and Microsoft Azure to ensure seamless deployment of AI models on their respective cloud platforms. These collaborations are strategically important for all participating companies whose hardware, software, and compute products are independent. The hardware manufacturers need their accelerators to work well with the dominant DL frameworks, AI companies need their models to be accessible via popular platforms like HF Hub, and cloud providers need to be able to support the deployment of models on their servers.

Strategic collaborations like the ones observed in Chapter 4 (RP1) are complemented by inter-company collaborations in OSAI projects hosted by vendor-neutral foundations. As shown in Chapter 5 (RP2), while companies often publicly frame their AI software donations as acts of AI democratisation, they instrumentalise the democratisation of their project's governance as a means to achieve social, economic, and technological goals. Across the board, companies seek to attract new contributors to their projects and reduce development costs, with 100% of survey respondents citing the recruitment of new contributors as an important incentive. For startups and smaller companies, foundation governance provides crucial stability offerings and enhanced credibility in the AI industry. Larger companies often seek to integrate with existing ecosystems and facilitate collaboration with other projects. A salient incentive, especially for industry leaders, is the ability to shape industry standards via the adoption of open governance. For example, the donation of PyTorch by Meta was viewed by respondents as aimed at establishing PyTorch as the de facto DL framework standard. The formation of the PyTorch Foundation's governing board, comprising major industry players like AMD, Intel, Nvidia, AWS, HF, and even Google Cloud, created a powerful alliance that promised to increase investments by these industry giants in the further development and adoption of PyTorch. As one respondent noted, this strategic alliance represented "a death knell to TensorFlow."

These findings provide a baseline for further research on commercial interests and involvement in OSAI development. In light of increasing commercial activity in OSS development (Germonprez et al., 2018), it is time for more, not less, scholarship on commercial interests, activity, and governance approaches. The findings in Chapter 4 (RP1) provide fruitful ground for future research on inter-company collaboration on OSS development in and beyond the AI industry. Beyond OSS, future research could investigate emerging forms of collaboration in OM development, which introduces unique aspects of collaboration on a wider set of components beyond just software. For example, the BigScience project, which involved over 1,000 contributors from more than 250 organisations, offers an example of community-led OM project that involved collaborations on the development, auditing, and governance of both data and software components as well as the resulting model. In addition, the taxonomy of commercial incentives in Chapter 5 (RP2) provides a foundation for studying other AI democratisation efforts, in particular the commercial release of OMs. Many of the identified incentives, such as increasing adoption, building or joining an ecosystem, and influencing standards, likely apply to model releases, too. Future research could build on these findings by investigating the commercial incentives and governance approaches associated with OM releases, exploring how they align with or diverge from those in the release or donation of AI OSS projects.

8.2.2.2 OM Ecosystems: A New Arena for Commercial Competition

A central finding emerging from Chapters 4-6 (RP1-3) is the importance of commercial ecosystem strategies in OSAI development, from industry giants that seek to control ecosystems to others that seek to join and collaborate in ecosystems. These divergent strategies reflect not only differences in the resources and market power of the respective companies but also distinct theories of value creation and capture. For industry giants, the pursuit of ecosystem control represents more than just a development strategy; it is an approach to establish or maintain their market dominance through open source. This was evident in Google's leaked memo expressing their explicit aim to "own the ecosystem" and "let open source work for us" (Patel & Ahmad, 2023), as well as in Mark Zuckerberg's public statements about creating self-reinforcing "flywheels" that could improve their Llama models (South Park Commons, 2024). In addition, Meta's donation of PyTorch exemplifies how governance can be weaponised in ecosystem battles between industry giants: by establishing a governing board comprising industry giants, Meta created a powerful corporate alliance that Google would struggle to compete with by itself. These ecosystem-building strategies align with ecosystem innovation and platform ownership strategies in the technology sector, where OSS or OM ecosystem ownership is understood to provide companies with competitive advantages that make it harder for market rivals to dislodge them (Gawer & Cusumano, 2014).

The emergence of OMs represents the next frontline of OSAI ecosystem competition. While OMs are often portrayed as means to level the playing field, Chapter 6 (RP3) reveals a different reality: the adoption and development of OMs is characterised by extreme imbalances, for example with 1% of OMs on HF Hub accounting for 99% of all downloads. Rather than equalise the field, the compute-intensive nature of AI development that only a handful of industry leaders can afford, coupled with commercial strategies that seek to increase adoption and engagement within their ecosystems, have entrenched the status quo of commercial dominance in AI R&D. These findings provide a baseline for future research on the social dynamics of OM development. Future research could explore the ecosystem-building strategies of industry leaders, in particular the governance and community involvement strategies that are employed to gather contributions by distributed researchers and developers, from model audits to model fine-tunes for specific use cases. In addition, future research could examine how the ecosystem strategies have influenced the trajectory of OM development, particularly regarding what types of OMs have been or could be built on top of the dominant OMs. Furthermore, future research could examine in what ways HF influences OM development practices and norms, as a start-up that does not have the resources of an industry giant yet owns the platform that AI researchers and developers, including industry giants, depend on to discover, share, and develop OMs. Understanding these dynamics will be crucial, as the OM ecosystem is increasingly becoming an important battleground for commercial influence in AI R&D.

8.2.2.3 Opportunities and Challenges for Community-led OSAI Development

The prevalence and intensity of commercial activity in OSAI development shown in Chapters 4-7 (RP1-3) raise questions about the feasibility and long-term sustainability of community-led OSAI development; that is, OSAI projects that are developed and governed by a community of diverse stakeholders, who are not collectively motivated by a commercial agenda. Chapter 7 (RP4) contributes to this debate with lessons on the role that both public and private funding can play in supporting development, maintenance, and the sustainability of community-led OSAI projects. Through the analysis of the funding model of scikit-learn, it showed that commercial sponsors often bring industry insights and understanding of maintenance needs to such projects which OSS developers appreciate, while public funding can support development as well as ensure the independence of such projects without imposing commercial agendas. Moreover, it highlights the benefits of diversified funding model for balancing diverse interests in community-led projects and mitigating the disproportionate influence of any single funder on a project, be it a government, a company, or other.

The material dependencies of AI development underline that only funding the development of OSS will not suffice to support and sustain the development of large OMs. Indeed, as the Public

AI Network contends, public investments will be necessary at each layer of the AI stack (Public.AI, 2024). The BigScience Project offers an example of how public funding for compute power has facilitated community-led OM development. Thanks to subsidies provided by the French government for access to a state-owned supercomputer, Jean-Z, this community-led project was able to develop BLOOM, a 176 billion parameter LLM that can generate text in 46 natural languages and 13 programming languages (Akiki et al., 2022). Moreover, governments from Singapore to Saudi Arabia have invested public funds in the development of open LLMs that are adapted to their languages and cultural norms. For example, SEA-LION by AI Singapore is a family of open LLMs that "better understands Southeast Asia diverse contexts, languages, and cultures" (SEA-LION.AI, 2024). Crucially, these examples demonstrate that, with the appropriate public support, it is indeed possible to develop open LLMs that are motivated by and serve wider societal benefits (Ding et al., 2023).

While it may be possible, we must also ask ourselves if the development of open LLMs with public funds is a sensible approach to advancing the public interest in OSAI. We have ample evidence that the current compute-intensive approach to AI development is not only extremely costly, but its supply chain causes significant harms to the environment and the rights of local communities (Valdivia, 2024). As I discuss in my concluding remarks, these facts ought to make us reconsider whether AI development of this kind, even if it is open source, is desirable or justifiable in the long-term.

8.3 Concluding Remarks: The Role of Open Source in Public AI

We are at a critical juncture in AI development. The current trajectory, characterised by industry concentrations and the "bigger-is-better" paradigm that prizes the development of large AI models at exorbitant economic and environmental costs, does not seem to be compatible with the public interest. It has cemented the narrow field of actors who can participate in the design, development, and governance of AI technologies. We may hear AI democratisation being trumpeted as an equalising or democratising force by corporate actors, but this thesis has shown that, rather than challenge the power structures of the AI industry, the trend has been that OSAI has provided an avenue for a handful of powerful companies to entrench the influence in AI R&D.

In light of this status quo, advocacy for alternative approaches to AI development that are better aligned with the public interest have been gaining momentum. Three proposals were published in the months prior to submitting this thesis. In August 2024, the Public AI Network published "Public AI: Infrastructure for the Common Good," outlining three essential features for AI development in the public interest: universal public access to key AI capabilities, public accountability in governance and development, and permanent public goods that resist corporate capture (Public.AI, 2024). In

September 2024, Marda et al. (2024, p.4) from the Mozilla Foundation published the "Public AI" manifesto, proposing the creation of "a robust ecosystem of initiatives that promote public goods, public orientation, and public use throughout every step of AI development and deployment." In October 2024, Varoquaux et al. (2024) proposed that we break away from the "bigger-is-better" paradigm altogether and instead focus on approaches that are more affordable, more environmentally friendly, and more conducive to democratic governance. For example, they propose that we focus on the development of smaller models that can run on widely-available hardware at moderate costs, which would not only reduce the environmental impact and barriers to participation, but also enable more diverse actors to shape how AI technologies are developed, used, and governed.

Open source can play an important role in advancing these pathways to public AI. While open source may not be in itself resistant to commercial dominance or corporate capture, it has promise as an approach to sharing, developing, and governing AI technologies. Indeed, open source can contribute to public AI in three vital ways: as a mode of open access, ensuring AI technologies remain freely available and adaptable; as a mode of collaborative development, enabling diverse stakeholders to contribute to and shape AI technologies; and as a mode of open governance, providing transparent and participatory mechanisms for decision-making about AI development. However, for open source to advance public interests in AI development, it cannot be confined to researchers and developers alone. As Marda et al. (2024, p.4) note, public AI "will need serious financial, community, and political backing to enable it to become a meaningful alternative to the closed, commercial ecosystem." Fortunately, there is a growing recognition of the roles and responsibilities of diverse stakeholders, from governments to companies, in nurturing and sustaining open source, as well as a growing evidence base of the role that public funding can play in supporting community-led OSAI initiatives, such as the scikit-learn project, the BigScience project, and the SEA-LION LLMs. Learning lessons from successful examples such as these, among others, will be crucial as we seek to build a more equitable and sustainable OSAI ecosystem that serves broader societal and public interests.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2016, March). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.* arXiv. Retrieved 2024-01-04, from http://arxiv.org/abs/1603.04467 (arXiv:1603.04467 [cs]) doi: 10.48550/arXiv.1603.04467

Abboud, L., Levingston, I., & Hammond, G. (2024, April). Mistral in talks to raise €500mn at €5bn valuation. *Financial Times*. Retrieved 2024-04-19, from https://www.ft.com/content/358fc1d8-8276-4e81-9ee6-e0391575d569

Aberbach, J. D., & Rockman, B. A. (2002). Conducting and Coding Elite Interviews. *PS, political science & politics*, *35*(4), 673–676. (Place: New York, USA Publisher: Cambridge University Press)

Acquisti, A., Agrawal, A., Athey, S., Autor, D., Bessen, J., Björkegren, D., ... Youssef, A. B. (2024, December). *Statement from Economists on the Importance of Open Source AI.* Retrieved from https://open.mozilla.org/economists/

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard business review*, *84*(4), 98.

Ahlawat, P., Boyne, J., Herz, D., Schmieg, F., & Stephan, M. (2021, April). *Why You Need an Open Source Software Strategy.* Retrieved 2023-04-04, from https://www.bcg.com/publications/2021/open-source-software-strategy-benefits

Ahmed, N., & Wahed, M. (2020, October). *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research.* arXiv. Retrieved 2024-07-05, from http://arxiv.org/abs/2010.15581 (arXiv:2010.15581 [cs]) doi: 10.48550/arXiv.2010.15581

Ahmed, N., Wahed, M., & Thompson, N. C. (2023, March). The growing influence of industry in AI research. *Science*, *379*(6635), 884–886. Retrieved 2023-03-09, from https://www.science.org/doi/10.1126/science.ade2420 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.ade2420

Ait, A., Izquierdo, J. L. C., & Cabot, J. (2023a, March). HFCommunity: A Tool to Analyze the Hugging Face Hub Community. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 728–732). Retrieved 2024-04-05, from https://ieeexplore.ieee.org/abstract/document/10123660?casa_token=YtxaFDwLH1kAAAAA:PiQ2oxCA6fSVp65VQf77TC1905pV92hQjQceVhxNfZtnsjyUc6X6p9kPsCA4qefgr8vojaAm9pU (ISSN: 2640-7574) doi: 10.1109/SANER56733.2023.00080

Ait, A., Izquierdo, J. L. C., & Cabot, J. (2023b, July). *On the Suitability of Hugging Face Hub for Empirical Studies.* arXiv. Retrieved 2024-04-05, from http://arxiv.org/abs/2307.14841 (arXiv:2307.14841 [cs]) doi: 10.48550/arXiv.2307.14841

Akiki, C., Pistilli, G., Mieskes, M., Gallé, M., Wolf, T., Ilić, S., & Jernite, Y. (2022, December). *BigScience: A Case Study in the Social Construction of a Multilingual Large Language Model.* arXiv. Retrieved 2023-10-06, from http://arxiv.org/abs/2212.04960 (arXiv:2212.04960 [cs]) doi: 10.48550/arXiv.2212.04960

Alrawashdeh, T. A., Elbes, M. W., Almomani, A., ElQirem, F., & Tamimi, A. (2020). User acceptance model of open source software: An integrated model of OSS characteristics and UTAUT. *Journal of Ambient Intelligence and Humanized Computing*, *11*, 3315–3327. (ISBN: 1868-5137

Publisher: Springer)

Amreen, S., Mockus, A., Zaretzki, R., Bogart, C., & Zhang, Y. (2020). ALFAA: Active Learning Fingerprint based Anti-Aliasing for correcting developer identity errors in version control systems. *Empirical software engineering : an international journal*, *25*(2), 1136–1167. (Place: New York Publisher: Springer US)

arXiv. (2024). *arXiv.org e-Print archive.* Retrieved 2024-04-19, from https://arxiv.org/

Ashcraft, C., McLain, B., & Eger, E. (2016). Women in Tech: The Facts. *National Centre for Women & IT*, 77.

Atkinson, P., & Delamont, S. (2010). Collecting Data from Elites and Ultra Elites: Telephone and Face-to-Face Interviews with Macroeconomists. *Qualitative research : QR*, *7*(2), 203–216. (ISBN: 9781849203784 Place: United Kingdom Publisher: SAGE Publications, Limited)

Bagozzi, R. P., & Dholakia, U. M. (2006). Open source software user communities: A study of participation in linux user groups: Open source software. *Management science*, *52*(7), 1099–1115. (Place: Linthicum, MD Publisher: Institute for Operations Research and the Management Sciences)

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512. (Publisher: American Association for the Advancement of Science)

Barr, A. (2023). The tech world is being divided into 'GPU rich' and 'GPU poor.' Here are the companies in each group. *Business Insider*. Retrieved from https://www.businessinsider.com/gpurich-vs-gpu-poor-tech-companies-ineach-group-2023-8

Basdevant, A., François, C., Storchan, V., Bankston, K., Bdeir, A., Behlendorf, B., ... Tunney, J. (2024, May). *Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence.* arXiv. Retrieved 2024-08-26, from http://arxiv.org/abs/2405.15802 (arXiv:2405.15802 [cs]) doi: 10.48550/arXiv.2405.15802

Basole, R. C. (2009). Visualization of interfirm relations in a converging mobile ecosystem. *Journal of Information Technology, suppl. Special Issue on Mobile Computing*, *24*(2), 144–159. Retrieved 2022-02-11, from https://www.proquest.com/docview/216195909/abstract/A28E71F8D2864653PQ/1 (Num Pages: 16 Place: London, United Kingdom Publisher: Sage Publications Ltd.) doi: http://dx.doi.org/10.1057/jit.2008.34

Batagelj, V., & Zaversnik, M. (2003, October). *An O(m) Algorithm for Cores Decomposition of Networks.* arXiv. Retrieved 2023-10-05, from http://arxiv.org/abs/cs/0310049 (arXiv:cs/0310049) doi: 10.48550/arXiv.cs/0310049

Bdeir, A., & François, C. (2024, March). *Introducing the Columbia Convening on Openness and AI.* Retrieved 2024-03-25, from https://blog.mozilla.org/en/mozilla/ai/introducing-columbia-convening-openness-and-ai/

Beitin, B. K. (2012a). Interview and sampling. *The SAGE handbook of interview research: The complexity of the craft*, 243–253. Retrieved from https://scholar.google.fr/scholar?hl=en&as_sdt=0%2C5&q=Beitin%2C+2012&btnG=#d=gs_cit&t=1710989831122&u=%2Fscholar%3Fq%3Dinfo%3Aa6CFhQu0tZIJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den (Publisher: Sage London)

Beitin, B. K. (2012b, February). Interview and sampling: How many and whom. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE Handbook of Interview Research: The Complexity of the Craft.* SAGE Publications. (Google-Books-ID: _hp1AwAAQBAJ)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-05-16, from https://doi.org/10.1145/3442188.3445922 doi: 10.1145/3442188.3445922

Bengtsson, M., & Kock, S. (2000). "Coopetition" in Business Networks—to Cooperate and Compete Simultaneously. *Industrial marketing management*. doi: 10.1016/S0019-8501(99)00067-X

Benkler, Y. (2002). Coase's Penguin, or, Linux and "The Nature of the Firm". *The Yale Law journal*,

112(3), 369–446. (Place: New Haven Publisher: The Yale Law Journal Company) doi: 10.2307/1562247

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven [Conn.: Yale University Press. Retrieved 2020-09-13, from https://hdl.handle.net/10535/7396

Bergquist, M., & Ljungberg, J. (2001). The power of gifts: organizing social relationships in open source communities. *Information systems journal (Oxford, England)*, *11*(4), 305–320. (Place: Oxford, UK Publisher: Blackwell Science Ltd) doi: 10.1046/j.1365-2575.2001.00111.x

Betker, J. (2023, June). *The "it" in AI models is the dataset. – Non_interactive – Software & ML.* Retrieved 2024-12-13, from https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/

Bezroukov, N. (1999). Open source software as a special type of academic research: Critique of the vulgar Raymondism. *First Monday*. Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/696

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006). Mining email social networks. In *International Conference on Software Engineering: Proceedings of the 2006 international workshop on Mining software repositories; 22-23 May 2006* (pp. 137–143). ACM.

Birkinbine, B. (2015, February). Conflict in the Commons: Towards a Political Economy of Corporate Involvement in Free and Open Source Software. *The Political Economy of Communication*, *2*(2). Retrieved 2021-10-17, from http://www.polecom.org/index.php/polecom/article/view/35 (Number: 2)

Birkinbine, B. (2018). Commons Praxis: Toward a Critical Political Economy of the Digital Commons. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*. Retrieved 2021-11-15, from https://www.academia.edu/50200120/Commons_Praxis_Toward_a_Critical_Political_Economy_of_the_Digital_Commons

Birkinbine, B. (2020). *Incorporating the Digital Commons: Corporate Involvement in Free and Open Source Software*. University of Westminster Press. Retrieved 2021-10-01, from https://uwestminsterpress.co.uk/site/books/10.16997/book39/ doi: 10.16997/book39

Bitzer, J., Schrettl, W., & Schröder, P. J. (2007). Intrinsic motivation in open source software development. *Journal of Comparative Economics*, *35*(1), 160–169. (Place: San Diego Publisher: Elsevier Inc) doi: 10.1016/j.jce.2006.10.001

Bitzer, J., & Schröder, P. J. H. (2006). The Economics of Open Source Software Development: An Introduction. In *The Economics of Open Source Software Development*. United Kingdom: Emerald Publishing Limited. doi: 10.1016/B978-044452769-1/50001-9

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., ... Weinbach, S. (2022, April). *GPT-NeoX-20B: An Open-Source Autoregressive Language Model*. arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/2204.06745 (arXiv:2204.06745 [cs]) doi: 10.48550/arXiv.2204.06745

Blind, K., Pätsch, S., Muto, S., Böhm, M., Schubert, T., Grzegorzewska, P., & Katz, A. (2021). *The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy* (Tech. Rep.). Brussels, Belgium: European Commission. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/29effe73-2c2c-11ec-bd8e-01aa75ed71a1/language-en

Bobba, S., Carrara, S., Huisman, J., Mathieux, F., & Pavel, C. (2020). *Critical raw materials for strategic technologies and sectors in the EU. A Foresight Study* (Tech. Rep.). Luxembourg, Luxembourg. Retrieved from https://rmis.jrc.ec.europa.eu/uploads/CRMs_for_Strategic_Technologies_and_Sectors_in_the_EU_2020.pdf (Publisher: Publications Office of the European Union)

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2022, July). *On the Opportunities and Risks of Foundation Models*. arXiv. Retrieved 2024-03-12, from http://arxiv.org/abs/2108.07258 (arXiv:2108.07258 [cs]) doi: 10.48550/arXiv.2108

.07258

Bonaccorsi, A., & Rossi, C. (2006). Comparing motivations of individual programmers and firms to take part in the open source movement: From community to business. *Knowledge, Technology & Policy*, 25. Retrieved from https://link.springer.com/article/10.1007/s12130-006-1003-9

Bonaccorsi, A., & Rossi, C. L. (2003, September). *Altruistic Individuals, Selfish Firms? The Structure of Motivation in Open Source Software* (SSRN Scholarly Paper No. ID 433620). Rochester, NY: Social Science Research Network. Retrieved 2022-01-09, from https://papers.ssrn.com/abstract=433620 doi: 10.2139/ssrn.433620

Bornstein, M., Appenzeller, G., & Casado, M. (2023, January). *Who Owns the Generative AI Platform?* Retrieved 2024-11-29, from https://a16z.com/who-owns-the-generative-ai-platform/

Bosu, A., & Sultana, K. Z. (2019, September). Diversity and Inclusion in Open Source Software (OSS) Projects: Where Do We Stand? In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (pp. 1–11). Retrieved 2024-04-27, from https://ieeexplore.ieee.org/abstract/document/8870179?casa_token=XMvd4JpI6S8AAAAA:EB5n0Y5F1AcA4N6r-oOub6ERbunUqGTZeaEwy8wlSWKTkhRp7LVaHptNg79dgPCB5KUnmlwHnDI (ISSN: 1949-3789) doi: 10.1109/ESEM.2019.8870179

Braesemann, F., Stoehr, N., & Graham, M. (2019, January). Global networks in collaborative programming. *Regional Studies, Regional Science*, *6*(1), 371–373. Retrieved 2020-04-02, from https://doi.org/10.1080/21681376.2019.1588155 doi: 10.1080/21681376.2019.1588155

Brandenburger, A. M., & Nalebuff, B. (1997). *Co-Opetition: A Revolution Mindset That Combines Competition and Cooperation*. Doubleday. Retrieved 2021-10-20, from https://silo.pub/co-opetition-a-revolution-mindset-that-combines-competition-and-cooperation-the-game-theory-strategy-thats-changing-the-game-of-business.html

Brandenburger, A. M., & Nalebuff, B. (2021). The rules of co-opetition. *Harvard Business Review*, *99*(1), 48–57. Retrieved from https://spinup-000d1a-wp-offload-media.s3.amazonaws.com/faculty/wp-content/uploads/sites/8/2020/12/The-Rules-of-Co-opetition.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101. (Place: London Publisher: Taylor & Francis Group) doi: 10.1191/1478088706qp063oa

Brinkmann, S., & Kvale, S. (2015). *InterViews: Learning the craft of qualitative research interviewing* (3rd ed.). London, UK: SAGE.

Broca, S. (2013). Utopie du logiciel libre. Du bricolage informatique à la réinvention sociale. *Lectures*. Retrieved 2022-05-16, from https://journals.openedition.org/lectures/14019 (Publisher: Liens Socio)

Broca, S. (2021, June). Communs et capitalisme numérique : histoire d'un antagonisme et de quelques affinités électives. *Terminal. Technologie de l'information, culture & société*(130). Retrieved 2022-02-14, from https://journals.openedition.org/terminal/7595 doi: 10.4000/terminal.7595

Brooke, S. (2019, August). "Condescending, Rude, Assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 172–180). Florence, Italy: Association for Computational Linguistics. Retrieved 2020-09-13, from https://www.aclweb.org/anthology/W19-3519 doi: 10.18653/v1/W19-3519

Brooke, S. (2021, October). Trouble in programmer's paradise: gender-biases in sharing and recognising technical knowledge on Stack Overflow. *Information, Communication & Society*, *24*(14), 2091–2112. Retrieved 2021-12-20, from https://doi.org/10.1080/1369118X.2021.1962943 (Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2021.1962943) doi: 10.1080/1369118X.2021.1962943

Bryman, A. (2006). Integrating quantitative and qualitative research: how is it done? *Qualitative research : QR*, *6*(1), 97–113. (Place: Thousand Oaks, CA Publisher: Sage Publications) doi:

10.1177/1468794106058877

Buolamwini, J., & Gebru, T. (2018, January). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR. Retrieved 2022-05-29, from https://proceedings.mlr.press/v81/buolamwini18a.html (ISSN: 2640-3498)

Burgess, R. G. (1984). *In the field: an introduction to field research* (No. 8). London ;: Allen & Unwin. doi: 10.4324/9780203418161

Burwell, F., & Propp, K. (2022, November). *Digital sovereignty in practice: The EU's push to shape the new global economy* (Tech. Rep.). Retrieved 2023-04-17, from https://www.atlanticcouncil.org/in-depth-research-reports/report/digital-sovereignty-in-practice-the-eus-push-to-shape-the-new-global-economy/

Butler, S., Gamalielsson, J., Lundell, B., Brax, C., Sjöberg, J., Mattsson, A., ... Lönroth, E. (2021, July). On Company Contributions to Community Open Source Software Projects. *IEEE Transactions on Software Engineering*, *47*(7), 1381–1401. (Conference Name: IEEE Transactions on Software Engineering) doi: 10.1109/TSE.2019.2919305

Butler, S., Gamalielsson, J., Lundell, B., Jonsson, P., Sjoberg, J., Mattsson, A., ... Lonroth, E. (2018). An Investigation of Work Practices Used by Companies Making Contributions to Established OSS Projects. In *2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP)* (pp. 201–210). ACM.

Calimaq. (2018, June). *Les Communs numériques sont-il condamnés à devenir des « Communs du capital » ?* Retrieved 2022-01-25, from https://scinfolex.com/2018/06/24/les-communs-numeriques-sont-il-condamnes-a-devenir-des-communs-du-capital/

Candela, J. Q. (2017, September). *Facebook and Microsoft introduce new open ecosystem for interchangeable AI frameworks - Meta Research | Meta Research.* Retrieved 2024-01-07, from https://research.facebook.com/blog/2017/9/facebook-and-microsoft-introduce-new-open-ecosystem-for-interchangeable-ai-frameworks/

Cao, X., & Chesbrough, H. W. (2022). *OSS research in organizational management: A literature review and critical reappraisal.* Harwood Center for Corporate Innovation, UC Berkeley.

Cap, C. M. (2023a). *Alphabet (Google) - Market capitalization.* Retrieved 2023-12-26, from https://companiesmarketcap.com/alphabet-google/marketcap/

Cap, C. M. (2023b). *Meta Platforms (META) - Market capitalization.* Retrieved 2023-12-26, from https://companiesmarketcap.com/meta-platforms/marketcap/

Carlo, A. (2021). Artificial Intelligence in the Defence Sector. In J. Mazal, A. Fagiolini, P. Vasik, & M. Turi (Eds.), *Modelling and Simulation for Autonomous Systems* (pp. 269–278). Cham: Springer International Publishing. doi: 10.1007/978-3-030-70740-8_17

Casari, A., Ferraioli, J., & Lovato, J. (2023, May). Beyond the Repository: Best practices for open source ecosystems researchers. *Queue*, *21*(2), Pages 30:14–Pages 30:34. Retrieved 2024-10-03, from https://dl.acm.org/doi/10.1145/3595879 doi: 10.1145/3595879

Castaño, J., Martínez-Fernández, S., Franch, X., & Bogner, J. (2024, February). *Analyzing the Evolution and Maintenance of ML Models on Hugging Face.* arXiv. Retrieved 2024-04-05, from http://arxiv.org/abs/2311.13380 (arXiv:2311.13380 [cs]) doi: 10.48550/arXiv.2311.13380

CERN. (2023). *The birth of the Web | CERN.* Retrieved 2023-10-16, from https://home.cern/science/computing/birth-web

Champion, K., & Hill, B. M. (2021, March). Underproduction: An Approach for Measuring Risk in Open Source Software. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 388–399). Retrieved 2024-10-31, from https://ieeexplore.ieee.org/abstract/document/9426043?casa_token=NM7KgPQYBRMAAAAA:0OOUis3khswZ1AIbHejCBmfsfZnoR7Xe9YWkbIjT52r-4MNtchS3W-loa5Fm5uE_hO9fGTqV-Q (ISSN: 1534-5351) doi: 10.1109/SANER50967.2021.00043

Chan, C. K. Y., & Hu, W. (2023, July). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher*

Education, *20*(1), 43. Retrieved 2024-08-29, from https://doi.org/10.1186/s41239-023-00411-8 doi: 10.1186/s41239-023-00411-8

CHAOSS. (2023, June). *Metric: Bus Factor.* Retrieved 2024-01-15, from https://chaoss.community/?p=3944

CHAOSS. (2024). *Community Health Analytics in Open Source Software.* Retrieved 2024-05-01, from https://chaoss.community/

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE.

Chatterjee, M., & Volpicelli, G. (2023, August). *France bets big on open-source AI.* Retrieved 2023-08-09, from https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264–75278. Retrieved 2024-08-29, from https://ieeexplore.ieee.org/abstract/document/9069875 (Conference Name: IEEE Access) doi: 10.1109/ACCESS.2020.2988510

Chen, M.-J., & Miller, D. (2015). Reconceptualizing competitive dynamics: A multidimensional framework. *Strategic management journal*, *36*(5), 758–775. (Publisher: Wiley Online Library)

Chesbrough, H. (2006). *Open business models: How to thrive in the new innovation landscape*. Harvard Business Press.

Chesbrough, H. (2023, March). *Measuring the Economic Value of Open Source* (Tech. Rep.). San Francisco, CA, USA: Linux Foundation. Retrieved 2023-03-06, from https://www.linuxfoundation.org/research/measuring-economic-value-of-os

Cihon, P. (2024, April). *Helping policymakers weigh the benefits of open source AI.* Retrieved 2024-04-12, from https://github.blog/2024-04-10-helping-policymakers-weigh-the-benefits-of-open-source-ai/

Clauset, A., Newman, M. E. J., & Moore, C. (2004, December). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111. Retrieved 2023-10-05, from http://arxiv.org/abs/cond-mat/0408187 (arXiv:cond-mat/0408187) doi: 10.1103/PhysRevE.70.066111

Clegg, N. (2023, July). Nick Clegg: Openness on AI is the way forward for tech. *Financial Times*. Retrieved 2023-10-08, from https://www.ft.com/content/ac3b585a-ce50-43d1-b71d-14dfe6dce999

CMA. (2024, April). *AI Foundation Models: Update paper* (Tech. Rep.). London, UK: Competition and Markets Authority. Retrieved 2024-04-12, from https://www.gov.uk/government/publications/ai-foundation-models-update-paper

Coase, R. H. (1995). The Nature of the Firm. In S. Estrin & A. Marin (Eds.), *Essential Readings in Economics* (pp. 37–54). London: Macmillan Education UK. Retrieved 2024-08-30, from https://doi.org/10.1007/978-1-349-24002-9_3 doi: 10.1007/978-1-349-24002-9_3

Coleman, E. G. (2012). *Coding Freedom: The Ethics and Aesthetics of Hacking [electronic resource]*. Princeton, NJ: Princeton University Press. Retrieved 2022-05-16, from https://ezproxy-prd.bodleian.ox.ac.uk/login?url=https://doi.org/10.1515/9781400845293

Commission, E. (2020, October). *The European Commission adopts its new Open Source Software Strategy 2020-2023* (Tech. Rep.). Brussels, Belgium: Directorate-General for Informatics, European Commission. Retrieved 2023-04-13, from https://commission.europa.eu/news/european-commission-adopts-its-new-open-source-software-strategy-2020-2023-2020-10-21_en

CommonCrawl. (2024, May). *Common Crawl - Open Repository of Web Crawl Data.* Retrieved 2024-05-01, from https://commoncrawl.org/

Conti, J. A., & O'Neil, M. (2007). Studying power: qualitative methods and the global elite. *Qualitative research : QR*, *7*(1), 63–82. (Place: Thousand Oaks, CA Publisher: Sage Publications)

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to Algorithms* (4th ed.). MIT Press. (Google-Books-ID: RSMuEAAAQBAJ)

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, UNITED STATES: Yale University Press. Retrieved 2021-10-08, from http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=6478659

Creswell, J. W. (2018). *Research design: qualitative, quantitative & mixed methods approaches* (5th edition.; International student edition. ed.). Los Angeles: SAGE.

Crowston, K., Annabi, H., Howison, J., & Masango, C. (2005). Effective Work Practices for FLOSS Development: A Model and Propositions. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 197a–197a). IEEE. (ISSN: 1530-1605)

Crowston, K., & Howison, J. (2005). The social structure of Open Source Software development teams. *First Monday*, *10*(2), 22. Retrieved from https://crowston.syr.edu/sites/crowston.syr.edu/files/icis2003sna.pdf

Crowston, K., & Howison, J. (2006, December). Hierarchy and centralization in free and open source software team communications. *Knowledge, Technology & Policy*, *18*(4), 65–85. Retrieved 2022-03-18, from http://link.springer.com/10.1007/s12130-006-1004-8 doi: 10.1007/s12130-006-1004-8

Crowston, K., & Scozzi, B. (2002). Open source software projects as virtual organisations: competency rallying for software development. *IEE Proceedings-Software*, *149*(1), 3–17. (Publisher: IET)

Crowston, K., Wei, K., Howison, J., & Wiggins, A. (2012). Free/Libre open-source software development: What we know and what we do not know. *ACM computing surveys*, *44*(2), 1–35. (Place: New York, NY Publisher: ACM) doi: 10.1145/2089125.2089127

Cruzes, D. S., & Dybå, T. (2011, September). Recommended Steps for Thematic Synthesis in Software Engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement* (pp. 275–284). (ISSN: 1949-3789) doi: 10.1109/ESEM.2011.36

Dagnino, G. B. (2009). Coopetition strategy: A new kind of interfirm dynamics for value creation. In *Coopetition Strategy*. Routledge. (Num Pages: 19)

Dahlander, L., & Magnusson, M. (2008, December). How do Firms Make Use of Open Source Communities? *Long Range Planning*, *41*(6), 629–649. Retrieved 2023-04-12, from https://www.sciencedirect.com/science/article/pii/S0024630108000836 doi: 10.1016/j.lrp.2008.09.003

Dahlander, L., & Magnusson, M. G. (2005, May). Relationships between open source software companies and communities: Observations from Nordic firms. *Research Policy*, *34*(4), 481–493. Retrieved 2022-03-18, from https://linkinghub.elsevier.com/retrieve/pii/S0048733305000405 doi: 10.1016/j.respol.2005.02.003

Dahlander, L., & Wallin, M. W. (2006, October). A man on the inside: Unlocking communities as complementary assets. *Research Policy*, *35*(8), 1243–1259. Retrieved 2022-03-18, from https://linkinghub.elsevier.com/retrieve/pii/S0048733306001387 doi: 10.1016/j.respol.2006.09.011

Daigle, K. (2023, November). *Octoverse: The state of open source and rise of AI in 2023.* Retrieved 2024-05-06, from https://github.blog/2023-11-08-the-state-of-open-source-and-ai/

David, C., & Paul, J. (2023, March). *ChatGPT and large language models: what's the risk?* Retrieved 2023-08-11, from https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk

David, P., & Shapiro, J. (2008, December). Community-based production of open-source software: What do we know about the developers who participate? *Information Economics and Policy*, *20*(4), 364–398. Retrieved 2022-01-27, from https://ezproxy-prd.bodleian.ox.ac.uk:2056/science/article/pii/S0167624508000553 (Publisher: North-Holland) doi: 10.1016/j.infoecopol.2008.10.001

David, P., Waterman, A., & Arora, S. (2003). *FLOSS-US: The Free/Libre/Open Source Software Survey for 2003* (Tech. Rep.). Stanford, CA, USA: Stanford Institute for Economic Policy Research.

Di Giacomo, D., Kudzmanaite, B., Deveny, V., Dussutour, C., & Shaikh, M. (2020). *Key success factors of sustainable open source communities* (Tech. Rep.). Brussels, Belgium: EU. Retrieved 2023-04-19, from https://joinup.ec.europa.eu/sites/default/files/inline-files/SC272_D05.01_Community_Success_Factors_vFINAL_3.pdf

Dillet, R. (2022, May). *Hugging Face nabs $100M to build the GitHub of machine learning.* Retrieved 2023-12-26, from https://techcrunch.com/2022/05/09/hugging-face-reaches-2-billion-valuation-to-build-the-github-of-machine-learning/

Ding, J., Akiki, C., Jernite, Y., Steele, A. L., & Popo, T. (2023, January). *Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives.* arXiv. Retrieved 2023-10-06, from http://arxiv.org/abs/2301.08488 (arXiv:2301.08488 [cs]) doi: 10.48550/arXiv.2301.08488

Dolan, M. (2023, October). *How open source foundations protect the licensing integrity of open source projects.* Retrieved 2024-09-23, from https://www.linuxfoundation.org/blog/how-open-source-foundations-protect-the-licensing-integrity-of-open-source-projects

Don-Yehiya, S., Burtenshaw, B., Astudillo, R. F., Osborne, C., Jaiswal, M., Kuo, T.-S., ... Choshen, L. (2024, September). *The Future of Open Human Feedback.* arXiv. Retrieved 2024-10-01, from http://arxiv.org/abs/2408.16961 (arXiv:2408.16961 [cs]) doi: 10.48550/arXiv.2408.16961

Drost, E. A. (2011). Validity and Reliability in Social Science Research. *Education Research and Perspectives*, *38*(1), 21.

Dulong de Rosnay, M., & Stalder, F. (2020, December). Digital commons. *Internet Policy Review*, *9*(4), 15 p. Retrieved 2024-09-01, from https://hal.science/hal-03080213 (Publisher: Alexander von Humboldt Institute for Internet and Society) doi: 10.14763/2020.4.1530

Dunbar-Hester, C. (2019). *Hacking Diversity: The Politics of Inclusion in Open Technology Cultures* (Vol. 19). Princeton: University Press.

During, B. (2006, July). Trouble in paradise: the open source project PyPy, EU-funding and agile practices. In *AGILE 2006 (AGILE'06)* (pp. 11 pp.–231). Retrieved 2023-11-13, from https://ieeexplore.ieee.org/abstract/document/1667583 doi: 10.1109/AGILE.2006.58

Easterbrook, S., Singer, J., Storey, M.-A., & Damian, D. (2008). Selecting Empirical Methods for Software Engineering Research. In F. Shull, J. Singer, & D. I. K. Sjøberg (Eds.), *Guide to Advanced Empirical Software Engineering* (pp. 285–311). London: Springer. Retrieved 2021-12-20, from https://doi.org/10.1007/978-1-84800-044-5_11 doi: 10.1007/978-1-84800-044-5_11

Economist, T. (2024, February). Why do Nvidia's chips dominate the AI market? *The Economist*. Retrieved 2024-12-02, from https://www.economist.com/the-economist-explains/2024/02/27/why-do-nvidias-chips-dominate-the-ai-market

Edwards, R., & Holland, J. (2013). How have qualitative interviews developed? In *What is Qualitative Interviewing?* (1st ed., pp. 11–28). London: Bloomsbury Academic. Retrieved 2021-10-07, from http://www.bloomsburycollections.com/book/what-is-qualitative-interviewing/ch2-how-have-qualitative-interviews-developed/

Eghbal, N. (2016, June). *Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure* (Tech. Rep.). Ford Foundation. Retrieved 2022-06-21, from https://www.fordfoundation.org/work/learning/research-reports/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/

Eghbal, N. (2017). *Where money meets open source.* O'Reilly. Retrieved from https://www.youtube.com/watch?v=bjAinwgvQqc

Eghbal, N. (2020). *Working in Public: The Making and Maintenance of Open Source Software*. San Francisco: Stripe Press.

Eghbal, N. (2022, August). *A handy guide to financial support for open source.* Retrieved 2023-05-04, from https://github.com/nayafia/lemonade-stand (original-date: 2016-06-16T01:45:55Z)

Eiras, F., Petrov, A., Vidgen, B., Schroeder, C., Pizzati, F., Elkins, K., ... Foerster, J. (2024, May). *Risks*

and Opportunities of Open-Source Generative AI. arXiv. Retrieved 2024-05-28, from http://arxiv.org/abs/2405.08597 (arXiv:2405.08597 [cs]) doi: 10.48550/arXiv.2405.08597

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management review*, *14*(4), 532–550. (Place: Ada, Ohio, etc Publisher: Academy of Management) doi: 10.2307/258557

EleutherAI. (2021). *EleutherAI Models*. Retrieved 2023-09-18, from https://www.eleuther.ai/releases

EU. (2024, June). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence*. Retrieved 2024-09-11, from https://eur-lex.europa.eu/eli/reg/2024/1689/oj (Doc ID: 32024R1689 Doc Sector: 3 Doc Title: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) Doc Type: R Usr_lan: en)

Evas, T., Sipinen, M., Ulbrich, M., Dalla Benetta, A., Sobolewski, M., & Nepelski, D. (2022). *AI Watch: Estimating AI Investments in the European Union* (Tech. Rep. No. EUR 31088). Luxembourg: Publications Office of the European Union. Retrieved 2023-03-09, from https://ai-watch.ec.europa.eu/publications/ai-watch-estimating-ai-investments-european-union_en

Eynon, R. (2022, February). *Digital Interviewing - Methods Option Course, MSc Social Science of the Internet*. Oxford, UK. Retrieved from https://www.oii.ox.ac.uk/study/msc-in-social-science-of-the-internet/

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999, August). On power-law relationships of the Internet topology. *ACM SIGCOMM Computer Communication Review*, *29*(4), 251–262. Retrieved 2024-04-24, from https://dl.acm.org/doi/10.1145/316194.316229 doi: 10.1145/316194.316229

Feller, J., & Fitzgerald, B. (2002). *Understanding Open Source Software development*. London: Pearson Education.

Ferraioli, J. (2022, October). *Bringing together a social model of open source*. Retrieved 2024-10-30, from https://leaddev.com/velocity/bringing-together-social-model-open-source

Fershtman, C., & Gandal, N. (2004, March). *The Determinants of Output Per Contributor in Open Source Projects: An Empirical Examination* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-04-06, from https://papers.ssrn.com/abstract=539783

Fielding, N. G. (2012). Triangulation and Mixed Methods Designs: Data Integration With New Research Technologies. *Journal of mixed methods research*, *6*(2), 124–136. (Place: Los Angeles, CA Publisher: SAGE Publications) doi: 10.1177/1558689812437101

Fink, M. (2003). *The business and economics of Linux and open source*. Prentice Hall Professional.

Finlay, L. (2002, April). "Outing" the Researcher: The Provenance, Process, and Practice of Reflexivity. *Qualitative Health Research*, *12*(4), 531–545. Retrieved 2024-03-21, from https://doi.org/10.1177/104973202129120052 (Publisher: SAGE Publications Inc) doi: 10.1177/104973202129120052

Finley, K. (2017). Diversity in Open Source Is Even Worse Than in Tech Overall. *Wired*. Retrieved 2021-12-20, from https://www.wired.com/2017/06/diversity-open-source-even-worse-tech-overall/ (Section: tags)

Fletcher, R. (2024). How many news websites block AI crawlers? Retrieved 2024-12-11, from https://ora.ox.ac.uk/objects/uuid:6b0653e7-4a3b-4448-b0bd-1bdbd55aa61e (Publisher: Reuters Institute for the Study of Journalism)

Ford, D., Harkins, A., & Parnin, C. (2017, October). Someone like me: How does peer parity influence participation of women on stack overflow? In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 239–243). (ISSN: 1943-6106) doi: 10.1109/VLHCC.2017.8103473

Foster, K. (2022, March). *PyTorch vs TensorFlow: Who has More Pre-trained Deep Learning Models?* Retrieved 2024-12-02, from https://hackernoon.com/pytorch-vs-tensorflow-who-has-more-pre-trained-deep-learning-models

Franck, E., & Jungwirth, C. (2002). Reconciling investors and donators - The governance structure of open source. *Working Papers*. Retrieved 2023-04-04, from https://ideas.repec.org//p/iso/wpaper/0008.html (Number: 0008 Publisher: University of Zurich, Institute for Strategy and Business Economics (ISU))

Franke, N., & Hippel, E. v. (2003). Satisfying heterogeneous user needs via innovation toolkits: the case of Apache security software. *Research policy*, *32*(7), 1199–1215. (Publisher: Elsevier B.V) doi: 10.1016/S0048-7333(03)00049-0

Freeman, L. C. (1978, January). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. Retrieved 2019-03-14, from http://linkinghub.elsevier.com/retrieve/pii/0378873378900217 doi: 10.1016/0378-8733(78)90021-7

Gabrys, J. (2013). *Digital Rubbish: A Natural History of Electronics*. University of Michigan Press. Retrieved 2024-11-24, from https://library.oapen.org/handle/20.500.12657/24019 (Accepted: 2019-11-09 03:00:32) doi: 10.3998/dcbooks.9380304.0001.001

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... Leahy, C. (2020, December). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/2101.00027 (arXiv:2101.00027 [cs]) doi: 10.48550/arXiv.2101.00027

Gawer, A., & Cusumano, M. A. (2002). *Platform Leadership*. USA: Harvard Business School Press.

Gawer, A., & Cusumano, M. A. (2014). Industry Platforms and Ecosystem Innovation. *Journal of Product Innovation Management*, *31*(3), 417–433. Retrieved 2024-11-26, from https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12105 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12105) doi: 10.1111/jpim.12105

Geiger, R. S., Howard, D., & Irani, L. (2021, April). The Labor of Maintaining and Scaling Free and Open-Source Software Projects. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 175:1–175:28. Retrieved 2023-10-13, from https://dl.acm.org/doi/10.1145/3449249 doi: 10.1145/3449249

Germonprez, M., Allen, J. P., Warner, B., Hill, J., & McClements, G. (2013, November). Open source communities of competitors. *Interactions*, *20*(6), 54–59. Retrieved 2022-03-17, from https://dl.acm.org/doi/10.1145/2527191 doi: 10.1145/2527191

Germonprez, M., Link, G. J., Lumbard, K., & Goggins, S. (2018, November). Eight Observations and 24 Research Questions About Open Source Projects: Illuminating New Realities. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 57:1–57:22. Retrieved 2024-01-11, from https://dl.acm.org/doi/10.1145/3274326 doi: 10.1145/3274326

Gerosa, M., Wiese, I., Trinkenreich, B., Link, G., Robles, G., Treude, C., ... Sarma, A. (2021, November). The Shifting Sands of Motivation: Revisiting What Drives Contributors in Open Source. In *Proceedings of the 43rd International Conference on Software Engineering* (pp. 1046–1058). Madrid, Spain: IEEE Press. Retrieved 2024-06-24, from https://doi.org/10.1109/ICSE43902.2021.00098 doi: 10.1109/ICSE43902.2021.00098

Ghosh, R. A. (1998, March). Interviews with Linus Torvalds: What motivates software developers. *First Monday*. Retrieved 2023-04-10, from https://firstmonday.org/ojs/index.php/fm/article/view/583 doi: 10.5210/fm.v3i2.583

Ghosh, R. A. (2007). Understanding Free Software Developers: Findings from the FLOSS Study. In J. Feller, B. Fitzgerald, S. Hissam, & K. Lakhani (Eds.), *Perspectives on Free and Open Source Software*. Cambridge, MA, USA: MIT Press.

Ghosh, R. A., Glott, R., Krieger, B., & Robles, G. (2002). Free/Libre and Open Source Software: Survey and Study. *International Institute of Infonomics*.

Ghosh, R. A., & Schmidt, J. P. (2006). Open source and open standards: A new Frontier for economic development? *UNU-MERIT Policy Brief*. (ISBN: 1814-8026 Publisher: United Nations

University)
Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & Van Den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research policy*, *37*(10), 1717–1731. (Publisher: Elsevier)
GitHub. (2017). *Open Source Survey*. Retrieved 2022-01-03, from https://opensourcesurvey.org/2017/
GitHub. (2022). *Build software better, together.* Retrieved 2022-04-19, from https://github.com
GitHub. (2023). *Sponsoring an open source contributor.* Retrieved 2023-05-05, from https://ghdocs-prod.azurewebsites.net/_next/data/n5Bdo88NSHaRKsrErpk-I/en/free-pro-team@latest/sponsors/sponsoring-open-source-contributors/sponsoring-an-open-source-contributor.json?versionId=free-pro-team%40latest&productId=sponsors&restPage=sponsoring-open-source-contributors&restPage=sponsoring-an-open-source-contributor
GitHub. (2024a, December). *Artificial Intelligence Repositories on GitHub.* Retrieved 2023-09-18, from https://github.com/search?q=ai%20OR%20artificial-intelligence&type=repositories
GitHub. (2024b, May). *GitHub Innovation Graph.* Retrieved 2024-05-06, from https://innovationgraph.github.com/
Gnyawali, D. R., & Park, B.-j. (2009). Co-opetition and Technological Innovation in Small and Medium-Sized Enterprises: A Multilevel Conceptual Model. *Journal of small business management*, *47*(3), 308–330. (Place: Malden, USA Publisher: Blackwell Publishing Inc) doi: 10.1111/j.1540-627X.2009.00273.x
Gnyawali, D. R., & Park, B.-J. R. (2011, June). Co-opetition between giants: Collaboration with competitors for technological innovation. *Research Policy*, *40*(5), 650–663. Retrieved 2023-12-07, from https://www.sciencedirect.com/science/article/pii/S0048733311000187 doi: 10.1016/j.respol.2011.01.009
Goeminne, M., & Mens, T. (2011). Evidence for the pareto principle in open source software activity. In *the Joint Porceedings of the 1st International workshop on Model Driven Software Maintenance and 5th International Workshop on Software Quality and Maintainability* (pp. 74–82). Citeseer. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=75780c99b5f30e13a7682b2900289cfff75807c4#page=78
Goeminne, M., & Mens, T. (2013). A comparison of identity merge algorithms for software repositories. *Science of computer programming*, *78*(8), 971–986. (Publisher: Elsevier B.V)
Goggins, S., Lumbard, K., & Germonprez, M. (2021, May). Open Source Community Health: Analytical Metrics and Their Corresponding Narratives. In *2021 IEEE/ACM 4th International Workshop on Software Health in Projects, Ecosystems and Communities (SoHeal)* (pp. 25–33). Retrieved 2024-04-27, from https://ieeexplore.ieee.org/abstract/document/9474775 doi: 10.1109/SoHeal52568.2021.00010
Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023, January). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/2301.04246 (arXiv:2301.04246 [cs]) doi: 10.48550/arXiv.2301.04246
Google. (2023). *TensorFlow.* Retrieved 2023-12-26, from https://www.tensorflow.org/overview
Gorman, G. E., Clayton, P. R., Shep, S. J., & Clayton, A. (2005). *Qualitative Research for the Information Professional: A Practical Handbook*. Facet Publishing. (Google-Books-ID: CNQqDgAAQBAJ)
Gorwa, R., & Veale, M. (2024, February). *Moderating Model Marketplaces: Platform Governance Puzzles for AI Intermediaries.* arXiv. Retrieved 2024-05-16, from http://arxiv.org/abs/2311.12573 (arXiv:2311.12573 [cs]) doi: 10.48550/arXiv.2311.12573
Graham, M. (2014, January). *The Knowledge Based Economy and Digital Divisions of Labour* (SSRN Scholarly Paper No. ID 2363880). Rochester, NY: Social Science Research Network. Retrieved 2020-04-13, from https://papers.ssrn.com/abstract=2363880

Graham, M. (2019). *Digital economies at global margins*. Cambridge, Mass: The MIT Press. (OCLC: 1113389059)

Gray, M. L., & Suri, S. (2019). *Ghost work: how to stop Silicon Valley from building a new global underclass* (International edition ed.). Boston: Houghton Mifflin Harcourt.

Green, L. (2000). *Economics of Open Source Software.* Retrieved from http://badtux.org/home/eric/editorial/economics.php

Gruman, G. (2020, July). *IT snapshot: Ethnic diversity in the tech industry.* Retrieved 2022-01-03, from https://www.computerworld.com/article/3567095/it-snapshot-ethnic-diversity-in-the-tech-industry.html

Gubrium, J. F., Holstein, J. A., Marvasti, A. B., & McKinney, K. D. (2012). *The SAGE Handbook of Interview Research: The Complexity of the Craft*. SAGE Publications. (Google-Books-ID: _hp1AwAAQBAJ)

Gulati, R. (1998). Alliances and networks. *Strategic management journal*, *19*(4), 293–317. (Publisher: Wiley Online Library)

Gulson, K. N., & Webb, P. T. (2021, September). Steering the mind share: technology companies, policy and AI research in universities. *Discourse: Studies in the Cultural Politics of Education*, 1–13. Retrieved 2021-10-01, from https://doi.org/10.1080/01596306.2021.1981828 doi: 10.1080/01596306.2021.1981828

Gururaja, S., Bertsch, A., Na, C., Widder, D., & Strubell, E. (2023, December). To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13310–13325). Singapore: Association for Computational Linguistics. Retrieved 2024-12-02, from https://aclanthology.org/2023.emnlp-main.822 doi: 10.18653/v1/2023.emnlp-main.822

Haddad, I. (2022, March). *Artificial Intelligence and Data in Open Source* (Tech. Rep.). Linux Foundation. Retrieved from https://8112310.fs1.hubspotusercontent-na1.net/hubfs/8112310/LF%20Research/Artificial%20Intelligence%20and%20Data%20in%20Open%20Source%20-%20Report.pdf

Hale, C. (2022, August). Google might actually be the best friend for open-source software right now. *TechRadar*. Retrieved 2023-05-05, from https://www.techradar.com/news/google-might-actually-be-the-best-friend-for-open-source-software-right-now

Hamel, G. (1991). Competition for competence and interpartner learning within international strategic alliances. *Strategic Management Journal*, *12*(S1), 83–103. Retrieved 2023-12-29, from https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.4250120908 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.4250120908) doi: 10.1002/smj.4250120908

Hammond, G. (2023, December). Big Tech outspends venture capital firms in AI investment frenzy. *Financial Times*. Retrieved 2024-11-24, from https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532

Hand, M., Hillyard, S., Pole, C., & Love, K. (2014). *Big Data?: Qualitative Approaches to Digital Research*. Bingley, UNITED KINGDOM: Emerald Publishing Limited. Retrieved 2022-01-17, from http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=1865241

Hann, I., Roberts, J., Slaughter, S., & Fielding, R. (2002). Economic Incentives for Participating in Open Source Software Projects. *Proceedings of the International Conference on Information Systems*.

Hardy, M. (2023, October). *Should we use open source licenses for ML/AI models?* Retrieved 2023-11-02, from https://opensource.org/deepdive/webinars/should-we-use-open-source-licenses-for-ml-ai-models/

Hars, A., & Ou, S. (2002). Working for Free? Motivations for Participating in Open-Source Projects. *International journal of electronic commerce*, *6*(3), 25–39. (Publisher: Routledge) doi: 10.1080/10864415.2002.11044241

Haruvy, E., Prasad, A., & Sethi, S. (2003, August). Harvesting Altruism in Open-Source Software Development. *Journal of Optimization Theory and Applications*, *118*(2), 381–416. Retrieved 2023-04-06, from https://doi.org/10.1023/A:1025455523489 doi: 10.1023/A:1025455523489

Harvey, W. S. (2011). Strategies for conducting elite interviews. *Qualitative research : QR*, *11*(4), 431–441. (Place: London, England Publisher: SAGE Publications)

Hawkins, R. E. (2004). The economics of open source software for a competitive firm: Why give it away for free? *Netnomics*, *6*(2), 103–117. (Place: Dordrecht Publisher: Springer Nature B.V) doi: 10.1007/s11066-004-2717-z

Hays, K. (2024, January). Zuck's GPU flex will cost Meta as much as $18 billion by the end of 2024. *Yahoo News*. Retrieved 2024-11-28, from https://ca.news.yahoo.com/zucks-gpu-flex-cost-meta-171750011.html

Hecker, F. (1999, February). Setting up shop: The business of open-source software. *IEEE Software*, *16*(1), 45–51. Retrieved 2022-02-10, from http://ieeexplore.ieee.org/document/744568/ doi: 10.1109/52.744568

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., ... Zilberman, N. (2024, March). *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*. arXiv. Retrieved 2024-11-24, from http://arxiv.org/abs/2403.08501 (arXiv:2403.08501 [cs]) doi: 10.48550/arXiv.2403.08501

Heltweg, P., & Riehle, D. (2023, December). A Systematic Analysis of Problems in Open Collaborative Data Engineering. *ACM Transactions on Social Computing*, *6*(3-4), 8:1–8:30. Retrieved 2024-05-30, from https://doi.org/10.1145/3629040 doi: 10.1145/3629040

Hemetsberger, A. (2002). Fostering Cooperation on the Internet: Social Exchange Processes in Innovative Virtual Consumer Communities. *ACR North American Advances*, *NA-29*. Retrieved 2023-04-07, from https://www.acrwebsite.org/volumes/8675/volumes/v29/NA-29/full

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research policy*, *35*(7), 953–969. (Place: Amsterdam Publisher: Elsevier B.V)

Herpig, S. (2023, May). *Fostering Open Source Software Security – Blueprint for a Government Cybersecurity Open Source Program Office* (Tech. Rep.). Berlin, Germany: Stiftung Neue Verantwortung. Retrieved from https://www.stiftung-nv.de/en/publication/fostering-open-source-software-security

Herriott, R. E., & Firestone, W. A. (1983). Multisite Qualitative Policy Research: Optimizing Description and Generalizability. *Educational researcher*, *12*(2), 14–19. (Publisher: American Educational Research Association) doi: 10.2307/1175416

Hertel, G., Niedner, S., & Herrmann, S. (2003, July). Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research Policy*, *32*(7), 1159–1177. Retrieved 2021-12-20, from https://linkinghub.elsevier.com/retrieve/pii/S0048733303000477 doi: 10.1016/S0048-7333(03)00047-7

Hoffmann, M., Boysel, S., Nagle, F., Peng, S., & Xu, K. (2024, October). *Generative AI and the Nature of Work* [SSRN Scholarly Paper]. Rochester, NY: Social Science Research Network. Retrieved 2024-11-28, from https://papers.ssrn.com/abstract=5007084 doi: 10.2139/ssrn.5007084

Hoffmann, M., Nagle, F., & Zhou, Y. (2024). *The Value of Open Source Software*. Retrieved 2024-07-25, from https://www.ssrn.com/abstract=4693148 doi: 10.2139/ssrn.4693148

Hoffmann, W. H. (2007). Strategies for managing a portfolio of alliances. *Strategic management journal*, *28*(8), 827–856. (Publisher: Wiley Online Library)

Hogan, B. (2022). *From Social Science to Data Science*. Los Angeles ; London: SAGE.

Hogan, B., Carrasco, J. A., & Wellman, B. (2007, May). Visualizing Personal Networks: Working with Participant-aided Sociograms. *Field Methods*, *19*(2), 116–144. Retrieved 2024-05-29, from https://doi.org/10.1177/1525822X06298589 (Publisher: SAGE Publications Inc) doi: 10.1177/1525822X06298589

Holmstrom, B. R., & Tirole, J. (1989, January). Chapter 2 The theory of the firm. In *Handbook of*

Industrial Organization (Vol. 1, pp. 61–133). Elsevier. Retrieved 2024-08-30, from https://www.sciencedirect.com/science/article/pii/S1573448X89010058 doi: 10.1016/S1573-448X(89)01005-8

Hossain, A. (2021). *Regional OSS Communities: The View From Dhaka, Bangladesh*. Retrieved from https://www.fordfoundation.org/media/6667/regional-foss-communities_final-report_ahossain-1.pdf

House, W. (2022, January). *Readout of White House Meeting on Software Security*. Retrieved 2023-11-13, from https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/13/readout-of-white-house-meeting-on-software-security/

Howison, J., & Herbsleb, J. D. (2011, March). Scientific software production: incentives and collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 513–522). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-04-25, from https://dl.acm.org/doi/10.1145/1958824.1958904 doi: 10.1145/1958824.1958904

Howison, J., & Herbsleb, J. D. (2013, February). Incentives and integration in scientific software production. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 459–470). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-07-04, from https://doi.org/10.1145/2441776.2441828 doi: 10.1145/2441776.2441828

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12). Retrieved from https://crowston.syr.edu/content/validity-issues-use-social-network-analysis-digital-trace-data doi: 10.5281/zenodo.913303

Howlett, M. (2021, January). Looking at the 'field' through a Zoom lens: Methodological reflections on conducting online research during a global pandemic. *Qualitative Research*. Retrieved 2022-01-17, from https://journals.sagepub.com/doi/10.1177/1468794120985691 doi: https://doi.org/10.1177/1468794120985691

HuggingFace. (2023a). *Hugging Face*. Retrieved 2023-10-23, from https://huggingface.co/huggingface

HuggingFace. (2023b). *Transformers*. Retrieved 2023-12-26, from https://huggingface.co/docs/transformers/index

HuggingFace. (2024a, April). *BigCode - Open and responsible development and use of LLMs for code*. Retrieved 2024-04-19, from https://www.bigcode-project.org/

HuggingFace. (2024b, April). *Datasets - Hugging Face Hub*. Retrieved 2024-04-19, from https://huggingface.co/datasets

HuggingFace. (2024c). *Evaluate - Hugging Face*. Retrieved 2024-04-19, from https://huggingface.co/docs/evaluate/index

HuggingFace. (2024d, January). *Hugging Face and Google partner for open AI collaboration*. Retrieved 2024-07-02, from https://huggingface.co/blog/gcp-partnership

HuggingFace. (2024e, April). *Hugging Face Hub*. Retrieved 2024-04-19, from https://huggingface.co/

HuggingFace. (2024f, April). *Meta LlaMa Models on the HF Hub*. Retrieved 2024-04-23, from https://huggingface.co/meta-llama

HuggingFace. (2024g, April). *Mistral AI Models on the HF Hub*. Retrieved 2024-04-23, from https://huggingface.co/mistralai

HuggingFace. (2024h, April). *Models - Hugging Face*. Retrieved 2024-04-19, from https://huggingface.co/models

HuggingFace. (2024i, April). *OpenAI Models on the HF Hub*. Retrieved 2024-04-23, from https://huggingface.co/openai

HuggingFace. (2024j, April). *Spaces - Hugging Face*. Retrieved 2024-04-19, from https://huggingface.co/spaces

Iansiti, M. (2004). Strategy as Ecology. *Harvard Business Review*.

ihmpavel. (2022, February). *free-email-domains-list/src/constants.ts at master · ihmpavel/free-email-domains-list*. Retrieved 2023-12-27, from https://github.com/ihmpavel/free-email-domains-list/blob/master/src/constants.ts

ImageNet. (2024, May). *ImageNet*. Retrieved 2024-05-01, from https://www.image-net.org/

Inria. (2020, January). *The 2019 Inria French Academy of Sciences Dassault Systèmes Innovation Prize: scikit-learn, a success story for machine learning free software*. Retrieved from https://www.inria.fr/en/2019-inria-french-academy-sciences-dassault-systemes-innovation-prize-scikit-learn-success-story

Insight, O. (2023, March). *OSS Insight - scikit-learn*. Retrieved 2023-03-16, from https://ossinsight.io/analyze/scikit-learn/scikit-learn

Jacobson, D., & Mustafa, N. (2019). Social Identity Map: A Reflexivity Tool for Practicing Explicit Positionality in Critical Qualitative Research. *International journal of qualitative methods*, *18*, 160940691987007–. (Place: Los Angeles, CA Publisher: SAGE Publications) doi: 10.1177/1609406919870075

James, N., & Busher, H. (2009). *Online Interviewing*. London: SAGE Publications Ltd. Retrieved from https://methods.sagepub.com/book/online-interviewing doi: 10.4135/9780857024503

Jansen, S., Brinkkemper, S., & Finkelstein, A. (2009). Business Network Management as a Survival Strategy: A Tale of Two Software Ecosystems. In *Proceedings of the first International Workshop on Software Ecosystems* (pp. 34–48). Retrieved 2022-03-18, from http://ceur-ws.org/Vol-505/iwseco09-5JansenBrinkkemperFinkelstein.pdf

Jernite, Y., Gallé, M., Sanh, V., Tan, S., Wolf, T., Ilic, S., & Mitchell, M. (2021, July). NLP needs to be open. 500+ researchers are trying to make it happen. *VentureBeat*. Retrieved 2022-05-05, from https://venturebeat.com/2021/07/14/nlp-needs-to-be-open-500-researchers-are-trying-to-make-it-happen/

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative science quarterly*, *24*(4), 602–611. (Place: Ithaca, N.Y Publisher: Cornell University Graduate School of Business and Public Administration) doi: 10.2307/2392366

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational researcher*, *33*(7), 14–26. (Place: Thousand Oaks, CA Publisher: American Educational Research Association) doi: 10.3102/0013189X033007014

Jokonya, O. (2015, January). Investigating Open Source Software Benefits in Public Sector. In *2015 48th Hawaii International Conference on System Sciences* (pp. 2242–2251). (ISSN: 1530-1605) doi: 10.1109/HICSS.2015.268

Kaggle. (2024). *Find Open Datasets and Machine Learning Projects | Kaggle*. Retrieved 2024-04-19, from https://www.kaggle.com/datasets

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... Narayanan, A. (2024, February). *On the Societal Impact of Open Foundation Models*. Retrieved from https://crfm.stanford.edu/open-fms/paper.pdf

Ke, W., & Zhang, P. (2008, July). Motivations for participating in open source software communities: Roles of Pyschological needs and altruism. *PACIS 2008 Proceedings*. Retrieved from https://aisel.aisnet.org/pacis2008/76

Keller, P. (2022, December). European Public Digital Infrastructure Fund White Paper. *Open Future*. Retrieved 2023-02-01, from https://openfuture.pubpub.org/pub/public-digital-infra-fund-whitepaper/release/1 (Publisher: Open Future Foundation)

Kelty, C. M. (2008). *Two bits: the cultural significance of free software*. Durham, N.C. ; London: Duke University Press.

Kendall, J. E., Kendall, K. E., & Germonprez, M. (2016, October). Game theory and open source contribution: Rationale behind corporate participation in open source software development. *Journal of Organizational Computing and Electronic Commerce*, *26*(4), 323–343. Retrieved 2024-02-25, from https://www.tandfonline.com/doi/full/10.1080/10919392.2016.1228360 doi:

10.1080/10919392.2016.1228360

King, N. (2009). *Interviews in qualitative research*. London: SAGE. (Book Title: Interviews in qualitative research)

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023, July). A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 17061–17084). PMLR. Retrieved 2024-02-29, from https://proceedings.mlr.press/v202/kirchenbauer23a.html (ISSN: 2640-3498)

Klinger, J., Mateos-Garcia, J. C., & Stathoulopoulos, K. (2020, September). *A Narrowing of AI Research?* (SSRN Scholarly Paper No. ID 3698698). Rochester, NY: Social Science Research Network. Retrieved 2021-10-01, from https://papers.ssrn.com/abstract=3698698 doi: 10.2139/ssrn.3698698

Kogut, B. (1988). Joint ventures: Theoretical and empirical perspectives. *Strategic Management Journal*, *9*(4), 319–332. Retrieved 2023-12-29, from https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.4250090403 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.4250090403) doi: 10.1002/smj.4250090403

Kollock, P. (1999). Economics of on line cooperation: Gift and public goods in cyberspace. In M. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (pp. 220–246). London, UK: Routledge.

Korkmaz, G., Santiago Calderón, J. B., Kramer, B. L., Guci, L., & Robbins, C. A. (2024, April). From GitHub to GDP: A framework for measuring open source software innovation. *Research Policy*, *53*(3), 104954. Retrieved 2024-08-21, from https://www.sciencedirect.com/science/article/pii/S0048733324000039 doi: 10.1016/j.respol.2024.104954

Kouters, E., Vasilescu, B., Serebrenik, A., & van den Brand, M. G. J. (2012). Who's who in Gnome: Using LSA to merge software repository identities. In *2012 28th IEEE International Conference on Software Maintenance (ICSM)* (pp. 592–595). IEEE. (ISSN: 1063-6773)

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020, February). Defining AI in Policy versus Practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 72–78). New York, NY, USA: Association for Computing Machinery. Retrieved 2020-07-26, from https://doi.org/10.1145/3375627.3375835 doi: 10.1145/3375627.3375835

Krishnamurthy, S. (2005a, May). An Analysis of Open Source Business Models. In *Perspectives on Free and Open Source Software*. The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/5326.003.0022 (_eprint: https://direct.mit.edu/book/chapter-pdf/2297464/9780262256124_cao.pdf) doi: 10.7551/mitpress/5326.003.0022

Krishnamurthy, S. (2005b, October). Cave or community? An empirical examination of 100 mature open source projects. *First Monday*. Retrieved 2022-05-06, from https://journals.uic.edu/ojs/index.php/fm/article/view/1477 doi: 10.5210/fm.v0i0.1477

Krishnamurthy, S. (2006, December). On the intrinsic and extrinsic motivation of free/libre/open source (FLOSS) developers. *Knowledge, Technology & Policy*, *18*(4), 17–39. Retrieved 2024-03-24, from https://doi.org/10.1007/s12130-006-1002-x doi: 10.1007/s12130-006-1002-x

Krishnamurthy, S., & Tripathi, A. K. (2009, March). Monetary donations to an open source software platform. *Research Policy*, *38*(2), 404–414. Retrieved 2023-04-07, from https://www.sciencedirect.com/science/article/pii/S0048733308002758 doi: 10.1016/j.respol.2008.11.004

Kuwabara, K. (2000, March). Linux: A bazaar at the edge of chaos. *First Monday*. Retrieved 2022-05-09, from https://firstmonday.org/ojs/index.php/fm/article/view/731 doi: 10.5210/fm.v5i3.731

Kvale, S. (1996). *InterViews: an introduction to qualitative research interviewing*. Thousand Oaks ; London: Sage.

Lakatos, S. (2023, December). *A Revealing Picture: AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business* (Tech. Rep.).

Retrieved 2024-02-29, from https://graphika.com/reports/a-revealing-picture
Lakhani, K., & von Hippel, E. (2003). How open source software works: "free" user-to-user assistance. *Research policy*, *32*(6), 923–943. (Place: Amsterdam Publisher: Elsevier B.V) doi: 10.1016/S0048-7333(02)00095-1
Lakhani, K. R., Wolf, B., Bates, J., & DiBona, C. (2002, July). *Release 0.73 In Cooperation with OSDN* (Tech. Rep.). Cambridge, MA, USA: The Boston Consulting Group.
Lakhani, K. R., & Wolf, R. G. (2003, September). *Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-04-04, from https://papers.ssrn.com/abstract=443040 doi: 10.2139/ssrn.443040
Lamont, M., & Swidler, A. (2014). Methodological Pluralism and the Possibilities and Limits of Interviewing. *Qualitative Sociology*, *37*(2), 153–171. Retrieved 2022-02-07, from https://www.proquest.com/docview/1519493837/abstract/88A8CF249B24825PQ/1 (Num Pages: 153-171 Place: New York, Netherlands Publisher: Springer Nature B.V.) doi: http://dx.doi.org/10.1007/s11133-014-9274-z
Langenkamp, M., & Yue, D. N. (2022, July). How Open Source Machine Learning Software Shapes AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 385–395). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-08-17, from https://doi.org/10.1145/3514094.3534167 doi: 10.1145/3514094.3534167
Lattemann, C., & Stieglitz, S. (2005, January). Framework for Governance in Open Source Communities. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 192a–192a). (ISSN: 1530-1605) doi: 10.1109/HICSS.2005.278
Laurila, J. (1997). Promoting research access and informant rapport in corporate settings: Notes from research on a crisis company. *Scandinavian journal of management*, *13*(4), 407–418. (Publisher: Elsevier Ltd)
Law, H., & Krier, S. (2023, August). Open-source provisions for large models in the AI Act. *Cambridge Journal of Science and Policy*, *4*. Retrieved 2023-08-09, from https://www.repository.cam.ac.uk/handle/1810/354175 (Publisher: Cambridge University Science and Policy Exchange)
Lawless, B., & Chen, Y.-W. (2019, January). Developing a Method of Critical Thematic Analysis for Qualitative Communication Inquiry. *Howard Journal of Communications*, *30*(1), 92–106. Retrieved 2022-04-07, from https://doi.org/10.1080/10646175.2018.1439423 (Publisher: Routledge _eprint: https://doi.org/10.1080/10646175.2018.1439423) doi: 10.1080/10646175.2018.1439423
LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved 2024-12-03, from https://www.nature.com/articles/nature14539 (Publisher: Nature Publishing Group) doi: 10.1038/nature14539
Lee, V., & Herstatt, C. (2015). How Firms Can Strategically Influence Open Source Communities: The Employment of 'Men on the Inside'. In *Open Source Innovation*. Routledge. (Num Pages: 35)
Lehdonvirta, V., Kässi, O., Hjorth, I., Barnard, H., & Graham, M. (2019, February). The Global Platform Economy: A New Offshoring Institution Enabling Emerging-Economy Microproviders. *Journal of Management*, *45*(2), 567–599. Retrieved 2022-06-30, from https://doi.org/10.1177/0149206318786781 (Publisher: SAGE Publications Inc) doi: 10.1177/0149206318786781
Lehdonvirta, V., Wu, B., & Hawkins, Z. (2023, December). *Cloud empires' physical footprint: How trade and security politics shape the global expansion of U.S. and Chinese data centre infrastructures* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2024-01-09, from https://papers.ssrn.com/abstract=4670764 doi: 10.2139/ssrn.4670764
Lehdonvirta, V., Wú, B., & Hawkins, Z. (2024, October). Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 828–838. Retrieved 2024-11-22, from

https://ojs.aaai.org/index.php/AIES/article/view/31683 doi: 10.1609/aies.v7i1.31683

Lerner, J., & Tirole, J. (2001, May). The open source movement: Key research questions. *European Economic Review*, *45*(4), 819–826. Retrieved 2022-01-09, from https://www.sciencedirect.com/science/article/pii/S0014292101001246 doi: 10.1016/S0014-2921(01)00124-6

Lerner, J., & Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal of Industrial Economics*, *50*(2), 197–234. Retrieved 2022-02-24, from https://onlinelibrary.wiley.com/doi/10.1111/1467-6451.00174 doi: 10.1111/1467-6451.00174

Lerner, J., & Tirole, J. (2005). The Economics of Technology Sharing: Open Source and Beyond. *The Journal of Economic Perspectives*, *19*(2), 99–120. Retrieved 2024-10-30, from https://www.jstor.org/stable/4134939 (Publisher: American Economic Association)

Lessig, L. (2003). *An Information Society: Free or Feudal?* World Summit on the Information Society. Retrieved from http://www.itu.int/net/wsis/docs/pc2/visionaries/lessig.pdf

Leswing, K. (2024, June). Nvidia dominates the AI chip market, but there's more competition than ever. *CNBC*. Retrieved 2024-12-02, from https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html (Section: Technology)

LFAI&Data. (2022). *Linux Foundation AI & Data Landscape.* Retrieved 2022-01-19, from https://landscape.lfai.foundation/

LFAI&Data. (2023). *LF AI & Data Foundation.* Retrieved 2023-06-10, from https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=327683

Li, H., Ajmani, L., Zhou, M., Vincent, N., Hwang, S., Piccardi, T., ... Veselovsky, V. (2022, November). Ethical Tensions, Norms, and Directions in the Extraction of Online Volunteer Work. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 273–277). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-07-04, from https://doi.org/10.1145/3500868.3560923 doi: 10.1145/3500868.3560923

Li, X., Zhang, Y., Osborne, C., Zhou, M., Jin, Z., & Liu, H. (2024, August). Systematic Literature Review of Commercial Participation in Open Source Software. *ACM Transactions on Software Engineering and Methodology*. Retrieved 2024-09-26, from https://dl.acm.org/doi/10.1145/3690632 (Just Accepted) doi: 10.1145/3690632

Liesenfeld, A., & Dingemanse, M. (2024, May). Rethinking open source generative AI: open-washing and the EU AI Act. ACM. Retrieved 2024-06-03, from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3588217

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023, July). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–6). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-08-18, from https://dl.acm.org/doi/10.1145/3571884.3604316 doi: 10.1145/3571884.3604316

Lin, B., Robles, G., & Serebrenik, A. (2017, May). Developer Turnover in Global, Industrial Open Source Projects: Insights from Applying Survival Analysis. In *2017 IEEE 12th International Conference on Global Software Engineering (ICGSE)* (pp. 66–75). Buenos Aires, Argentina: IEEE. Retrieved 2021-12-20, from http://ieeexplore.ieee.org/document/7976690/ doi: 10.1109/ICGSE.2017.11

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills ; London: Sage.

Lindman, J., Juutilainen, J.-P., & Rossi, M. (2009). Beyond the Business Model: Incentives for Organizations to Publish Software Source Code. In C. Boldyreff, K. Crowston, B. Lundell, & A. I. Wasserman (Eds.), *Open Source Ecosystems: Diverse Communities Interacting* (pp. 47–56). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-02032-2_6

Link, G., & Germonprez, M. (2016). Understanding open source communities as complex adaptive systems: A case of the R Project community. In *AMCIS 2016 Proceedings.* Retrieved 2024-02-25,

from https://core.ac.uk/download/pdf/301368981.pdf

Linåker, J., Link, G. J. P., & Lumbard, K. (2024, August). *Sustaining Maintenance Labor for Healthy Open Source Software Projects through Human Infrastructure: A Maintainer Perspective.* arXiv. Retrieved 2024-10-30, from http://arxiv.org/abs/2408.06723 (arXiv:2408.06723 [cs]) doi: 10.48550/arXiv.2408.06723

Linåker, J., Papatheocharous, E., & Olsson, T. (2022, September). How to characterize the health of an Open Source Software project? A snowball literature review of an emerging practice. In *Proceedings of the 18th International Symposium on Open Collaboration* (pp. 1–12). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-10-03, from https://dl.acm.org/doi/10.1145/3555051.3555067 doi: 10.1145/3555051.3555067

Linåker, J., Regnell, B., & Damian, D. (2020, March). A method for analyzing stakeholders' influence on an open source software ecosystem's requirements engineering process. *Requirements Engineering*, *25*(1), 115–130. Retrieved 2024-05-30, from https://doi.org/10.1007/s00766-019-00310-3 doi: 10.1007/s00766-019-00310-3

Linåker, J., Rempel, P., Regnell, B., & Mäder, P. (2016). How Firms Adapt and Interact in Open Source Ecosystems: Analyzing Stakeholder Influence and Collaboration Patterns. In M. Daneva & O. Pastor (Eds.), *Requirements Engineering: Foundation for Software Quality* (pp. 63–81). Cham: Springer International Publishing. doi: 10.1007/978-3-319-30282-9_5

Loebbecke, C., & Angehrn, A. A. (2003). Open Source Platforms Under Co-opetition: A Comparative Analysis of SourceForge and 'CodeX' (Xerox) as Two 'Co-opetitive Learning and Knowledge Exchange Networks' (CoLKENs). In *ECIS 2003 Proceedings.*

Long, Y., & Siau, K. (2007). Social Network Structures in Open Source Software Development Teams. *Journal of Database Management*, *18*(2), 25–33,35–40. Retrieved 2022-03-18, from https://www.proquest.com/docview/199603797/abstract/8D3CCF38B9474C8DPQ/1 (Num Pages: 15 Place: Hershey, United States Publisher: IGI Global) doi: http://dx.doi.org/10.4018/jdm.2007040102

Lopez-Fernandez, L. (2004). Applying social network analysis to the information in CVS repositories. In "*International Workshop on Mining Software Repositories (MSR 2004)" W17S Workshop - 26th International Conference on Software Engineering* (Vol. 2004, pp. 101–105). Edinburgh, Scotland, UK: IEE. Retrieved 2021-10-22, from https://digital-library.theiet.org/content/conferences/10.1049/ic_20040485 doi: 10.1049/ic:20040485

Luccioni, A. S., Jernite, Y., & Strubell, E. (2024, June). Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 85–99). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-08-29, from https://dl.acm.org/doi/10.1145/3630106.3658542 doi: 10.1145/3630106.3658542

Lumbard, K., & Germonprez, M. (2017). Open Source Communities as Liminal Ecosystems. In *MWAIS 2017* (p. 6). University of Illinois Springfield. Retrieved from http://aisel.aisnet.org/mwais2017/45

Lumbard, K., Germonprez, M., & Goggins, S. (2024). An empirical investigation of social comparison and open source community health. *Information Systems Journal*, *34*(2), 499–532. (Publisher: Wiley Online Library)

Luthiger, B., & Jungwirth, C. (2007, January). Pervasive fun. *First Monday*. Retrieved 2023-04-06, from https://firstmonday.org/ojs/index.php/fm/article/view/1422 doi: 10.5210/fm.v12i1.1422

Madey, G., Freeh, V., & Tynan, R. (2002). The Open Source Software Development Phenomenon: An Analysis based on Social Network Theory. In *AMCIS 2002 Proceedings.* Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1606&context=amcis2002

Maffulli, S. (2023, July). *Meta's LLaMa 2 license is not Open Source.* Retrieved 2023-08-11, from https://blog.opensource.org/metas-llama-2-license-is-not-open-source/

Mahanti, A., Carlsson, N., Mahanti, A., Arlitt, M., & Williamson, C. (2013, January). A tale of the

tails: Power-laws in internet measurements. *IEEE Network*, *27*(1), 59–64. Retrieved 2024-04-24, from https://ieeexplore.ieee.org/abstract/document/6423193 (Conference Name: IEEE Network) doi: 10.1109/MNET.2013.6423193

Marda, N., Sun, J., & Surman, M. (2024, September). *Public AI* (Tech. Rep.). Mozilla Foundation. Retrieved 2024-12-06, from https://foundation.mozilla.org/en/research/library/public-ai/

Margolis, J., & Fisher, A. (2001). *Unlocking the Clubhouse: Women in Computing*. Cambridge, MA, USA: MIT Press.

Markus, L., Manville, B., & Agres, C. (2000, October). What Makes a Virtual Organization Work: Lessons From the Open-Source World. *MIT Sloan Management Review*. Retrieved 2021-11-01, from https://shop.sloanreview.mit.edu/store/what-makes-a-virtual-organization-work-lessons-from-the-open-source-world

Marlow, J., & Dabbish, L. (2013). Activity traces and signals in software developer recruitment and hiring. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13* (p. 145). San Antonio, Texas, USA: ACM Press. Retrieved 2021-12-20, from http://dl.acm.org/citation.cfm?doid=2441776.2441794 doi: 10.1145/2441776.2441794

Marsan, J., Paré, G., & Beaudry, A. (2012, December). Adoption of open source software in organizations: A socio-cognitive perspective. *The Journal of Strategic Information Systems*, *21*(4), 257–273. Retrieved 2023-04-14, from https://www.sciencedirect.com/science/article/pii/S0963868712000200 doi: 10.1016/j.jsis.2012.05.004

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., ... Clark, J. (2024, April). *The AI Index 2024 Annual Report* (Tech. Rep.). Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Mateos-Garcia, J., Klinger, J., Stathoulopoulos, K., & Winch, R. (2019, November). *A Semantic Analysis of the Recent Evolution of AI Research* (Tech. Rep.). Retrieved from https://media.nesta.org.uk/documents/A_Semantic_Analysis_of_the_Recent_Evolution_of_AI_Research.pdf

McAuley, J. J., Costa, L. d. F., & Caetano, T. S. (2007, August). The rich-club phenomenon across complex network hierarchies. *Applied Physics Letters*, *91*(8), 084103. Retrieved 2023-10-05, from http://arxiv.org/abs/physics/0701290 (arXiv:physics/0701290) doi: 10.1063/1.2773951

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U Test. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). John Wiley & Sons, Ltd. Retrieved 2024-05-16, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470479216.corpsy0524) doi: 10.1002/9780470479216.corpsy0524

Medappa, P. K., Tunc, M. M., & Li, X. (2023, June). *Sponsorship Funding in Open-Source Software: Effort Reallocation and Spillover Effects in Knowledge-Sharing Ecosystems* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-11-13, from https://papers.ssrn.com/abstract=4484403 doi: 10.2139/ssrn.4484403

Meehan, E. R., Mosco, V., & Wasko, J. (1993). Rethinking Political Economy: Change and Continuity. *Journal of Communication*, *43*(4), 105–16.

Mehra, A. (2011). Firms as Incubators of Open-Source Software. *Information Systems Research*, *22*(1), 18. doi: 10.1287/isre.l090.0276

Menand, L. (1997). *Pragmatism: a reader* (1st ed. ed.). New York: Vintage Books. Retrieved 2022-04-05, from http://catdir.loc.gov/catdir/enhancements/fy0701/97009328-t.html

Meta. (2022). *Announcing the PyTorch Foundation: A new era for the cutting-edge AI framework.* Retrieved 2022-09-12, from https://ai.facebook.com/blog/pytorch-foundation/

Meta. (2023, July). *Meta and Microsoft Introduce the Next Generation of Llama.* Retrieved 2023-10-08, from https://about.fb.com/news/2023/07/llama-2/

Metz, C. (2015, November). Google Just Open Sourced the Artificial Intelligence Engine at the Heart

of Its Online Empire. *WIRED*. Retrieved 2020-04-05, from https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/

Microsoft. (2018). *Microsoft acquires GitHub*. Retrieved 2022-04-12, from https://news.microsoft.com/announcement/microsoft-acquires-github/

Mikecz, R. (2012). Interviewing Elites: Addressing Methodological Issues. *Qualitative Inquiry*, *18*(6), 482–493. Retrieved from https://doi.org/10.1177/1077800412442818 (_eprint: https://doi.org/10.1177/1077800412442818) doi: 10.1177/1077800412442818

Miller, C. (2022). *Chip War: The Fight for the World's Most Critical Technology*. Simon and Schuster.

Mitchell, T. M. (1997). *Machine learning* (Vol. 1) (No. 9). McGraw-hill New York.

Mockus, A., Fielding, R., & Herbsleb, J. (2002). Two case studies of open source software development: Apache and Mozilla. *ACM transactions on software engineering and methodology*, *11*(3), 309–346. (Publisher: ACM)

Molina, J., Maya-Jariego, I., & McCarty, C. (2014, January). Giving Meaning to Social Networks: Methodology for Conducting and Analyzing Interviews based on Personal Network Visualizations. In B. Hollstein & S. Dominguez (Eds.), *Mixed methods in social network research. Design and applications* (pp. 305–335). USA: Cambridge University Press.

Mosco, V. (2009). *The Political Economy of Communication* (2nd ed. ed.). London: Sage Publications. (Book Title: The political economy of communication)

Mozilla. (2023, March). *Introducing Mozilla.ai: Investing in trustworthy AI | The Mozilla Blog*. Retrieved 2023-10-30, from https://blog.mozilla.org/en/mozilla/introducing-mozilla-ai-investing-in-trustworthy-ai/

Mugrage, K. (2022, April). The changing economics of open source. *MIT Technology Review*. Retrieved 2023-04-04, from https://www.technologyreview.com/2022/04/21/1050788/the-changing-economics-of-open-source/

Muldoon, J., Graham, M., & Cant, C. (2024). *Feeding the Machine The Hidden Human Labour Powering AI*. Canongate Books. Retrieved 2024-08-29, from https://canongate.co.uk/books/5234-feeding-the-machine-the-hidden-human-labour-powering-ai/

Musser, M. (2023, August). *A Cost Analysis of Generative Language Models and Influence Operations*. arXiv. Retrieved 2024-02-29, from http://arxiv.org/abs/2308.03740 (arXiv:2308.03740 [cs]) doi: 10.48550/arXiv.2308.03740

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023, May). Auditing large language models: a three-layered approach. *AI and Ethics*. Retrieved 2024-08-30, from https://doi.org/10.1007/s43681-023-00289-2 doi: 10.1007/s43681-023-00289-2

Nagle, F. (2019, March). *Government Technology Policy, Social Value, and National Competitiveness* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-10-13, from https://papers.ssrn.com/abstract=3355486 doi: 10.2139/ssrn.3355486

Nagle, F., Wheeler, D. A., Lifshitz-Assaf, H., Ham, H., & Hoffman, J. L. (2020). *Report on the 2020 FOSS Contributor Survey* (Tech. Rep.). San Francisco, CA, USA: The Linux Foundation. Retrieved 2023-03-24, from https://www.linuxfoundation.org/resources/publications/foss-contributor-2020

Nakasai, K., Hata, H., & Matsumoto, K. (2019, May). Are Donation Badges Appealing?: A Case Study of Developer Responses to Eclipse Bug Reports. *IEEE Software*, *36*(3), 22–27. (Conference Name: IEEE Software) doi: 10.1109/MS.2018.2874568

Nakasai, K., Hata, H., Onoue, S., & Matsumoto, K. (2017, March). Analysis of Donations in the Eclipse Project. In *2017 8th International Workshop on Empirical Software Engineering in Practice (IWESEP)* (pp. 18–22). doi: 10.1109/IWESEP.2017.19

NetworkX. (2023a). *Betweenness Centrality*. Retrieved 2023-10-05, from https://networkx.org/documentation/stable/auto_examples/algorithms/plot_betweenness_centrality.html#betweenness-centrality

NetworkX. (2023b). *Density*. Retrieved 2023-10-05, from https://networkx.org/documentation/stable/reference/generated/networkx.classes.function.density.html#density

NetworkX. (2023c). *out_degree_centrality*. Retrieved 2023-10-05, from https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.out_degree_centrality.html#networkx.algorithms.centrality.out_degree_centrality

NetworkX. (2023d). *pagerank*. Retrieved 2023-10-05, from https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

NetworkX. (2023e). *Reciprocity*. Retrieved 2023-10-05, from https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.reciprocity.reciprocity.html#networkx.algorithms.reciprocity.reciprocity

Newman, M. (2005, September). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. Retrieved 2025-03-17, from https://doi.org/10.1080/00107510500052444 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00107510500052444) doi: 10.1080/00107510500052444

Newman, M. E. J. (2002, November). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 208701. doi: 10.1103/PhysRevLett.89.208701

Newman, M. E. J. (2006, June). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. Retrieved 2019-03-14, from http://www.pnas.org/cgi/doi/10.1073/pnas.0601602103 doi: 10.1073/pnas.0601602103

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... Nguyen, C. M. (2022, October). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding*, *223*, 103525. Retrieved 2023-08-09, from http://arxiv.org/abs/1909.11573 (arXiv:1909.11573 [cs, eess]) doi: 10.1016/j.cviu.2022.103525

Nguyen Duc, A., Cruzes, D. S., Hanssen, G. K., Snarby, T., & Abrahamsson, P. (2017). Coopetition of Software Firms in Open Source Software Ecosystems. In A. Ojala, H. Holmström Olsson, & K. Werder (Eds.), *Software Business* (pp. 146–160). Cham: Springer International Publishing. doi: 10.1007/978-3-319-69191-6_10

Nguyen Duc, A., Cruzes, D. S., Terje, S., & Abrahamsson, P. (2019). Do software firms collaborate or compete? A model of coopetition in community-initiated OSS projects. *e-Informatica*, *13*(1). Retrieved 2022-03-17, from https://jyx.jyu.fi/handle/123456789/63053 (Accepted: 2019-03-07T08:47:38Z Publisher: Wroclaw University of Science and Technology) doi: 10.5277/e-Inf190102

Nguyen-Duc, A., Cruzes, D. S., Terje, S., & Abrahamsson, P. (2019). Do software firms collaborate or compete? A model of coopetition in community-initiated OSS projects. *e-Informatica Vol. XIII*. Retrieved 2023-12-29, from http://arxiv.org/abs/1808.06489 (arXiv:1808.06489 [cs]) doi: 10.5277/e-Inf190102

Nolan, M. (2023, July). Llama and ChatGPT Are Not Open-Source - IEEE Spectrum. *IEEE Spectrum*. Retrieved 2023-08-18, from https://spectrum.ieee.org/open-source-llm-not-open

NTIA. (2024, July). *Dual-Use Foundation Models with Widely Available Model Weights Report | National Telecommunications and Information Administration* (Tech. Rep.). Washington D.C., USA: National Telecommunications and Information Administration. Retrieved 2024-08-26, from https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report

O'Brien, D. (2019a, November). *The FOSS Contributor Fund: Forming a Community of Adopters*. Retrieved 2023-03-31, from https://medium.com/indeed-engineering/the-foss-contributor-fund-at-indeed-f164125c1ca0

O'Brien, D. (2019b). *Sustaining FOSS Projects By Democratizing The Sponsorship Process*. FOSDEM. Retrieved 2023-03-31, from https://archive.fosdem.org/2019/schedule/event/community_sustaining_foss_projects_democratizing_sponsorship/

O'Connor, R. (2021, December). *PyTorch vs TensorFlow in 2023*. Retrieved 2024-01-07, from https://www.assemblyai.com/blog/pytorch-vs-tensorflow-in-2023/

OECD. (2023). A blueprint for building national compute capacity for artificial intelligence. *OECD Digital Economy Papers*(350). Retrieved from https://doi.org/10.1787/876367e3-en (Publisher: OECD Publishing, Paris) doi: 10.1787/876367e3-en

OECD. (2024). *Artificial Intelligence in Society*. OECD. Retrieved 2024-09-02, from https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en doi: 10.1787/eedfee77-en

Okoli, C., & Oh, W. (2007, April). Investigating recognition-based performance in an open content community: A social capital perspective. *Information & Management*, *44*(3), 240–252. Retrieved 2023-04-06, from https://www.sciencedirect.com/science/article/pii/S0378720607000195 doi: 10.1016/j.im.2006.12.007

O'Mahony, S., & Bechky, B. A. (2008, September). Boundary Organizations: Enabling Collaboration among Unexpected Allies. *Administrative Science Quarterly*, *53*(3), 422–459. Retrieved 2021-11-16, from https://doi.org/10.2189/asqu.53.3.422 doi: 10.2189/asqu.53.3.422

O'Mahony, S., & Ferraro, F. (2007, October). The Emergence of Governance in an Open Source Community. *Academy of Management Journal*, *50*(5), 1079–1106. Retrieved 2022-02-04, from http://journals.aom.org/doi/10.5465/amj.2007.27169153 doi: 10.5465/amj.2007.27169153

on Digital Commons, E. W. T. (2022, June). *Towards a Sovereign Digital Infrastructure of Commons: Report of the European Working Team on the Digital Commons* (Tech. Rep.). OpenFuture. Retrieved 2023-02-01, from https://openfuture.eu/wp-content/uploads/2022/12/report_of_the_european_working_team_on_digital_commons_digital_assembly_june_2022_wnetherlands_cle843dbf.pdf

OpenAI. (2022, November). *Introducing ChatGPT*. Retrieved 2024-10-11, from https://openai.com/index/chatgpt/

OpenAI. (2024, May). *A landmark multi-year global partnership with News Corp.* Retrieved 2024-12-11, from https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/

OpenForum Europe. (2023, February). *What's on the agenda for OSS funders?* Brussels, Belgium. Retrieved 2024-08-31, from https://www.youtube.com/watch?v=MoyEjw7AORs

OpenSSF. (2022, May). *OSS Mobilization Plan* (Tech. Rep.). Washington D.C., USA: Open Source Security Foundation, Linux Foundation. Retrieved 2022-12-30, from https://openssf.org/oss-security-mobilization-plan/

Oreg, S., & Nov, O. (2008, September). Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in Human Behavior*, *24*(5), 2055–2073. Retrieved 2023-04-06, from https://www.sciencedirect.com/science/article/pii/S0747563207001537 doi: 10.1016/j.chb.2007.09.007

Orucevic-Alagic, A., & Host, M. (2014, August). Network Analysis of a Large Scale Open Source Project. In *2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications* (pp. 25–29). Verona, Italy: IEEE. Retrieved 2022-03-18, from http://ieeexplore.ieee.org/document/6928785/ doi: 10.1109/SEAA.2014.50

Osborne, C. (2024a, May). *Public-private funding models in open source software development: A case study on scikit-learn.* arXiv. Retrieved 2024-09-26, from http://arxiv.org/abs/2404.06484 (arXiv:2404.06484 [cs]) doi: 10.48550/arXiv.2404.06484

Osborne, C. (2024b, April). *Python Scripts for Mining Research Data from the Hugging Face Hub.* Oxford, UK. Retrieved 2024-04-21, from https://github.com/ccosborne/hf-hub-mining/tree/main

Osborne, C. (2024c). Why Companies Democratise Artificial Intelligence: The Case of Open Source Software Donations. *arXiv preprint arXiv:2409.17876*.

Osborne, C., Boehm, M., & Jimenez Santamaria, A. (2023, September). *The European Public Sector Open Source Opportunity: Challenges and Recommendations for Europe's Open Source Future* (Tech. Rep.). Brussels, Belgium: The Linux Foundation. Retrieved 2023-09-18, from https://

www.linuxfoundation.org/research/european-public-sector-opportunity?hsLang=en
Osborne, C., Daneshyan, F., He, R., Ye, H., Zhang, Y., & Zhou, M. (2024a, October). *Characterising Open Source Co-opetition in Company-hosted Open Source Software Projects: The Cases of PyTorch, TensorFlow, and Transformers.* arXiv. Retrieved 2024-11-01, from http://arxiv.org/abs/2410.18241 (arXiv:2410.18241 [cs]) doi: 10.48550/arXiv.2410.18241
Osborne, C., Daneshyan, F., He, R., Ye, H., Zhang, Y., & Zhou, M. (2024b, October). *Python Scripts for Mining Research Data from GitHub for Research Paper on Open Source Co-opetition.* Oxford, UK. Retrieved 2024-10-22, from https://github.com/ccosborne/paper-open-source-co-opetition (original-date: 2024-10-22T19:15:21Z)
Osborne, C., Ding, J., & Kirk, H. R. (2024, June). The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. *Journal of Computational Social Science*. Retrieved from https://doi.org/10.1007/s42001-024-00300-8 doi: 10.1007/s42001-024-00300-8
Osborne, C., Sharratt, P., Foster, D., & Boehm, M. (2024, November). *A Toolkit for Measuring the Impacts of Public Funding on Open Source Software Development.* arXiv. Retrieved 2024-11-22, from http://arxiv.org/abs/2411.06027 (arXiv:2411.06027 [cs]) doi: 10.48550/arXiv.2411.06027
OSI. (2007, March). *The Open Source Definition (v1.9).* Retrieved 2023-04-10, from https://opensource.org/osd/
OSI. (2023, October). *Deep Dive: AI.* Retrieved 2023-11-02, from https://opensource.org/deepdive/webinars/
OSI. (2024a, July). *Cailean Osborne: Voices of the Open Source AI Definition.* Retrieved 2024-09-11, from https://opensource.org/blog/cailean-osborne-voices-of-the-open-source-ai-definition
OSI. (2024b, September). *Is "Open Source" ever hyphenated?* Retrieved 2024-11-26, from https://opensource.org/blog/is-open-source-ever-hyphenated
OSI. (2024c, October). *The Open Source AI Definition – version 1.0.* Retrieved 2024-10-28, from https://opensource.org/ai/open-source-ai-definition
Ostander, S. A. (1993). "Surely you're not in this just to be helpful: Access, Rapport, and Interviews in Three Studies of Elites. *Journal of contemporary ethnography*, *22*(1), 7–27. (Place: Thousand Oaks Publisher: SAGE Periodicals Press)
Osterloh, M., Rota, S., & Kuster, B. (2003, January). Trust and Commerce in Open Source — A Contradiction? In (pp. 129–141). doi: 10.1007/978-3-7091-6088-6_8
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge, United Kingdom: Cambridge University Press.
Overney, C., Meinicke, J., Kästner, C., & Vasilescu, B. (2020, October). How to not get rich: an empirical study of donations in open source. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (pp. 1209–1221). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-04-07, from https://doi.org/10.1145/3377811.3380410 doi: 10.1145/3377811.3380410
O'Neil, M., Muselli, L., Raissi, M., & Zacchiroli, S. (2021, May). 'Open source has won and lost the war': Legitimising commercial–communal hybridisation in a FOSS project. *New Media & Society*, *23*(5), 1157–1180. Retrieved 2022-05-16, from https://doi.org/10.1177/1461444820907022 (Publisher: SAGE Publications) doi: 10.1177/1461444820907022
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*. Retrieved from http://ilpubs.stanford.edu:8090/422/
Panezi, A., Feldman, J., & Bernholz, L. (2020, August). *Open Source Projects as Critical Digital Infrastructure.* Stanford, CA, USA. Retrieved 2023-04-11, from https://vimeo.com/449102698
PaperswithCode. (2023). *Papers with Code.* Retrieved 2023-09-18, from https://paperswithcode.com/trends

Parnin, C., Treude, C., Grammel, L., & Storey, M.-A. (2012). *Crowd Documentation: Exploring the Coverage and the Dynamics of API Discussions on Stack Overflow* (Tech. Rep.). Atlanta, GA: Georgia Institute of Technology.

Patel, D., & Ahmad, A. (2023, May). *Google "We Have No Moat, And Neither Does OpenAI"*. Retrieved 2023-07-27, from https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., ... Dean, J. (2022, July). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, *55*(7), 18–28. Retrieved 2024-11-28, from https://ieeexplore.ieee.org/abstract/document/9810097?casa_token=dTRkHHohKAgAAAAA:vFdy8HpbIkpa1yWHIDL0GNfyAsC51JA4giRXBktAguqqR6vLM7siqXhGxNaS5PTgle2gljRQUQ (Conference Name: Computer) doi: 10.1109/MC.2022.3148714

Paul, K. (2023, December). Meta used copyrighted books for AI training despite its own lawyers' warnings, authors allege. *Reuters*. Retrieved 2024-12-30, from https://www.reuters.com/technology/meta-used-copyrighted-books-ai-training-despite-its-own-lawyers-warnings-authors-2023-12-12/

Pavlicek, R. C. (2000, October). *Keys to Effective Linux Advocacy Within Your Organization*. Retrieved 2023-04-04, from http://linuxprofessionalsolutions.com/pavlicek/oreilly/1999/als-fullpaper-1999.txt

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023, February). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. arXiv. Retrieved 2024-11-28, from http://arxiv.org/abs/2302.06590 (arXiv:2302.06590) doi: 10.48550/arXiv.2302.06590

Perez, M. (2022). Hugging Face Pulls in $100M Series C to Hire, Develop Product | Built In NYC. *Built in NYC*. Retrieved 2023-12-26, from https://www.builtinnyc.com/2022/05/09/hugging-face-raises-100m-series-c-hiring

Pfaff, B., & David, K. (1998). *Open source software is better for society than proprietary closed source software*. Retrieved 2023-04-04, from http://pintos.benpfaff.org/writings/anp/oss-is-better.html

Pilz, K., & Heim, L. (2023, November). *Compute at Scale: A Broad Investigation into the Data Center Industry*. arXiv. Retrieved 2024-11-24, from http://arxiv.org/abs/2311.02651 (arXiv:2311.02651 [cs]) doi: 10.48550/arXiv.2311.02651

Pipatanakul, K., Jirabovonvisut, P., Manakul, P., Sripaisarnmongkol, S., Patomwong, R., Chokchainant, P., & Tharnpipitchai, K. (2023, December). *Typhoon: Thai Large Language Models*. arXiv. Retrieved 2024-02-29, from http://arxiv.org/abs/2312.13951 (arXiv:2312.13951 [cs]) doi: 10.48550/arXiv.2312.13951

Pitt, L. F., Watson, R. T., Berthon, P., Wynn, D., & Zinkhan, G. (2006, April). The Penguin's Window: Corporate Brands From an Open-Source Perspective. *Journal of the Academy of Marketing Science*, *34*(2), 115–127. Retrieved 2023-04-12, from https://journals.sagepub.com/doi/abs/10.1177/0092070305284972 (Publisher: SAGE Publications Inc) doi: 10.1177/0092070305284972

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology. *Administrative Science Quarterly*, *41*(1), 116–145. Retrieved 2023-12-29, from https://www.jstor.org/stable/2393988 doi: 10.2307/2393988

Public.AI. (2024, August). *Public AI: Infrastructure for the Common Good* (Tech. Rep.). Library of Congress: Public AI Network. Retrieved from https://publicai.network/whitepaper doi: 10.5281/zenodo.13914560

PyTorch. (2023a, October). *About the PyTorch Foundation*. Retrieved 2023-10-03, from https://www.pytorch.org

PyTorch. (2023b). *PyTorch documentation — PyTorch 2.1 documentation*. Retrieved 2023-12-26, from https://pytorch.org/docs/stable/index.html

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145–151). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-05-29, from https://doi.org/10.1145/3375627.3375820

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022, January). AI in health and medicine. *Nature Medicine*, *28*(1), 31–38. Retrieved 2024-08-29, from https://www.nature.com/articles/s41591-021-01614-0 (Publisher: Nature Publishing Group) doi: 10.1038/s41591-021-01614-0

Raymond, E. S. (2001a). *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol: O'Reilly Media, Incorporated.

Raymond, E. S. (2001b). Homesteading the Noosphere. In *The Cathedral and the Bazaar*. United States: O'Reilly Media, Incorporated.

Raymond, E. S. (2001c, August). Why Microsoft smears-and fears-open source. *IEEE Spectrum*, *38*(8), 14–15. Retrieved 2024-04-12, from https://ieeexplore.ieee.org/abstract/document/938720?casa_token=swj_n0t2nz8AAAAA:cMuP5-aFZsSXQr5nVpL1veVwVdphFHpKZKn2BjgEq6qhjw5gV5n8thQmDwkgmOAzZeQVn8mTmA (Conference Name: IEEE Spectrum) doi: 10.1109/6.938720

Riehle, D. (2007, April). The Economic Motivation of Open Source Software: Stakeholder Perspectives. *Computer*, *40*(4), 25–32. (Conference Name: Computer) doi: 10.1109/MC.2007.147

Roberts, H., Ziosi, M., Osborne, C., Saouma, L., Belias, A., Buchser, M., ... Zeng, Y. (2023, February). *A Comparative Framework for AI Regulatory Policy* (Tech. Rep.). The International Centre of Expertise on Artificial Intelligence in Montreal. Retrieved from https://ceimia.org/en/projet/a-comparative-framework-for-ai-regulatory-policy/

Roberts, J., Hann, I.-H., & Slaughter, S. (2006, July). Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects. *Management Science*, *52*(7), 984–999. Retrieved 2021-12-20, from https://www.proquest.com/docview/213173475?parentSessionId=9eC3qbiS9o7Y%2BjzIHvg5YExtVjv4aNVgXx8n96cVwyA%3D&pq-origsite=primo&accountid=13042

Robles, G., & Gonzalez-Barahona, J. (2005). Developer identification methods for integrated data from various sources. In *International Conference on Software Engineering: Proceedings of the 2005 international workshop on Mining software repositories : St. Louis, Missouri; 17-17 May 2005* (pp. 1–5). ACM. (ISSN: 0163-5948)

Robson, C., & McCartan, K. (2016). *Real World Research : A Resource for Users of Social Research Methods in Applied Settings* (Fourth Edition ed.). Chichester.

Roth, E. (2024, August). *Authors sue Anthropic for training AI using pirated books.* Retrieved 2024-12-30, from https://www.theverge.com/2024/8/20/24224450/anthropic-copyright-lawsuit-pirated-books-ai

Runeson, P., & Höst, M. (2008). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering : an international journal*, *14*(2), 131–164. (Place: Boston Publisher: Springer US) doi: 10.1007/s10664-008-9102-8

Runeson, P., Höst, M., Rainer, A., & Regnell, B. (2012). *Case Study Research in Software Engineering: Guidelines and Examples*. John Wiley & Sons. (Google-Books-ID: T7rXoaxqPIAC)

Russell, S. J. (2021). *Artificial Intelligence: A Modern Approach* (4ª ed. ed.). Upper Saddle River: Pearson Education.

Salkever, A. (2023, July). *Open Source Maintainers: Exploring the people, practices, and constraints facing the world's most critical open source software projects* (Tech. Rep.). Linux Foundation. Retrieved 2023-09-18, from https://www.linuxfoundation.org/research/open-source-maintainers

Santos, C., Kuk, G., Kon, F., & Pearson, J. (2013, March). The attraction of contributors in free and open source software projects. *The Journal of Strategic Information Systems*, *22*(1), 26–45. Retrieved 2021-10-04, from https://www.sciencedirect.com/science/article/pii/

S0963868712000340 doi: 10.1016/j.jsis.2012.07.004

Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., & Kertész, J. (2007, February). Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E*, *75*(2), 027105. Retrieved from https://link.aps.org/doi/10.1103/PhysRevE.75.027105 (Publisher: American Physical Society) doi: 10.1103/PhysRevE.75.027105

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., ... Coyle, D. (2024, February). *Computing Power and the Governance of Artificial Intelligence.* arXiv. Retrieved 2024-11-24, from http://arxiv.org/abs/2402.08797 (arXiv:2402.08797 [cs]) doi: 10.48550/arXiv.2402.08797

Savić, M., Ivanović, M., & Jain, L. C. (2019). *Complex Networks in Software, Knowledge, and Social Systems* (Vol. 148). Cham: Springer International Publishing. Retrieved 2019-10-21, from http://link.springer.com/10.1007/978-3-319-91196-0 doi: 10.1007/978-3-319-91196-0

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... Wolf, T. (2023, June). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/2211.05100 (arXiv:2211.05100 [cs]) doi: 10.48550/arXiv.2211.05100

Schaarschmidt, M., Walsh, G., & von Kortzfleisch, H. F. O. (2015, April). How do firms influence open source software communities? A framework and empirical analysis of different governance modes. *Information and Organization*, *25*(2), 99–114. Retrieved 2024-07-16, from https://www.sciencedirect.com/science/article/pii/S1471772715000111 doi: 10.1016/j.infoandorg.2015.03.001

Schofield, A., & Cooper, G. (2006, June). Participation in Free and Open Source Communities: An Empirical Study of Community Members' Perceptions. In *Open Source Systems: IFIP Working Group 2.13 Foundation on Open Source Software.* Como, Italy.

Schoonmaker, S. (2018). *Free Software, the Internet, and Global Communities of Resistance*. New York: Routledge. doi: 10.4324/9781315672786

scikit learn. (2023). *scikit-learn.* Retrieved 2023-09-18, from https://scikit-learn/stable/about.html

Scott, S., Brackett, S. A., Herr, T., & Hamin, M. (2023, February). *Avoiding the Success Trap: Toward Policy for Open-Source Software as Infrastructure* (Tech. Rep.). Washington D.C., USA: Atlantic Council.

SEA-LION.AI. (2024, December). *SEA-LION.AI.* Retrieved 2024-12-30, from https://sea-lion.ai/

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... Gupta, A. (2023, September). *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives.* arXiv. Retrieved 2024-02-12, from http://arxiv.org/abs/2311.09227 (arXiv:2311.09227 [cs]) doi: 10.48550/arXiv.2311.09227

Seger, E., Ovadya, A., Garfinkel, B., Siddarth, D., & Dafoe, A. (2023, March). *Democratising AI: Multiple Meanings, Goals, and Methods.* arXiv. Retrieved 2023-03-23, from http://arxiv.org/abs/2303.12642 (arXiv:2303.12642 [cs]) doi: 10.48550/arXiv.2303.12642

Sen, R., Ahmad, S., Phokeer, A., Farooq, Z. A., Qazi, I. A., Choffnes, D., & Gummadi, K. P. (2017, October). Inside the Walled Garden: Deconstructing Facebook's Free Basics Program. *SIGCOMM Comput. Commun. Rev.*, *47*(5), 12–24. Retrieved 2025-03-16, from https://dl.acm.org/doi/10.1145/3155055.3155058 doi: 10.1145/3155055.3155058

SGPI. (2021, November). *France 2030: Stratégie National pour l'Intelligence Artificielle.* Secrétariat général pour l'investissement. Retrieved 2022-04-13, from https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-11/dossier-de-presse---strat-gie-nationale-pour-l-intelligence-artificielle-2e-phase-14920.pdf

Shah, S. K. (2006). Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. *Management science*, *52*(7), 1000–1014. doi: 10.1287/mnsc.1060

.0553

Shaw, J. (2020). *Hyperreal estate: the production of new urban real estate markets* (http://purl.org/dc/dcmitype/Text, University of Oxford). Retrieved 2023-03-07, from https://ora.ox.ac.uk/objects/uuid:995bea51-7f36-4464-b233-238446e0bb41

Shevlane, T. (2022, January). *Sharing Powerful AI Models | GovAI Blog*. Retrieved 2022-05-16, from https://www.governance.ai/post/sharing-powerful-ai-models

Shull, F., Singer, J., & Sjøberg, D. I. K. (2008). *Guide to advanced empirical software engineering*. London: Springer.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359. (Publisher: Nature Publishing Group)

Singh, P. V. (2010, September). The small-world effect: The influence of macro-level properties of developer collaboration networks on open-source project success. *ACM Transactions on Software Engineering and Methodology*, *20*(2), 6:1–6:27. Retrieved 2023-12-27, from https://dl.acm.org/doi/10.1145/1824760.1824763 doi: 10.1145/1824760.1824763

Smilkov, D., & Kocarev, L. (2010, June). Rich-club and page-club coefficients for directed graphs. *Physica A: Statistical Mechanics and its Applications*, *389*(11), 2290–2299. Retrieved 2023-10-06, from https://www.sciencedirect.com/science/article/pii/S0378437110000944 doi: 10.1016/j.physa.2010.02.001

Smith, K. E. (2006). Problematising power relations in 'elite' interviews. *Geoforum*, *37*(4), 643–653. (Publisher: Elsevier Ltd)

Snarby, T. (2013). *Collaboration Patterns among Commercial Firms in Community-Based OSS Projects*. Institutt for datateknikk og informasjonsvitenskap.

Solaiman, I. (2023, February). *The Gradient of Generative AI Release: Methods and Considerations*. arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/2302.04844 (arXiv:2302.04844 [cs]) doi: 10.48550/arXiv.2302.04844

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... Wang, J. (2019, November). *Release Strategies and the Social Impacts of Language Models*. arXiv. Retrieved 2023-08-09, from http://arxiv.org/abs/1908.09203 (arXiv:1908.09203 [cs]) doi: 10.48550/arXiv.1908.09203

Sonnenburg, S., Braun, M. L., Cheng, S. O., Bengio, S., Bottou, L., Holmes, G., ... Williamson, R. C. (2007, October). The Need for Open Source Software in Machine Learning. *Journal of Machine Learning Research*, *8*, 2443–2466. Retrieved 2022-04-10, from http://www.scopus.com/inward/record.url?scp=35748939406&partnerID=8YFLogxK

SourceForge. (2022). *About*. Retrieved 2022-04-19, from https://sourceforge.net/about

South Park Commons. (2024, August). *Mark Zuckerberg on Llama, AI, & Minus One*. Retrieved 2024-08-29, from https://www.youtube.com/watch?v=02fBBoZa9l4

Spaeth, S., Haefliger, S., von Krogh, G., & Renzl, B. (2008, March). Communal Resources in Open Source Software Development. *Information Research: An International Electronic Journal*, *13*(1). Retrieved 2023-04-06, from https://eric.ed.gov/?id=EJ837230 (Publisher: Thomas D ERIC Number: EJ837230)

Srikumar, M., Chmielinski, K., & Chang, J. (2024, July). *Risk Mitigation Strategies for the Open Foundation Model Value Chain* (Tech. Rep.). San Francisco, CA, USA: Partnership on AI. Retrieved 2024-08-26, from https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

Srnicek, N. (2017). *Platform capitalism*. Cambridge: Polity Press. (Book Title: Platform capitalism)

Srnicek, N. (2022, May). Data, Compute, Labor. In M. Graham & F. Ferrari (Eds.), *Digital work in the planetary market*. Retrieved 2022-05-26, from https://direct.mit.edu/books/oa-edited-volume/5319/chapter/3800166/Data-Compute-Labor

Stability.AI. (2022, August). *Stable Diffusion Public Release*. Retrieved 2023-08-09, from https://stability.ai/blog/stable-diffusion-public-release

Stallman, R. (1984). *The GNU Manifesto - GNU Project - Free Software Foundation*. Retrieved 2023-04-04, from https://www.gnu.org/gnu/manifesto.en.html

Stallman, R. (2002). *Free software, free society : selected essays of Richard M. Stallman*. Boston, MA : Free Software Foundation. Retrieved 2022-05-16, from http://archive.org/details/freesoftwarefree00rich

Stanford. (2021, August). *Introducing the Center for Research on Foundation Models (CRFM)*. Retrieved 2024-12-12, from https://hai.stanford.edu/news/introducing-center-research-foundation-models-crfm

Statista. (2022a). *Alphabet: number of employees 2022*. Retrieved 2023-12-26, from https://www.statista.com/statistics/273744/number-of-full-time-google-employees/

Statista. (2022b). *Meta workforce*. Retrieved 2023-12-26, from https://www.statista.com/statistics/273563/number-of-facebook-employees/

Stewart, K. J., & Gosain, S. (2006). The Impact of Ideology on Effectiveness in Open Source Software Development Teams. *MIS Quarterly*, *30*(2), 291–314. Retrieved 2023-04-06, from https://www.jstor.org/stable/25148732 (Publisher: Management Information Systems Research Center, University of Minnesota) doi: 10.2307/25148732

STF. (2022). *Sovereign Tech Fund*. Retrieved 2023-03-12, from https://sovereigntechfund.de/en

Strasser, C., Hertweck, K., Greenberg, J., Taraborelli, D., & Vu, E. (2022, November). Ten simple rules for funding scientific open source software. *PLOS Computational Biology*, *18*(11), e1010627. Retrieved 2023-11-13, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010627 (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1010627

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge: University Press.

Streicher, M. (2020, October). What is Commercial Open Source Software. *Webiny*. Retrieved 2022-05-10, from https://www.webiny.com/blog/what-is-commercial-open-source

Stuart, T. E. (1998). Network Positions and Propensities to Collaborate: An Investigation of Strategic Alliance Formation in a High-Technology Industry. *Administrative Science Quarterly*, *43*(3), 668–698. Retrieved 2023-12-07, from https://www.jstor.org/stable/2393679 (Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University]) doi: 10.2307/2393679

Stuart, T. E. (2000). Interorganizational alliances and the performance of firms: a study of growth and innovation rates in a high-technology industry. *Strategic Management Journal*, *21*(8), 791–811. Retrieved 2023-12-07, from https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0266%28200008%2921%3A8%3C791%3A%3AAID-SMJ121%3E3.0.CO%3B2-K (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-0266%28200008%2921%3A8%3C791%3A%3AAID-SMJ121%3E3.0.CO%3B2-K) doi: 10.1002/1097-0266(200008)21:8<791::AID-SMJ121>3.0.CO;2-K

Stuart, T. E., Hoang, H., & Hybels, R. C. (1999). Interorganizational Endorsements and the Performance of Entrepreneurial Ventures. *Administrative science quarterly*, *44*(2), 315–349. (Place: Ithaca, N.Y Publisher: Cornell University Samuel Curtis Johnson Graduate School of Management) doi: 10.2307/2666998

Subramanyam, R., & Xia, M. (2008). Free/Libre Open Source Software development in developing and developed countries: A conceptual framework with an exploratory study. *Decision Support Systems*, *46*(1), 173–186. (Place: Amsterdam Publisher: Elsevier B.V) doi: 10.1016/j.dss.2008.06.006

Synopsys. (2023). *Open Source Security and Analysis Report* (Tech. Rep.). Mountain View, CA, USA. Retrieved 2023-05-04, from https://www.synopsys.com/software-integrity/resources/analyst-reports/open-source-security-risk-analysis.html

Szymański, K., & Ochodek, M. (2023, September). On the Applicability of the Pareto Principle to Source-Code Growth in Open Source Projects. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)* (pp. 781–789). Retrieved 2023-11-19, from

https://ieeexplore.ieee.org/document/10306152 doi: 10.15439/2023F5221
Takhteyev, Y. (2012). *Coding places: software practice in a South American city*. Cambridge, Mass.: MIT Press.
Takhteyev, Y., & Hilts, A. (2010, January). *Investigating the Geography of Open Source Software through Github*.
Tan, X., Chen, Y., Wu, H., Zhou, M., & Zhang, L. (2023, February). *Is It Enough to Recommend Tasks to Newcomers? Understanding Mentoring on Good First Issues*. arXiv. Retrieved 2023-10-24, from http://arxiv.org/abs/2302.05058 (arXiv:2302.05058 [cs]) doi: 10.48550/arXiv.2302.05058
Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T. B. (2023, March). *Alpaca: A Strong, Replicable Instruction-Following Model*. Retrieved 2023-08-14, from https://crfm.stanford.edu/2023/03/13/alpaca.html
Tapscott, D. (2011). *Wikinomics: How mass collaboration changes everything*. London: Atlantic Books.
Tarkowski, A. (2023, August). *The Mirage of Open-Source AI: Analyzing Meta's Llama 2 Release Strategy*. Retrieved 2023-09-18, from https://openfuture.eu/blog/the-mirage-of-open-source-ai-analyzing-metas-llama-2-release-strategy
Teixeira, J. (2014, August). Understanding Coopetition in the Open-Source Arena: The Cases of WebKit and OpenStack. In *Proceedings of The International Symposium on Open Collaboration* (pp. 1–5). New York, NY, USA: Association for Computing Machinery. Retrieved 2021-10-14, from https://doi.org/10.1145/2641580.2641627 doi: 10.1145/2641580.2641627
Teixeira, J., & Lin, T. (2014). Collaboration in the open-source arena: The WebKit case. *Proceedings of the 52nd ACM conference on Computers and people research - SIGSIM-CPR '14*, 121–129. Retrieved 2021-10-21, from http://arxiv.org/abs/1401.5996 (arXiv: 1401.5996) doi: 10.1145/2599990.2600009
Teixeira, J., Mian, S. Q., & Hytti, U. (2016). Cooperation among competitors in the open-source arena. In *IS in Organizations and Society* (p. 38). Dublin, Ireland. Retrieved from https://arxiv.org/abs/1612.09462
Teixeira, J., Robles, G., & González-Barahona, J. M. (2015, July). Lessons learned from applying social network analysis on an industrial Free/Libre/Open Source Software ecosystem. *Journal of Internet Services and Applications*, *6*(1), 14. Retrieved 2021-10-21, from https://doi.org/10.1186/s13174-015-0028-2 doi: 10.1186/s13174-015-0028-2
Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., & Stallings, J. (2017, May). Gender differences and bias in open source: pull request acceptance of women versus men. *PeerJ Computer Science*, *3*, e111. Retrieved 2021-12-20, from https://peerj.com/articles/cs-111 (Publisher: PeerJ Inc.) doi: 10.7717/peerj-cs.111
Thiel, D., Stroebel, M., & Portnoff, R. (2023, June). *Generative ML and CSAM: Implications and Mitigations* (Tech. Rep.). Stanford University: Stanford University. Retrieved from https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations
Thomas, A. C. (2020, February). *Open AI just choose PyTorch over TensorFlow*. Retrieved 2024-12-03, from https://medium.com/the-ultimate-engineer/open-ai-just-choose-pytorch-over-tensorflow-d1385133d534
Tidelift. (2020). *2020: The managed open source survey* (Tech. Rep.). Author. Retrieved 2024-04-26, from https://tidelift.com/subscription/2020-managed-open-source-survey
Torvalds, L. (2001). *Just for fun: the story of an accidental revolutionary*. New York ; London: Texere.
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023, July). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv. Retrieved 2024-07-04, from http://arxiv.org/abs/2307.09288 (arXiv:2307.09288 [cs]) doi: 10.48550/arXiv.2307.09288
Trinkenreich, B., Guizani, M., Wiese, I., Sarma, A., & Steinmacher, I. (2020, October). Hidden Figures: Roles and Pathways of Successful OSS Contributors. *Proceedings of the ACM on Human-*

Computer Interaction, *4*(CSCW2), 180:1–180:22. Retrieved 2024-04-27, from https://doi.org/10.1145/3415251 doi: 10.1145/3415251

Tsamados, A., Floridi, L., & Taddeo, M. (2023, September). *The Cybersecurity Crisis of Artificial Intelligence: Unrestrained Adoption and Natural Language-Based Attacks* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-10-08, from https://papers.ssrn.com/abstract=4578165 doi: 10.2139/ssrn.4578165

Tsing, A. (2009, April). Supply Chains and the Human Condition. *Rethinking Marxism*, *21*(2), 148–176. Retrieved 2024-11-26, from https://doi.org/10.1080/08935690902743088 (Publisher: Routledge _eprint: https://doi.org/10.1080/08935690902743088) doi: 10.1080/08935690902743088

Tubaro, P., Ryan, L., & D'angelo, A. (2016, May). The Visual Sociogram in Qualitative and Mixed-Methods Research. *Sociological Research Online*, *21*(2), 180–197. Retrieved 2024-05-29, from https://doi.org/10.5153/sro.3864 doi: 10.5153/sro.3864

Valdivia, A. (2024, October). The supply chain capitalism of AI: a call to (re)think algorithmic harms and resistance through environmental lens. *Information, Communication & Society*, *0*(0), 1–17. Retrieved 2024-10-31, from https://doi.org/10.1080/1369118X.2024.2420021 (Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2024.2420021) doi: 10.1080/1369118X.2024.2420021

Valiev, M., Vasilescu, B., & Herbsleb, J. (2018, October). Ecosystem-level determinants of sustained activity in open-source projects: a case study of the PyPI ecosystem. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 644–655). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-03, from https://dl.acm.org/doi/10.1145/3236024.3236062 doi: 10.1145/3236024.3236062

Varoquaux, G. (2018, September). *A foundation for scikit-learn at Inria.* Retrieved 2023-05-18, from https://gael-varoquaux.info/programming/a-foundation-for-scikit-learn-at-inria.html

Varoquaux, G. (2021, September). *Hiring someone to develop scikit-learn community and industry partners.* Retrieved 2022-04-13, from https://softbranchdevelopers.com/gal-varoquaux-hiring-someone-to-develop-scikit-learn-community-and-industry-partners/

Varoquaux, G., Luccioni, A. S., & Whittaker, M. (2024, September). *Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI.* arXiv. Retrieved 2024-09-24, from http://arxiv.org/abs/2409.14160 (arXiv:2409.14160 [cs]) doi: 10.48550/arXiv.2409.14160

Vasilescu, B., Capiluppi, A., & Serebrenik, A. (2014, September). Gender, Representation and Online Participation: A Quantitative Study. *Interacting with Computers*, *26*(5), 488–511. Retrieved 2021-12-20, from https://doi.org/10.1093/iwc/iwt047 doi: 10.1093/iwc/iwt047

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023, August). *Attention Is All You Need.* arXiv. Retrieved 2024-03-12, from http://arxiv.org/abs/1706.03762 (arXiv:1706.03762 [cs]) doi: 10.48550/arXiv.1706.03762

Vaughan-Nichols, S. (2021a, June). Hard work and poor pay stresses out open-source maintainers. *ZDNET*. Retrieved 2023-04-07, from https://www.zdnet.com/article/hard-work-and-poor-pay-stresses-out-open-source-maintainers/

Vaughan-Nichols, S. (2021b, December). *Log4Shell: We Are in So Much Trouble.* Retrieved 2023-11-13, from https://thenewstack.io/log4shell-we-are-in-so-much-trouble/

Vipra, J., & Myers West, S. (2023, September). *Computational Power and AI.* Retrieved 2024-12-02, from https://ainowinstitute.org/publication/policy/compute-and-ai

von Hippel, E. (2001, July). Innovation by User Communities: Learning from Open Source Software. *Sloan Management Review*.

von Hippel, E. (2003). Open Source Projects as Horizontal Innovation Networks - By and for users. *IDEAS Working Paper Series from RePEc*. Retrieved 2023-04-04, from https://search.proquest.com/docview/1698779037?pq-origsite=primo (Place: St. Louis Publisher: Fed-

eral Reserve Bank of St Louis)

von Krogh, G., Haefliger, S., Spaeth, S., & Wallin, M. W. (2012). Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS quarterly*, *36*(2), 649–676. doi: 10.2307/41703471

Wagstrom, P. A. (2009). *Vertical interaction in open software engineering communities* (Doctoral dissertation, Carnegie Insitute of Technology/School of Computer Science, Carnegie Mellon University). Retrieved from https://patrick.wagstrom.net/thesis/wagstrom-phd-thesis-single.pdf

Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024, November). E-waste challenges of generative artificial intelligence. *Nature Computational Science*, *4*(11), 818–823. Retrieved 2024-11-24, from https://www.nature.com/articles/s43588-024-00712-6 (Publisher: Nature Publishing Group) doi: 10.1038/s43588-024-00712-6

Warren, T. (2020, May). Microsoft: we were wrong about open source. *The Verge*. Retrieved 2022-04-12, from https://www.theverge.com/2020/5/18/21262103/microsoft-open-source-linux-history-wrong-statement

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.

Weaver, O. (2020, December). *Beware: Over Half of the GitHub Public Repositories are Not Open Source Licensed!* Retrieved 2023-11-02, from https://openweaver.medium.com/beware-over-half-of-the-github-public-repositories-are-not-open-source-licensed-23c7d2b5b621

West, J. (2003, July). How open is open enough?: Melding proprietary and open source platform strategies. *Research Policy*, *32*(7), 1259–1285. Retrieved 2023-12-29, from https://www.sciencedirect.com/science/article/pii/S0048733303000520 doi: 10.1016/S0048-7333(03)00052-0

West, J., & Gallagher, S. (2006). Challenges of open innovation: the paradox of firm investment in open-source software. *R&D Management*, *36*(3), 319–331. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9310.2006.00436.x (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9310.2006.00436.x) doi: https://doi.org/10.1111/j.1467-9310.2006.00436.x

West, J., & O'Mahony, S. (2005). Contrasting Community Building in Sponsored and Community Founded Open Source Projects. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 196c–196c). Big Island, HI, USA: IEEE. Retrieved 2023-10-25, from http://ieeexplore.ieee.org/document/1385638/ doi: 10.1109/HICSS.2005.166

White, M., Haddad, I., Osborne, C., Xiao-Yang, Liu, Abdelmonsef, A., & Varghese, S. (2024, March). *The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI*. arXiv. Retrieved 2024-03-21, from http://arxiv.org/abs/2403.13784 (arXiv:2403.13784 [cs]) doi: 10.48550/arXiv.2403.13784

Whittaker, M. (2021, November). The steep cost of capture. *Interactions*, *28*(6), 50–55. Retrieved 2021-11-16, from https://doi.org/10.1145/3488666 doi: 10.1145/3488666

Wichmann, T. (2002, July). *Firms' Open Source Activities: Motivations and Policy Implications* (Tech. Rep.). Berlin, Germany: Berlecon Research GmbH. Retrieved 2023-04-04, from https://www.math.unipd.it/~bellio/FLOSS%20Final%20Report%20-%20Part%202%20-%20Firms'%20Open%20Source%20Activities%20-%20Motivations%20and%20Policy%20Implications.pdf

Widder, D. G., Gururaja, S., & Suchman, L. (2024). Basic Research, Lethal Effects: Military AI Research Funding as Enlistment. *arXiv preprint arXiv:2411.17840*.

Widder, D. G., West, S., & Whittaker, M. (2023, August). *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-08-18, from https://papers.ssrn.com/abstract=4543807

Widder, D. G., Whittaker, M., & West, S. M. (2024, November). Why 'open' AI systems are actually closed, and why this matters. *Nature*, *635*(8040), 827–833. Retrieved 2024-12-02, from

https://www.nature.com/articles/s41586-024-08141-1 (Publisher: Nature Publishing Group) doi: 10.1038/s41586-024-08141-1

Wiggers, K. (2023, October). *5 investors on the pros and cons of open source AI business models.* Retrieved 2024-04-19, from https://techcrunch.com/2023/10/18/pros-cons-open-source-ai-business-models/

Wiggers, K. (2024, June). AI training data has a price tag that only Big Tech can afford. *TechCrunch*. Retrieved 2024-12-13, from https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/

Williams, A. (2023, January). *Enabling Global Collaboration* (Tech. Rep.). San Francisco, CA, USA: Linux Foundation. Retrieved 2023-10-31, from https://www.linuxfoundation.org/research/open-source-fragmentation

Wladawsky-Berger, I. (2023, June). *Are Open AI Models Safe?* Retrieved 2023-06-13, from https://www.linuxfoundation.org/blog/are-open-ai-models-safe

Woods, D., & Guliani, G. (2005). *Open Source for the Enterprise: Managing Risks, Reaping Rewards*. Sebastopol: O'Reilly Media, Incorporated.

Wright, A. (2023, March). Bloomberg Launches FOSS Fund to Support Free and Open Source Projects. *Bloomberg L.P.*. Retrieved 2023-03-31, from https://www.bloomberg.com/company/stories/bloomberg-ospo-launches-foss-contributor-fund/

Wu, C.-G., Gerlach, J. H., & Young, C. E. (2007, April). An empirical analysis of open source software developers' motivations and continuance intentions. *Information & Management*, *44*(3), 253–262. Retrieved 2023-04-06, from https://www.sciencedirect.com/science/article/pii/S0378720607000067 doi: 10.1016/j.im.2006.12.006

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... Hazelwood, K. (2022, April). Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems*, *4*, 795–813. Retrieved 2024-11-24, from https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html

Xia, X., Wang, W., Zhao, S., Bian, S., & Wang, R. (2023, April). Lessons Learned From the Ant Group Open Source Program Office. *Computer*, *56*(4), 92–97. (Conference Name: Computer) doi: 10.1109/MC.2022.3219638

Xu, B., Jones, D. R., & Shao, B. (2009, April). Volunteers' involvement in online community based software development. *Information & Management*, *46*(3), 151–158. Retrieved 2023-04-06, from https://www.sciencedirect.com/science/article/pii/S0378720609000044 doi: 10.1016/j.im.2008.12.005

Xu, J., Christley, S., & Madey, G. (2006, January). 12 - Application of Social Network Analysis to the Study of Open Source Software. In J. Bitzer & P. J. H. Schröder (Eds.), *The Economics of Open Source Software Development* (pp. 247–269). Amsterdam: Elsevier. Retrieved 2023-10-04, from https://www.sciencedirect.com/science/article/pii/B9780444527691500123 doi: 10.1016/B978-044452769-1/50012-3

Yamashita, K., McIntosh, S., Kamei, Y., Hassan, A. E., & Ubayashi, N. (2015). Revisiting the applicability of the pareto principle to core development teams in open source software projects. In *Proceedings of the 14th International Workshop on principles of software evolution* (pp. 46–55). ACM.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., & Lin, D. (2023, October). *Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models.* arXiv. Retrieved 2024-02-29, from http://arxiv.org/abs/2310.02949 (arXiv:2310.02949 [cs])

Ye, Y., & Kishida, K. (2003, May). Toward an understanding of the motivation of open source software developers. In *25th International Conference on Software Engineering, 2003. Proceedings.* (pp. 419–429). (ISSN: 0270-5257) doi: 10.1109/ICSE.2003.1201220

Yeasmin, S. (2019, May). Benefits of Artificial Intelligence in Medicine. In *2019 2nd International Conference on Computer Applications & Information Security (IC-*

CAIS) (pp. 1–6). Retrieved 2024-08-29, from https://ieeexplore.ieee.org/abstract/document/8769557?casa_token=cGYaFinJHVEAAAAA:ePOsmPQygpv2_jSyJ2358WxZOA1kDXNCQkzIY3D0pFHmfD0gpHHNlPHgbTeoOb_ZIu9HHCeR doi: 10.1109/CAIS.2019.8769557

Yin, R. K. (2018). *Case study research and applications: design and methods* (Sixth edition. ed.). Los Angeles, USA: Sage.

Young, J.-G., Casari, A., McLaughlin, K., Trujillo, M. Z., Hébert-Dufresne, L., & Bagrow, J. P. (2021, May). Which contributions count? Analysis of attribution in open source. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)* (pp. 242–253). Retrieved 2024-10-30, from https://ieeexplore.ieee.org/abstract/document/9463079?casa_token=wfs3p2ztKVIAAAAA:r_eH_CvTIhWWjYcPYPP9mgwq9aEqmEJLQ1AG4jgi8ElXBKOPOKLSwDlSISnSf94GGFV2bJwBuA (ISSN: 2574-3864) doi: 10.1109/MSR52588.2021.00036

Yu, J., Jiang, Z., & Chan, H. C. (2007). Knowledge contribution in problem solving virtual communities: the mediating role of individual motivations. In *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce* (pp. 144–152). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-04-06, from https://doi.org/10.1145/1235000.1235034 doi: 10.1145/1235000.1235034

Yue, D., & Nagle, F. (2024, September). *Igniting Innovation: Evidence from PyTorch on Technology Control in Open Collaboration* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2024-09-24, from https://papers.ssrn.com/abstract=4960578 doi: 10.2139/ssrn.4960578

Zeitlyn, D. (2003, July). Gift economies in the development of open source software: anthropological reflections. *Research Policy*, *32*(7), 1287–1291. Retrieved 2023-04-06, from https://www.sciencedirect.com/science/article/pii/S0048733303000532 doi: 10.1016/S0048-7333(03)00053-2

Zemlin, J. (2022, September). *Welcoming PyTorch to the Linux Foundation*. Retrieved 2022-09-12, from https://linuxfoundation.org/blog/welcoming-pytorch-to-the-linux-foundation/

Zhang, X., Wang, T., Yu, Y., Zeng, Q., Li, Z., & Wang, H. (2021, November). Who, What, Why and How? Towards the Monetary Incentive in Crowd Collaboration: A Case Study of Github's Sponsor Mechanism. *arXiv:2111.13323 [cs]*. Retrieved 2022-02-23, from http://arxiv.org/abs/2111.13323 (arXiv: 2111.13323)

Zhang, Y., Qin, M., Stol, K.-J., Zhou, M., & Liu, H. (2024, April). How Are Paid and Volunteer Open Source Developers Different? A Study of the Rust Project. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (pp. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-11-06, from https://dl.acm.org/doi/10.1145/3597503.3639197 doi: 10.1145/3597503.3639197

Zhang, Y., Stol, K.-J., Liu, H., & Zhou, M. (2022, November). Corporate dominance in open source ecosystems: a case study of OpenStack. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 1048–1060). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-18, from https://dl.acm.org/doi/10.1145/3540250.3549117 doi: 10.1145/3540250.3549117

Zhang, Y., Tan, X., Zhou, M., & Jin, Z. (2018). Companies' domination in FLOSS development: an empirical study of OpenStack. In *Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings* (pp. 440–441). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-10-20, from https://doi.org/10.1145/3183440.3195047 doi: 10.1145/3183440.3195047

Zhang, Y., Zhou, M., Mockus, A., & Jin, Z. (2021, October). Companies' Participation in OSS Development–An Empirical Study of OpenStack. *IEEE Transactions on Software Engineering*, *47*(10), 2242–2259. Retrieved 2023-10-08, from https://ieeexplore.ieee.org/abstract/document/8862903 (Conference Name: IEEE Transactions on Software Engineering) doi:

10.1109/TSE.2019.2946156

Zhang, Y., Zhou, M., Stol, K.-J., Wu, J., & Jin, Z. (2020). How do companies collaborate in open source ecosystems? an empirical study of OpenStack. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (pp. 1196–1208). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-10-20, from https://doi.org/10.1145/3377811.3380376 doi: 10.1145/3377811.3380376

Zhou, M., & Mockus, A. (2010, November). Developer fluency: achieving true mastery in software projects. In *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering* (pp. 137–146). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-12-26, from https://dl.acm.org/doi/10.1145/1882291.1882313 doi: 10.1145/1882291.1882313

Zhou, M., & Mockus, A. (2015). Who Will Stay in the FLOSS Community? Modeling Participant's Initial Behavior. *IEEE transactions on software engineering*, *41*(1), 82–99. (Place: New York Publisher: IEEE)

Zhou, M., Mockus, A., Ma, X., Zhang, L., & Mei, H. (2016). Inflow and Retention in OSS Communities with Commercial Involvement: A Case Study of Three Hybrid Projects. *ACM Transactions on Software Engineering and Methodology*, *25*(2), 13:1–13:29. Retrieved 2022-10-20, from https://doi.org/10.1145/2876443 doi: 10.1145/2876443

Zhou, S., & Mondragon, R. (2004, March). The rich-club phenomenon in the Internet topology. *IEEE Communications Letters*, *8*(3), 180–182. Retrieved 2023-10-08, from https://ieeexplore.ieee.org/abstract/document/1278314 (Conference Name: IEEE Communications Letters) doi: 10.1109/LCOMM.2004.823426

Zhu, J., & Wei, J. (2019, May). An Empirical Study of Multiple Names and Email Addresses in OSS Version Control Repositories. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (pp. 409–420). Retrieved 2024-01-12, from https://ieeexplore.ieee.org/abstract/document/8816766?casa_token=1eVm4qoYyHIAAAAA:X-a6L5pkzrI2occLAX072wajQZiyZwlQNGJOHHIgx0qnQHZn2uT1jcmVoPPOxNec3PlzS1muRQ (ISSN: 2574-3864) doi: 10.1109/MSR.2019.00068

Zineldin, M. (2004). Co-opetition: the organisation of the future. *Marketing Intelligence & Planning*, *22*(7), 780–790. (Publisher: Emerald Group Publishing Limited)

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. New York, NY: PublicAffairs.

Zuckerman, H. (1972). INTERVIEWING AN ULTRA-ELITE. *Public opinion quarterly*, *36*(2), 159–175. (Place: Princeton, N.J Publisher: Oxford University Press)

Ågerfalk, P. J., & Fitzgerald, B. (2008, June). Outsourcing to an unknown workforce: exploring opensourcing as a global sourcing strategy. *MIS Quarterly*, *32*(2), 385–409.