



POLITECNICO
MILANO 1863

RE.PUBLIC@POLIMI

This is the accepted version of:

Baruffaldi, S. H., Harhoff, D., & Liebender-Luc, A. S. (2024). Scientific Paradigms, Graphics Processing Units and the Evolution of Artificial Intelligence. In Academy of Management Proceedings (Vol. 2024, No. 1, p. 20222). Valhalla, NY 10595: Academy of Management.

Published Journal Article available at:

<https://journals.aom.org/doi/abs/10.5465/AMPROC.2024.286bp>

Access to the published version may require subscription.

When citing this work, cite the original published paper.

Permanent link to this version: <https://hdl.handle.net/11311/1272826>

SCIENTIFIC PARADIGMS, GRAPHICS PROCESSING UNITS AND THE EVOLUTION OF ARTIFICIAL INTELLIGENCE

STEFANO BARUFFALDI

Politecnico di Milano, School of Management
Via Raffaele Lambruschini, 4C, 20156 Milano MI
Max Planck Institute for Innovation and Competition

DIETMAR HARHOFF

Max Planck Institute for Innovation and Competition

ANNA-SOPHIE LIEBENDER-LUC

Max Planck Institute for Innovation and Competition
Ludwigs-Maximilians-University Munich

INTRODUCTION

In recent years, many nations and enterprises have embarked on a race to become leaders in AI technology, and many AI applications have been developed. The recent emergence of large language models, such as the highly functional (yet controversial) ChatGPT, may just be the most visible incarnations of new AI applications. Despite these efforts, the implications of this technological success story for innovation scholars, as well as decision-makers in public and private organizations, are poorly understood.

The paper seeks to contribute to explaining the emergence of AI research and innovation across regions worldwide, focusing on three important elements: scientific paradigms, enabling technologies and diffusion processes. The first refers to the presence of multiple schools of thought, or paradigms, for most of the history of AI (Domingos, 2015; Cardon et al., 2018). Scientific paradigms are frameworks for understanding and approaching complex problems. In the context of AI, paradigms have aided researchers in identifying and solving key challenges, coordinating efforts around often radically different approaches. The connectionist paradigm and the symbolic paradigm, in particular, had a significant impact.

Second, we observe that enabling technologies also play a critical role in the development of AI, providing the necessary hardware and software infrastructure to run and scale AI algorithms. Graphics processing units (GPUs) are the technology that has been a major factor in the development of deep learning. In particular, GPUs have emerged as an unexpected complementary technology that has substantially enhanced the performance of neural networks.

The third element to be considered is diffusion processes, which starting point is often characterized by regional heterogeneity. Regions are likely to have very different starting conditions in the process of adopting and modifying a new technology. Familiarity with related technologies and concepts is likely to give some regions an advantage; the lack thereof will constitute a handicap. Thus, to understand the diffusion of knowledge in the field of AI, it is necessary to consider the interplay between regional specialization towards different paradigms, the command of enabling technologies and the regional endowment with critical human capital (Dosi, 1982; Zucker et al., 1998, 2002).

We postulate that the change in scientific and technological paradigm offers a window of opportunity for different regions to catch up with AI development and innovation. This is, we expect regions initially endowed with higher levels of GPU expertise to be at an advantage. In particular, a strong regional endowment of GPU-savvy researchers (even though most of them were not AI researchers) favored regions that were previously dedicated to the development of GPU technology, for historical and seemingly exogenous reasons. Once the potential of GPUs has been revealed, we expect these regions to shift more readily towards the new scientific paradigm of NNs and related AI innovation activity, particularly compared with regions with a dominant orientation towards AI philosophies other than Neural Networks (NN) (such as symbolic AI).

We provide empirical evidence based on large comprehensive data capturing AI research and innovation activity across regions worldwide. We use functional urban areas (FUAs) as the units of analysis, over 20 years. We first identify AI-related research publications, patents, and startups based on existing databases and established classification criteria (Baruffaldi et al., 2020). We geolocate all publications, patents, startups, and other activities to map them to Functional Urban Areas (FUAs), a geographic delineation that can be applied globally to major urban areas (Dijkstra et al., 2019). We consider the time from 2000 to 2018 for all of these variables. We thus obtain data enabling us to capture the global distribution of AI-related knowledge and innovation activity.

In a Difference-in-differences regression framework, we demonstrate that regions endowed with higher levels of GPU human capital have seen a stronger uptake in AI research and innovation after 2012. We measure human capital relevant to a field of knowledge in a region as the number of active researchers weighted by their stock of publications in that field. In IV analysis, we use the level of human capital related to video game technologies as instruments for the level of GPU-related human capital. This analysis is premised on the technological “disconnect” as GPU technology evolved in the early 2000s without being tied to AI research. Rather, video games have been historically the main area of application of GPUs and we assume that expertise in this area had no direct influence on AI. In addition, we leverage information on papers that combined both AI and GPU before 2010 to exclude from the estimation sample regions that have been pioneering in the attempt to recombine these technologies.

A BRIEF INTRODUCTION OF AI AND GPUS HISTORY

Artificial Intelligence

We present a brief review of the history of AI where we highlight the dynamics that are more closely related to central tenants in the literature on innovation. First, the concept of scientific paradigms is central to a vast literature in the sociology of science and technology development and there is evidence of their role in various technologies (Kuhn, 2012; Dosi, 1982). In AI, the distinction between different approaches is particularly well delineated. Second, it is common that a scientific or technological breakthrough is accompanied by a somewhat unexpected role of existing technologies, which turn out crucial. In the case of machine learning, there is one clear candidate: GPUs.

The two key approaches that can be identified in AI are the "symbolic" approach and the "connectionist" approach or “neural AI” (Cardon et al., 2018). The symbolic approach was the first framework for AI and was used as the basis for developing the ENIAC, the historically

significant first programmable, electronic, general-purpose digital computer. The "connectionist" approach to artificial intelligence is based on early cybernetics and involves conducting parallel calculations across a neural network. This model was later shown to represent learning, especially in visual recognition.

The two approaches have taken turns having the upper hand in technological performance, gaining the attention of the industry and the political sphere. The first period is characterized by the symbolic approach, which was overcome by the connectionist approach. Despite this initial success, the connectionist approach eventually became a minority field within AI research at that time. In the 1970s and 1980s, AI research hit the "AI winter" where funding and interest in the field decreased. However, in the late 1980s, after the "AI winter", the field began to revive with the advent of machine learning, and in the mid-1990s, the connectionist approaches around Geoffrey Hinton once again became dominant in scientific publications, under the banner of deep learning (Cardon et al., 2018).

The success of neural networks and GPUs

Broad access to GPUs allowed for several advancements in deep learning in the late 2000s. GPUs were initially intended for use in game engineering for visualization and computation, as CPUs were not yet sufficiently capable of this task. The costs of GPUs were low at approximately USD 600, and people had already begun experimenting with their capabilities by 2003. In reaction to this development, NVIDIA started to facilitate the programming of GPUs. By introducing a programming system (CUDA), the company paved the way for broad applications of GPUs in high-performance computing. Advancements in deep learning, including those from the early 2000s to 2011, were numerous. Already in 2003, neural networks slightly outperformed other machine learning techniques in certain domains, such as image recognition, with Convolutional Neural Networks achieving new records in a low Credit Assignment Path depth. Around 2006, the term "Deep Learning" was coined when unsupervised pre-training of deep Feedforward Neural Networks helped accelerate subsequent supervised learning through backpropagation. Although research shows that knowledge of GPUs and their potential had been present for some time (Schmidhuber, 2014), it was not until 2012 that the connection to artificial intelligence (AI) became widely recognized.

In 2012, a team of researchers (Geoffrey Hinton, Alex Krizhevsky, and Ilya Sutskever) achieved a major breakthrough in image recognition using a convolutional neural network (CNN) architecture known as AlexNet. They trained a convolutional neural network on a subset of ImageNet, which consists of more than 15 million high-resolution images, divided into over 22,000 categories (Krizhevsky et al., 2012). The breakthrough made them the winners of the highly visible ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The hardware required for this achievement was relatively modest: it consisted of only a PC with two NVIDIA cards (Metz, 2021). The success of AlexNet paved the way for the development of other deep learning models, including recurrent neural networks (RNNs) and generative adversarial networks (GANs).

The application of GPUs in the development of neural networks was unexpected by most, considering their original purpose of accelerating graphics rendering. The scientific community and then the wider public recognized the GPUs' ability to conduct parallel computing only after researchers around Geoffrey Hinton experimented with them for neural network training and took the upper hand in the 2012 ImageNet competition (Metz, 2021).

We conclude our brief history of recent developments in AI by summarizing expectations for our empirical analysis, which will be conducted at the level of urban areas. Our reading of the literature suggests that the vast majority of researchers had not anticipated the results of Krizhevsky et al. (2012). Moreover, regional research on GPU was not yet concentrated at that point on the implementation of neural networks or other AI approaches. But once the potential of GPUs was revealed, those regions that had a historically accumulated advantage in GPU-related human capital should have been able to grow faster in AI outcomes such as patents and startups than other, less advantaged regions. In the next section, we describe the data that will allow us to test these expectations.

DATA

We obtain basic information on conferences, the conference series they belong to, papers listed in their proceedings, and other scientific publications, from the dblp computer science bibliography (dblp). Important to our purpose, dblp also provides a high-quality disambiguation of all authors active in CS (Kim, 2018). We complement dblp with Scopus data on authors' affiliations and abstracts' text. We incorporate information on conference series quality and CS research sub-fields curated by the Computing Research and Education Association of Australasia (CORE). CORE relies on expert assessment to classify all relevant CS conference series into subfields and the quality-rank levels A^* , A , B , and C .

To measure inventive and business activity, we use patent data and information on startup funding date and location, respectively. We obtain patent data from PATSTAT, which is provided by the European Patent Office (EPO). Additionally, we gather data on startup funding from Crunchbase, a prominent platform for startup information and insights. This data source provides us with detailed information on startups, their location, industry, and funding date.

One cornerstone of our data is the identification of scientific and innovative activity related to AI and, within the body of scientific literature, the distinction between Neural AI research as opposed to other AI-related research. For patents, we rely on a classification provided by the Organization for Economic Cooperation and Development (OECD) for identifying patents on AI (Baruffaldi et al., 2020). Notably, although a category for AI and ML-related patents exists in the Cooperative Patent Classification (CPC), it is not precise and, in particular, AI-related inventions may exist outside this classification. Baruffaldi et al. (2020) offer a method based on a combination of CPC-relevant codes and expert-defined keywords to improve this selection. To identify startups operating in AI we rely on classification provided in the Crunchbase database itself. Finally, for conference paper publications, we rely on AI and ML categories as classified by CORE, and, within these categories, we deploy a keyword-based search to identify neural networks and symbolic-related research. Within the larger CS literature, we identify GPU-related and video games-related publications based on keyword search. The dataset comprises information for 2088 distinct regions (FUAs), from the period 1959 to 2022, to which we remove regions for which no single publication in CS is observed over the entire period, for a final number of 2088 regions. The final selection of the data also depends on the consistency of data availability across all variables. We consider the period of observation from 2000-2018 (extended to 1980-2020 for some descriptive evidence). This also guarantees completeness among variables until most recent years, which would otherwise suffer from truncation (i.e. due to more limited coverage of patents and startups). We restrict the sample of regions to those that show some activity in the broad area of CS, based on scientific publications, at least in one year.

ANALYSIS AND RESULTS

In descriptive analyses, we document the ascendancy of neural AI and its association with GPU expertise. Symbolic AI approaches were prevalent in AI literature in the 80s and maintained a sizable share of publications up to the late 90s. Between 2000 and 2010 ML approaches emerged but the relative share of NN publications remains limited. After 2012, AI literature in top conferences doubled over five years and tripled by 2020. This remarkable increase is almost entirely explained by NN research.

We design a DID analysis to study the extent to which regions endowed with a higher level of GPU human capital at the outset of the Neural AI breakthrough gained an advantage in the development of related science and innovation activities. We measure human capital relevant to a field of knowledge in a region as the number of active researchers weighted by their count of publications in that field, up to the focal year.

In IV analysis, we use the level of human capital related to video games' relevant technologies as instruments for the level of GPU-related human capital. As our IV we use the level of Videogame-related human capital in 2010. Video games have been historically the main area of application of GPUs and we assume that expertise in this area had no direct influence on AI. As the history of AI suggests, the discovery of GPUs as an enabling technology for AI emerged at least in part as a discovery and recombination process.

We further consider one main threat to identification. As the history of AI suggests, the discovery of GPUs as an enabling technology for AI emerged at least in part as a discovery and recombination process. As such, some regions may have experimented with GPU technology before 2010 precisely with the intent of resolving existing shortcomings in the AI field. For that, we leverage information on papers that combined both AI and GPU before 2010 to exclude from the estimation sample regions that have been pioneering in the attempt to recombine these technologies. Therefore, we exclude in one set of estimates observations for which the identification assumption is less plausible by excluding pioneering regions in which experimentation with GPUs started very early. We do so by identifying the geolocation of papers that combine both GPU and AI research before 2010. We then refine our estimation sample by dropping the regions in which early experimentation happened and estimate the role of GPU knowledge in the remaining sample of regions for which there was no evidence of anticipation of the adoption of GPUs as an enabling technology for AI.

In our preferred IV model specification, we find evidence that GPU human capital has been instrumental in catching up with Neural AI research as of 2014. This is, regions with higher levels of GPU human capital in 2011 do not show significant trends in Neural AI publications up to 2013 included, compared to regions with lower levels, while they diverge significantly later. Interpreting these estimates as elasticity, their magnitude implies that doubling the GPU human capital endowment in 2011 led to an increase reaching 50% more Neural AI publications in the years up to 2020.

The comparison with the OLS model results is informative and reveals the existence of an anticipation effect when not accounting for the potential endogeneity of GPU human capital. In fact, the related estimates already diverge significantly from the 2000 level immediately. In the following years. This suggests that some regions with high levels of GPU human capital in 2011 actually started taking the lead in Neural AI research already in the prior years. Analyses where we focus on Neural AI research in top conferences as a dependent variable confirm the results

discussed with some interesting variations. In fact, we encounter a faster uptake of regions with high levels of GPU human capital.

CONCLUSION

We show that human capital in general AI research is instrumental to the emergence of AI innovation, but regions endowed with specific knowledge in neural AI and GPU-related knowledge may have gained an advantage unforeseen before 2012. The evolution of scientific developments in AI and the novel recombination with GPU technology allowed for a shift towards neural AI, becoming the main approach and scientific paradigm. GPUs had been previously used in very different applications, such as video games. The work by Krizhevsky et al. (2012), proving GPUs as the enabling technology for the successful deployment of deep learning networks, caused a breakthrough, resulting in a sharp increase of publications on neural AI shortly thereafter. What had been a minority field prior to this publication suddenly became the most prolific AI research ground.

We intend to contribute to the literature on AI and enrich the understanding of how new scientific and technological breakthroughs evolve and determine the geographic emergence of new industries. At times, competing approaches within a given scientific discipline play an important role in advancing its knowledge frontier. As the combinatorial process of ideas and technologies unfolds, novel recombination may prove successful in a seemingly serendipitous manner, disrupting the relative value of different knowledge foundations. Moreover, knowledge related to enabling technologies, tools, and equipment, may become indispensable and scarce. As regions will differ, for historical reasons, in the extent to which they dispose of relevant human capital, their ability to secure and gather complementary knowledge assets and reconfigure their innovation systems will determine their fate in the race for technological leadership.

REFERENCES AVAILABLE FROM THE AUTHORS