Systems of Change:
A Study on the Nature of ICT and AI and
Their Impact on Industrial Trajectories

Dissertation
Zur Erlangung des akademischen Grades
doctor rerum politicarum
(Dr. rer. pol.)

vorgelegt dem Rat der Wirtschaftswissenschaftlichen
Fakultät der Friedrich–Schiller–Universität Jena

am 8. April 2021

von M.Sc. Ekaterina Prytkova
geboren am 6. November 1992 in Rybinsk,
Russische Föderation

Gutachter: Prof. Dr. Uwe Cantner,
Friedrich–Schiller–Universität Jena

Gutachter: Prof. Dr. William Edward Steinmueller,
SPRU, University of Sussex

Gutachter: Prof. Dr. Isabel Almudí,
Universidad de Zaragoza

Datum der Verteidigung: 24. September 2021

Deutsche Zusammenfassung

Die Dissertation beleuchtet technologische Aspekte von Informations- und Kommunikationstechnologien (IKT) und ihrer Rolle bei der Verknüpfung von Industrien sowohl innerhalb als auch außerhalb von IKT-Clustern oder -Systemen. Das übergeordnete Ziel der Arbeit ist es, technologische und systemische Aspekte in ökonomischen Mechanismen zu berücksichtigen, die den Prozess der Digitalisierung in verschiedenen Industrien vorantreiben. Die Arbeit besteht aus vier Hauptkapiteln und hat eine Trichterartige-Struktur; sie beginnt mit der Untersuchung der Makro-Produktivitätsdynamik, welche sich in einer intersektoralen Arbeitsmobilität zeigt, welche sich von Industrien mit erschöpften techno-ökonomischen Möglichkeiten wegbewegt. Die Arbeit fährt fort mit der Untersuchung von zwei Technologiesystemen - IKT und KI - als Quellen neuer technologischer Möglichkeiten und wirtschaftlicher Aktivitäten, welche die zuvor beschriebene Dynamik in Gang setzen. Es ist wichtig, das Ausmaß und die Struktur der Einbettung der IKT auf Volkswirtschaften abzuschätzen und die Natur der KI zu verstehen, um bessere Vorhersagen in Bezug auf ihre Auswirkungen auf etablierte technologische Pfade, Produktivität, Beschäftigung und langfristiges Wachstum machen zu können. So ist die abschließende Studie ein Beispiel für die Vorhersage möglicher zukünftiger Szenarien für die Halbleiterindustrie, deren technologischer Höhenflug durch das Aufkommen von KI unterbrochen wird ("Disruptive Innovation"). Die verwendeten Methoden umfassen: analytische Modellierung, Textanalyse, Produktivitätszerlegung, Produktivitätsanalyse, Netzwerkanalyse und Maße wirtschaftlicher Komplexität wie "Relatedness". Ich arbeitete mit verschiedenen bestehenden Datenbanken wie PATSTAT, REGPAT, COR&DIP, STAN OECD, sowie mit Primärdaten, die durch Webscraping und Textmining gesammelt wurden.

Kapitel 2 der Arbeit beginnt mit einer Makro-zu-Meso-Perspektive, die sich auf Produktivität als Wirtschaftsindikator konzentriert, welche die Auswirkungen sowohl der technologischen als auch der wirtschaftlichen Triebkräfte des Wachstums erfasst. Die Studie deckt 10 Länder ab und schlägt die Kompositionalität des Produktivitätswachstums als neue Erklärung für das

Produktivitätsparadoxon jenseits von säkularer Stagnation und Messproblemen vor. Unter Anwendung einer dynamischen Dekomposition zerlegen wir die Produktivität auf Makroebene in einen wettbewerbsbedingten Effekt, d.h. einen Between-Effekt, und einen technologischen Fortschritt, d.h. einen Within-Effekt, auf der jeweiligen Branchen-Ebene. Unsere Ergebnisse deuten darauf hin, dass die Gesamtdynamik der Produktivität größtenteils durch die so genannte Baumol-Disease getrieben wird, wobei der Anteil der Dienstleistungen (wo die Produktivität naturgemäß begrenzt ist) in einer Volkswirtschaft wächst und das Produktivitätswachstum reduziert. Die positive Entwicklung, aber sinkende Dynamik der technologischen Komponente deutet auf abnehmende Erträge aus innovationsbasierten Verbesserungen hin, eine Tatsache, die wir mit der Möglichkeit einer fortlaufenden Erschöpfung der technologischen Möglichkeiten und einer zeitlichen Verzögerung bis zum Eintreffen eines neuen Schubs, zum Beispiel in Form einer neuen General Purpose Technology (GPT), in Verbindung bringen. Um dies zu untersuchen, ergänzen wir unsere Produktivitätsanalyse mit Belegen für Trends der Innovationsverlangsamung, wobei wir aggregierte und kompositorische Trends betrachten. Wir untersuchen die Innovationsverlangsamung anhand einer Reihe von Indikatoren, die auf dem Konzept der "Ideen-TFP" basieren, und zeigen, dass es eine verallgemeinerte Evidenz für das gleichzeitige Auftreten von Produktivitäts- und Innovationsverlangsamungen gibt.

In Kapitel 3 werden die Informations- und Kommunikationstechnologien näher betrachtet, da sie für die großen strukturellen Veränderungen des 20. Jahrhunderts verantwortlich sind, indem sie die Produktivitätsdynamik befeuerten. Rückblickend ist der Einfluss der IKT immens und offensichtlich, aber komplex und oft nicht-linear in seiner Ausbreitung. Dies könnte die Bedeutung der IKT verschleiern oder im Gegenteil überhöhen. Eine Abhilfe ist die Betrachtung von IKT anstelle eines groben Monolithen als ein Cluster von unterschiedlichen, aber miteinander verbundenen Technologien. Kapitel 3 nimmt diese systemische Sichtweise von IKT auf, um ein feinkörniges Netzwerk von Verbindungen zwischen Branchen und verschiedenen IKT-Technologien zu konstruieren. Die Studie versucht, Muster in der Dynamik der industriellen Durchdringung durch IKT über den Zeitraum von 1977 bis 2020 in 28 EU-Mitgliedsstaaten zu identifizieren. Methodisch besteht der IKT-Cluster aus 13 patentbasierten Gruppen von Technologien, die an-

schließend mit 74 Industrien mittels Text Mining und der Algorithmic Links with Probabilities (ALP) Methode verknüpft werden. Verschiedene Netzwerkmaße und Relatedness-Indikatoren werden angewandt, um Industrie-Technologie-Matrizen zu konstruieren, die Einblicke in die Verbreitung von IKTs und deren ähnlichkeit in Bezug auf Wissens- und Anwendungsgrundlagen geben. Um die Lücke in der Literatur zur ökonomie der künstlichen Intelligenz (KI) zu schließen, wird ein besonderer Fokus der Analyse auf die KI-Technologie innerhalb des IKT-Clusters gelegt, indem diese Technologie in einen Kontext gestellt und KI in Beziehung und im Vergleich mit anderen Technologien untersucht wird.

Künstliche Intelligenz ist unter den IKT-Clustern das "nächste große Ding" in Bezug auf die erwarteten Auswirkungen auf Produktionsprozesse, Beschäftigung und der Durchdringung des Konsums. Angesichts der hohen Investitionen in KI, welche auf dem Spiel stehen, ist das richtige Verständnis der Technologie entscheidend, um ihre Vorteile zu nutzen und ihre Risiken zu mindern. In Kapitel 4 versuchen wir daher, den am besten geeigneten Rahmen zu finden, um die wesentlichen, aber komplexen Eigenschaften von KI zu erfassen, den Ursprung neuartiger und ungesehener Effekte zu beschreiben die sie hervorruft sowie Unternehmen und politischen Entscheidungsträgern Ratschläge zu geben, die auf der voraussichtlichen Entwicklungsrichtung von KI basieren. Wir argumentieren, dass es, obwohl KI einige Merkmale einer GPT aufweist, ein Fehler sein könnte, sie als solche zu bezeichnen; anstelle eines eigenständigen, komponentenartigen Charakters einer GPT scheint KI alle Chancen zu haben, sich zu einem sogenannten Großen Technischen System (LTS) mit infrastrukturähnlichem Charakter zu entwickeln. In beiden Fällen generiert KI bereits jetzt und auch in Zukunft komplementäre Innovationen und neuartige Anwendungen. Die Art und Weise, wie diese Prozesse ablaufen und verwaltet werden können, unterscheidet sich jedoch erheblich, wenn KI als LTS und nicht als GPT behandelt wird. Ein unvollständiges Verständnis und Missmanagement von KI-Technologien könnte zu suboptimalen Situationen mit beeinträchtigtem Wettbewerb (hohe Konzentration in Märkten, die KI-Komponenten liefern), fehlender effektiver Kompatibilität und Fragmentierung, Verletzung der Privatsphäre, erhöhter Umweltbelastung und sogar einem neuen "KI-Winter" führen, der alle KI-getriebenen Möglichkeiten für wirtschaftliches Wachstum zum Stillstand bringt.

Aufbauend auf meinen vorherigen Arbeiten untersuche ich in Kapitel 5 die Beziehung zwischen den zwei Bereichen, welche die KI-Technologie ausmachen: Software und Hardware. Nach Jahrzehnten stetiger Entwicklung entlang des technologischen Pfades mit dem Mooreschen Gesetz im Zentrum erlebt die Halbleiterindustrie derzeit einen exogenen Schock: den Aufstieg moderner KI-Technologien, deren zugrundeliegende Berechnungslogik sich von der Mainstream-Logik unterscheidet, die in der Mehrzahl der Algorithmen realisiert wird. Damit steht die Halbleiterindustrie vor einer fundamentalen Herausforderung, die Innovationen in der Chip-Architektur erfordert, um neuartige technische Spezifikationen zu adressieren und dennoch die hohen Herstellungs- und Produktionskosten mit den potenziellen Einnahmen in Einklang zu bringen. Die kommende Zeit ist für die Halbleiterindustrie wirklich anders, da sie noch nie in großem Umfang von der von-Neumann-Architektur, dem dominierenden Design der Industrie, abgewichen ist. Die Verbindung zwischen KI und der Halbleiterindustrie, die im dritten Kapitel empirisch aufgedeckt und im vierten Kapitel begründet wird, ist in Bezug auf die technologischen Input-Output Beziehungen zirkulär und unterliegt vielfältigen techno-ökonomischen Trade-offs. Dies führt zu Nichtlinearitäten und Unregelmäßigkeiten bei der Bereitstellung von technologischem Know-how durch diese Bereiche, was sich auf die Produktivitätsdynamik in Industrien, die eine der beiden Technologien einsetzen, auswirken kann. Bisher wurde die Volatilität der Software-Hardware-Bindung aufgrund der hohen Konzentration in beiden Branchen und damit ihrer relativen Stabilität in Bezug auf die Anzahl der Akteure, die Vielfalt der Produkte und die Entwicklungsmuster weitgehend verschleiert. Die Erhöhung der höchstmöglichen Dichte von Elementen auf einem Chip bei minimalen Kosten war der prominenteste techno-ökonomische Kompromiss, den die Chiphersteller eingegangen sind. Das Aufkommen der KI und die sich nähernde Grenze der Miniaturisierung brachten die Halbleiterindustrie aus dem Gleichgewicht und durchbrachen die etablierten technologischen Pfade. Wie in Kapitel 4 beschrieben, stellt dies eine umgekehrte Entwicklung dar, die ihren Ursprung in der Hardware-Domäne hat und auf die Software-Domäne (Algorithmen) des KI-LTS zurückwirkt. Insgesamt liefert die ausführliche Fallstudie in Kapitel 5 einen überblick über die Mechanismen und Faktoren, welche die Halbleiterindustrie in der Vergangenheit sowohl auf der Angebots- als auch auf der Nachfrageseite geprägt haben sowie über neu entstandene Kräfte. Die Ana-

lyse führt zur Konstruktion eines industriellen Organisationsmodells und möglicher Szenarien für die Entwicklung der Halbleiterindustrie.

Die Arbeit liefert einen Beitrag zur ökonomie der Digitalisierung und zur ökonomie der KI als jüngste Ausprägung der IKT; sie gehört auch zum Strang der Literatur über Technologiesysteme, da sowohl IKT als auch KI als solche betrachtet werden. Schließlich trägt die Arbeit aufgrund der Betonung der dynamischen und technologischen Aspekte zur ökonomie des technologischen Wandels bei.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"In short, computer technology offers the possibility of incorporating intelligent behavior in all the nooks and crannies of our world. With it, we could build an enchanted land."

— Alan Newell

1.1 A System-level Approach in Digital Economics

Contemporary economies and societies are increasingly reliant on the concept of information and the technologies that deal with it. The discovery of information properties and the possibility to encode, transmit, process, and use it to control physical devices was a true scientific revolution which, in turn, has paved the way to a technological revolution. In fact, "the conversion of analogue data and processes into a machine readable format" (OECD, 2019), or *digitisation*, has enabled the introduction of radical innovations such as computer platforms (Bresnahan and Greenstein, 1999) and eventually Information and Communication Technologies (ICT). Decades of development and diffusion of ICTs begot new economic activities and have changed the way old ones are performed bringing the so-called digital transformations upon them.

The economic study of information, ICTs, and digital technologies cuts through micro, meso and macro levels of analysis: from information asymmetries and expectations, to features of production processes of information goods and services such as near-zero marginal costs (Shapiro et al., 1998) and associated novel business models, to eventually productivity dynamics, automation and long-term growth. The ubiquitous diffusion of information and digital technologies has spawned new digital products, business models, and marketplaces. However, from a pure economic viewpoint, the study of digital economics does not require brand new economics. As Goldfarb et al. (2019) point out, "(s)imple microeconomic models with zero marginal cost are not so different from models with positive marginal cost. The demand curve slopes downward and firms price where marginal revenue equals zero". Digital technologies change the relative cost structure of activities such as search, replication, transportation, tracking and verification, and in doing so they shape incentives structure, which in turn affects economic and innovative conducts and performance.

The radical novelty of information technologies lies in the technological side of the equation, namely in the specific functions they execute, and the transformative power resides in their systemness, the fundamental complementarity among building blocks making ICTs a fabric that underlies socio-economic processes at every level of (dis)integration. From this perspective, ICTs constitute the fundamental *infrastructure* at the core of the current technological paradigm, defined by Perez (2010) as "a collectively shared logic at the convergence of technological potential, relative costs, market acceptance, functional coherence and other factors". Despite the inherent complementarity among ICTs and their dual importance, both economic and technological, ICTs have been studied from a *systemic* perspective to a very limited extent. As the technological side tends to be omitted in economics studies, this led to representation of ICT as a coarse, faceless monolith. To an even smaller extent, research effort has been invested in studying distinct ICT technologies in the broader context of the whole ICT system. Our understanding of the immense impact of ICT largely consists of studies that are snippets from a broader network of economic activities, actors, products and services related to ICT.

The recent breakthroughs in Artificial Intelligence (AI) renewed the attention of economists to the ICT system. AI is a field of research that combines computer science with other disciplines such as engineering, neurobiology, psychology, linguistics, philosophy and ethics but it also represents the latest wave of ICT innovation. The diffusion of AI as a technology is tied to AI's value as a commodity and utility for businesses and final consumers, and, at the same time, AI's value is a function of what AI as a technology can do. Thus, the technological aspect is instrumental to rationalise the mechanisms AI sets in motion and understand its economic impact. While the identification of what intelligence is remains a complex territory, the field of *artificial* intelligence "is concerned with intelligent behaviour in artefacts" (Nilsson, 2009); in other words, the creation of virtual or physical machines capable of executing tasks usually associated with natural intelligence. The modern state of AI is called "weak AI" for being "*idiot savant*" excelling only at a particular task and with no understanding of what it does (Dartnall, 1994). For this reason, AI algorithms remain confined within the prediction (elaboration of information) stage of a task and haven't permeated yet into judgement and decision-making (Agrawal et al., 2017). As with ICTs, the expectations, prediction and paradoxes emerging regarding the uses and impact of AI might stem from a partial or coarse representation of AI, inflating its technical capabilities, neglecting limitations and focusing only on the most prominent part of this *system* technology. The dissertation magnifies over this coarse view, and unpacks ICTs and AI as structured technologies, or "systems of change".

1.2 Structure of the Dissertation

The dissertation emphasises the role of the technological side of ICTs in establishing industrial links both within and beyond ICT cluster or system. The overarching aim of the dissertation is to bring technological and systemic aspects into the economic mechanisms driving the process of digitalisation in different industries. The dissertation consists of four chapters, each providing increasing magnification along the level of analysis. Chapter 2 looks at the macro productivity dynamics that emerges from the industry

level and is rooted in technology flows and the production of innovation. Chapter 3 focuses on ICT technology system as a source of influential technologies that affect productivity by studying relatedness and pervasiveness of distinct ICTs. Chapter 4 and Chapter 5 are devoted to AI, as a particular ICT, rationalising its nature and investigating the mechanisms that govern the interaction between the two domains AI is build upon: hardware and software.

The dissertation employs a wide range of methods, which include analytical modeling, text analysis, productivity decomposition exercise, efficiency analysis, network analysis, and metrics of economic complexity such as relatedness. The author worked with various sources of secondary data such as PATSTAT, REGPAT, COR&DIP, STAN OECD, as well as with primary data collected through webscraping and text mining.

The author developed the chapters benefiting from the scientific guidance provided by colleagues at the Chair of Microeconomics, Friedrich Schiller University Jena and by fellow researchers during a visiting period at the Science Policy Research Unit (SPRU), University of Sussex. The author presented the results of her research in a number of internal seminars (the Jena Economic Research Workshops and the Jena Summer Academies on Innovation and Uncertainty), invited workshops and seminars (at the Universities of Sussex, Strasbourg and the AI3SD Network+ of the University of Southampton) and international conferences, among which the EMAEE Conference 2017 in Strasbourg, the Summer School on "Knowledge Dynamics, Industrial Evolution, Economic Development" in Nice, the 17th ISS Conference "Innovation, Catch-up, and sustainable development" in Seoul, and the EMAEE Conference 2019 in Brighton.

1.2.1 Chapter 2

Chapter 2 of the dissertation, titled "The Compositional Nature of Productivity and Innovation Slowdown", studies the sources of the widely discussed slowdown in productivity growth and returns on research effort by identifying how much the industrial composition of economies drives such

outcomes. The analysis starts with a macro-to-meso perspective focusing on an economic indicator such as productivity that captures the impact of both technological and economic drivers of growth. We split these two dimensions with the decomposition exercise to uncover persistent patterns of the corresponding productivity components and interpret these patterns through the lens of technological input-output relationship. The chapter suggests to go beyond secular stagnation and mismeasurement problems as an explanation of the productivity slowdown. We argue that the downward macroeconomic trend of productivity growth emerges from the aggregation of diverse industry-level productivity trends and because of more general technological transformations transmitted along knowledge and technology flows among groups of industries classified according to Pavitt taxonomy (Pavitt, 1984). In turn, these technological changes associated with either nearing or just accomplished transition between techno-economic paradigms can manifest themselves through decreasing returns on innovative activities. To investigate what we call "innovation-productivity nexus", we complement our productivity analysis with evidence on innovation slowdown trends using an array of indicators based on the notion of "idea-TFP" (Bloom et al., 2017) and show that there is a consistent evidence for productivity and innovation slowdowns co-occurrence.

Among the literature on productivity that dates back to 1980s, many explanations of the productivity paradox had been suggested ranging from mismeasurement to decreasing dynamism, from stagnation of technological progress to almost opposite argument of so quickly renewed capital that it causes implementation lags. Compositionality of the productivity's growth as an explanation for the slowdown is a unique contribution of this chapter. By classifying industries according to Pavitt taxonomy, the composition of productivity growth consists of industry groups that reflect sources and recipients of technologies and knowledge. Covering ten countries allows estimating if there is a systematic correlation between an industry group and its contribution to the productivity growth. This novel approach introduces an additional dimension related to technologies and knowledge flows that helps to unpack further the productivity paradox.

1.2.2 Chapter 3

In Chapter 3, titled "ICT's Wide Web: a System-Level Analysis of ICT's Industrial Diffusion with Algorithmic Links", I delve into the ICT cluster as the locus of the influential technologies that fuel productivity fluctuations. ICTs have been frequently viewed as engines of growth for the industries implementing them (Bresnahan and Trajtenberg, 1995). In retrospect, the impact of ICT is immense and visible but complex and often non–linear in how it propagates. These complexity and non–linearity could obscure ICT's significance or, on the contrary, exaggerate it. One remedy is the consideration of ICT as a set of distinct but interrelated technologies, instead of a coarse monolith. This chapter adopts this systemic view of ICT to construct a fine–grained network of connections between industries and distinct ICT technologies. Using the terminology of complexity economics, ICTs reside in the realm of capabilities approximated in the Chapter 3 with the technological knowledge base; industries represent economic activities that rely on different combinations of knowledge. Mapping economic activities onto capabilities remotely echoes the identification of sources and receivers of technological knowledge in Chapter 2: in both cases the rationale behind the analysis is the localisation of the origins of technological changes and the directions of their diffusion along linkages established with dependent industries. The difference lies in the focus of Chapter 3 on a subset of knowledge base represented by ICTs, but studying ICT quantitatively as a technology system to address complexity and achieve a greater level of detail in the understanding of ICTs' pervasiveness.

For what concerns the industry-ICT relation, by highlighting the heterogeneity of the ICT system, the Chapter allows placing each ICT in context by looking at the scale, scope and pace of diffusion of not only each individual technology but also simultaneously of the cluster of ICT technologies in comparison with each other. In a way, this reveals the composition of the ICT system by identifying which technologies are pervasive and which are key technologies. The chapter proceeds with the estimation of within-cluster relations among ICTs using relatedness metrics, deepening the analysis of the ICT cluster. Technological relatedness and the newly introduced application relatedness represent two distinct media through which ICTs might

be connected. Apart from further refinement the understanding of ICT as a heterogeneous system rather than an undifferentiated monolith, this provides a useful framework to identify potential loci of adoption externalities and drivers of development for each pair of ICT technologies.

Among the ICT cluster, the Chapter places a special focus on AI technology. The aim is to look at AI in perspective and construct a more informed representation of the technology. On the one hand, AI is a novel, fast–growing technology that enters the commercial phase and is subject to intensive development and experimentation. With digital infrastructure widely deployed (Greenstein, 2019), AI has a great potential for automation, enables multiple applications, and already generates billions in revenues. On the other hand, studying AI's development trajectories, commercial value and directions of diffusion without considering related technologies and industries risks to form a misleading representation of AI, inflate expectations at this stage and overestimate AI's aggregate growth prospects (Bresnahan, 2019a).

The Chapter builds on studies that focus on the economic impact of ICT (Brynjolfsson and Hitt, 2000; Van Ark et al., 2003; Brynjolfsson et al., 2019), and offers a contribution by tracing back the source of this impact to particular ICT technologies. Another contribution lies in the field of sectoral patterns of innovation (Malerba, 2002; Castellacci, 2008) by showing the structure of industrial connections through shared technological knowledge base. In this context, the relatedness indicators represent an instantiation of the research on the principle of relatedness (Hidalgo et al., 2018) once again underscoring the multiplicity and importance of linkages between economic activities and technologies. As the analysis takes a closer look at AI technologies, the Chapter is a contribution to the Economics of AI (Agrawal et al., 2019b). In particular, to the best of the author's knowledge, this work is the first to study together AI diffusion among industries' knowledge base and the complementarity of AI with other ICT technologies.

1.2.3 Chapter 4

Present among the ICT cluster, Artificial Intelligence (AI) is poised to be "the next big thing" in terms of expected impact on production processes, workplace, and pervasiveness of consumption. Given the large bid on AI at stake, a proper understanding of the technology is crucial to harness its benefits and mitigate its risks. Thereby, in Chapter 4, we zoom even further on AI and apply to it two frameworks – General Purpose Technology (GPT) and Large Technical System (LTS) – to compare their goodness of fit with implications on the governance of AI technologies.

The Chapter starts with an overview of GPTs and the identification strategies to recognise a technology as a GPT. This is required for the subsequent mapping of AI onto GPT characteristics. The rationale behind this exercise is that since the Deep Learning (DL) breakthrough scholars settled too quickly with the GPT interpretation of AI, mostly because of its multiple applications. However, what constitutes GPT is more than many uses, and the Chapter provides a thorough analysis of "the AI equals GPT" formula along micro characteristics and macro effects characterising a GPT. In making the case for the AI LTS, we gather pieces of evidence from the studies on, among others, impact of AI on labor, production processes and automation, technology diffusion, strategic management, and we provide our own empirical findings. As an intermediate result, we arrive at the conclusion, drawing a parallel with econometrics, that GPT is a misspecified model of AI: some factors are over– or underestimated, relevant ones are excluded while irrelevant included, and connections among building blocks are partially or incorrectly accounted for. These problems might lead to poor predictions of pace, directions and conditions of the technology's diffusion, contributing to uncertainty with regard to estimations of scope and scale of AI's impact. The latter is an important issue for the governance of AI by public regulators as well as by companies producing and using AI technologies.

An alternative approach to understanding AI discussed and applied in the Chapter is Large Technical System (Hughes et al., 1987; van der Vleuten, 2009). As the name suggests, the offered perspective to look at AI is of a systemic kind rather than a stand-alone artefact. This builds on the ideas

and findings of Chapter 3, where AI is considered in the context of the whole ICT cluster and related technologies are identified. The LTS framework, despite sharing some aspects with GPT, captures what we claim is an essential property of AI and that the GPT interpretation misses, namely the *infrastructural nature* of the technology. Furthermore, the LTS framework equips a researcher or a policy-maker not only with a knowledge of the technical components that constitute AI, but also with an encompassing mapping of the whole circuit of actors, system's fault lines, size and boundaries of the system, as well as AI's development phases and corresponding driving forces at each stage.

The position of this Chapter in the literature is on the crossroad of two fields: economics of AI and economics of technological change. On the one hand, it provides a "thick description" of AI revealing its structural and relational properties as a system. This representation exposes mechanisms, factors and problems obscured in the GPT framework. On the other hand, the main driver of studying AI is its potential for profound socio-economic transformations as it could initiate a new industrial revolution, re-domaining the economy, and doing so produce "great surges of development" and a shift to a new techno-economic paradigm (Perez, 2010). As AI's expected impact is an ultimate reason for interest in the technology, misrepresentation of AI is likely to fail in estimations of its effects. Thus, the key contribution of the Chapter is an expansion of perspectives on how AI is viewed, from an individual component to an infrastructure, which we believe will help to construct correct expectations of AI.

1.2.4 Chapter 5

Building on my previous work, in the last Chapter I study the relationship between two domains that constitute AI technology: software and hardware. After decades of steady development along the technological trajectory with the Moore's law at the core, the semiconductor industry is currently experiencing an exogenous shock: the rise of modern AI technologies, whose underlying computation logic is different from the mainstream one realised in the majority of algorithms. Thus, the semiconductor industry faces a

fundamental challenge that calls for innovation in the chips architecture to address novel technical specifications and yet balance high costs of fabrication and production with potential revenues. This time is genuinely different for the semiconductor industry, as it has never departed at scale from the von Neumann architecture, the industry's dominant design.

The connection between AI and the semiconductor industry, empirically detected in the second chapter and rationalised in the third chapter, is circular in terms of technological input-output and subject to multiple techno-economic trade-offs. This results in non-linearity and irregularities in delivering technological know-how by these domains, which might reverberate on productivity dynamics in industries that implement either of the two technologies. Before the current AI shock, the volatility of the software-hardware bond has been largely obfuscated because of high concentration in both industries, which ensured their relative stability with regard to set of players, variety of products and patterns of development. Increasing the highest possible density of elements on a chip *at minimum costs* was the most prominent techno-economic compromise made by chipmakers. The arrival of AI and the approaching of the physical limit of miniaturisation disrupted the semiconductor industry breaking established technological trajectory. Using terms introduced in Chapter 4, this situation represents a reverse salient that originates in the hardware domain and reverberates on the software (algorithms') domain of the AI LTS. Chapter 5 is an in-depth case study that provides an overview of the mechanisms and factors that shaped the semiconductor industry in the past from both supply and demand side as well as newly emerged forces. The analysis results in the construction of an industrial organisation model and the derivation of two potential scenarios for the development of the semiconductor industry.

The Chapter is a multidisciplinary study as the line of arguments is built upon diverse literature: (i) the strategic management of semiconductor firms (Burgelman, 2002; Gawer and Henderson, 2007) and research on platform products (Baldwin and Clark, 2000), (ii) AI and computer science (Russell, 2019; Hooker, 2020) as well as computation theory and integrated circuits design (Borkar and Chien, 2011), (iii) the economics of network products and software as a supporting service (Church and Gandal, 1992; Chou and Shy,

1993) and (iii) the economics of technological change, industrial dynamics, and systems of innovations, as we study the forces that support and contest the established technological trajectory (Dosi, 1982) of chip production (Steinmueller, 1992) and the factors driving the evolution of the semiconductor industry (Malerba et al., 2008; Brown and Linden, 2011; Adams et al., 2013).

1.3 Final Overview

The dissertation has a funnel-like structure; it starts with a study on macro productivity dynamics captured with intersectoral labor mobility moving away from industries with depleted techno-economic opportunities. The dissertation continues with the research on two technology systems – ICT and AI – as the sources of new technological opportunities and economic activities setting in motion the dynamics described before. It is important to estimate the magnitude and structure of the ICT's "grip" over the countries' economies and rationalise AI's nature to construct better predictions with regard to their impact on established technological trajectories, productivity, employment and long-term growth. Thus, the concluding study is an example of envisaging potential futures for the semiconductor industry, whose technological trajectory is disrupted by the emergence of AI.

The dissertation provides a contribution to the economics of digitalisation and the economics of AI as ICT's latest instantiation; it also belongs to the strand of literature on technology systems as both ICT and AI are considered as such. Finally, given the emphasis placed on dynamics and the systemic and technological aspects, the dissertation contributes to the economics of innovation and technological change.

Dissertation Summary

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Title	The Compositional Nature of Productivity and Innovation Slowdown	ICT's Wide Web: a System-Level Analysis of ICT's Industrial Diffusion with Algorithmic Links	Artificial Intelligence's New Clothes? From General Purpose Technology to Large Technical System	On the Basis of Brain: Neural-Network-Inspired Change in General Purpose Chips
Co-authors	Uwe Cantner, Holger Graf, Simone Vannuccini	-	Simone Vannuccini	Simone Vannuccini
Focus	System's effects	System's properties	System's properties	System's effects
Theoretical foundation/ Field	Productivity slowdown, industrial dynamics, replicator dynamics, economics of innovation	Economics of ICT, technology system, technology diffusion, complexity economics	Economics of AI, large technical system, general purpose technology, science and technology policy	Economics of AI, network industries, industrial organisation, technological trajectories, economics of innovation
Methodology	Non-parametric dynamic decomposition of labor productivity	Text analysis, Algorithmic Links with Probabilities, Network analysis, Relatedness	Individualising comparative analysis of theoretical frameworks	IO model: augmented Hotelling with network effects
Level	Macro-meso	Meso, System	System	Meso-micro
Spatial coverage	10 OECD countries	EU28	-	-
Observation period	1970-2015	1977-2020	-	-
Data	STAN OECD	PATSTAT, REGPAT, primary (text mining)	Primary (web scraping)	COR&DIP
Data type	Panel	Constructed cross-section over time	Constructed cross-section over time	Cross-section over time
Contribution	Significant contribution to the design of the study, data collection, theoretical elaboration and empirical analysis and interpretation of results		Significant contribution to the design of the study, data collection, theoretical elaboration and empirical analysis and interpretation of results	Significant contribution to the design of the study, data collection, theoretical elaboration and empirical analysis and interpretation of results
Status	Has been reviewed in Journal of the European Economic Association. Available as working paper in JERP: No. 2018-006	Under review in Research Policy. Available as working paper in JERP: No. 2021-005	After revisions submitted in The Journal of Information Technology. Available at SSRN: Vannuccini&Prytkova (2021)	Second round of revisions in Industrial and Corporate Change. Available as working paper in SWPS: Prytkova&Vannuccini (2020)

Chapter 2

The Compositional Nature of Productivity and Innovation Slowdown

2.1 Introduction

A growing number of studies identify a generalized slowdown in labor productivity growth (Syverson, 2017). This trend, coupled with evidence of decline in the pace of business dynamism of US firms (Decker et al., 2014) ignited a series of academic debates, revolving around two main issues. The first, more macroeconomic in nature, confronts the hypothesis of secular stagnation (Teulings and Baldwin, 2014) with the 'mismeasurement' one (Syverson, 2017); the related debate is summarized in the confrontation between, respectively, 'techno–pessimists' and 'techno–optimists' (Gordon, 2016; Mokyr, 2014). The second issue, with a more microeconomic flavor, has to do with the 'black box' of the nature of productivity (Syverson, 2011; Bartelsman, 2010); in this context, scholars are starting to uncover how firms' heterogeneity and 'granularity' (Gabaix, 2011) and uncorrelated shocks at the micro–level reverberate up to the macro–level (through the structure of production networks) and produce aggregate dynamic effects — including on productivity.

In this Chapter, we take a position in–between these two corners of the research on productivity dynamics, as we address questions usually related to the macroeconomic side of the debate while at the same time shifting the focus to the slightly more granular 'meso'–level of industries. From the macroeconomic perspective, the analysis of the slowdown is usually based on aggregated measures or on productivity decompositions assessing the contribution of the different production factors (in a source–of–growth framework of analysis) or macro–sectors like IT–producing and IT–using. Here, the arguments to explain productivity slowdown usually echo the classical 'Solow productivity paradox' and suggest that IT–related industries are not producing the expected productivity gains or, when productivity growth occurs, it is driven by a faster decline in the denominator of the productivity ratio, rather than in any increase in output or efficiency gain[1]. Therefore, a more in–deep analysis of the structural patterns leading to productivity slowdown is needed.

In the Chapter, we combine two potential explanations of the productivity slowdown and we posit that (a) the composition of aggregate productivity matters, and that (b) productivity dynamics is rooted into more general technological dynamics. To support the first argument, we decompose productivity at the meso level of analysis to show the heterogeneous contribution of industries resulting in aggregate productivity slowdown. To support the second argument, we employ an array of indicators to detect a potential 'innovation slowdown'. Finally, we relate these two phenomena exploring what we label the 'innovation–productivity nexus'.

To deal with (a), we dissect the structural composition of the productivity slowdown, checking if it emerges from the aggregation of diverse — and with different weights — industry–level productivity trends. In this first step, we engage in an assessment of the determinants of the slowdown; while complementary research explores firm–level determinants of productivity dynamics (Bartelsman, 2010) and the role played by skill–biased technical change

[1] According to (Acemoglu et al., 2014, p.399), when a differential productivity growth driven by IT–related industries is detected, "...it is driven by declining relative output accompanied by even more rapid declines in employment. It is difficult to square these output declines with the notion that computerization and IT embodied in new equipment are driving a productivity revolution, at least in U.S. manufacturing."

(Acemoglu and Autor, 2011), ours is an exercise in detection of structural dynamics. Indeed, we focus on capturing the underlying distribution of meso trends and changes resulting in the aggregate productivity slowdown.

As concerns our methodological contribution, in order to understand the compositional nature of the aggregate productivity slowdown, we decompose the growth of labor productivity for a sample of 10 OECD countries using non–parametric decomposition techniques (Cantner and Krüger, 2008; Foster et al., 2001; Castaldi, 2009). By doing that, we assess (i) if the generalized productivity slowdown is pulled by the productivity dynamics in a subset of industries and (ii) what is the role played by structural change — namely by the reallocation of labor between industries.

To deal with (b), we limit our insight on the determinants of the slowdown to a conjecture about the role played by profound technological transformations. In fact, the technological (supply–side) explanation of the productivity slowdown should not be overlooked. Indeed, a generalized exhaustion of technological opportunities, one for example characterizing the transition between techno–economic paradigms (Dosi, 1982) could affect the pace of productivity growth.[2] In this sense, the productivity slowdown could be the proximate outcome of a technological transition yet in the making. This argument goes in line with the recent contribution of Bloom et al. (2017) that suggests how research productivity is not constant but decreasing and finds evidence of these decreasing returns of research (and employed researchers) for the whole U.S. economy at a rather disaggregate level. An alternative — though related — explanation is that a technological transition has already taken place, but has yet to show its effects due to implementation lags similar to those to be expected during the establishment of a novel general purpose technology (Brynjolfsson et al., 2019; Cantner and Vannuccini, 2012); according to both views, either a new profound transformation has just occurred or has yet to happen, the productivity slowdown is mirroring decreasing returns in innovative activities within the current established direction of technological development.

To investigate the idea that an exhaustion of technological opportunities or

[2] "In other words, whenever the technological paradigm changes, one has got to start (almost) from the beginning in the problem–solving activity." (Dosi, 1982)

that delayed adoption may be a major driver of the productivity slowdown, we complement our decomposition of aggregate productivity growth with an analysis of trends in innovation–related variables. Being aware that the dynamics and turbulence of the economic and technological domains may not be perfectly synchronous (Cantner and Krüger, 2004), we look for a comparable slowdown in the 'productivity' of innovative activities — what we label the 'innovation slowdown'.

We find evidence that (i) the compositional nature of the productivity slowdown is a common trend in the countries under analysis; that (ii) industry–specific productivity improvements prevail in their magnitude over the effect of structural change; that (iii) the trend of industry–specific productivity improvements is, in general, a declining one. Findings (i)–(iii) suggest that the aggregate productivity slowdown is a result of heterogeneous contributions of constituting industries. Regarding technological dynamics, we find (iv) there is a generalized (across OECD countries) and compositional innovation slowdown taking place in parallel with the productivity slowdown. All the evidence points to the possibility that the slowdown is driven by the exploitation of established technological opportunities (or implementation lags) in knowledge–intensive industries, coupled with structural shifts of economic activities towards services.

The originality of our contribution is threefold: first, we apply decomposition techniques at the industry level; while this is not an absolute novelty (see Holm (2014) and Castaldi (2009), who conducted a similar analysis to assess the role played by manufacturing and services in determining aggregate labor productivity), we are the first to link productivity decompositions and the productivity slowdown literature. In doing so, we apply methods and notions used in the literature on market selection to a novel domain. Second, we exploit the most up–to–date available datasets to gain a throughout understanding of recent productivity dynamics. Third, we combine descriptive evidence on the productivity and innovation slowdowns to explore the innovation–productivity nexus.

The Chapter proceeds as follows. Section 2.2 discusses the debate revolving around the productivity slowdown. Section 2.3 introduces our exploratory

analysis and discusses the findings. Section 2.4 extends our analysis to the innovation slowdown and relates the two phenomena. Section 2.5 concludes.

2.2 Productivity Slowdown: What We Know, and What We Don't

Recently, a number of studies has detected a slowdown in labor productivity growth, starting in the year 2004 (Syverson, 2017; Fernald, 2015). The beginning of this trend predates the beginning of the economic crisis and the great recession, and cannot be fully explained by market 'bubbles'–related arguments. Furthermore, the slowdown is experienced by most of advanced economies, thus it is not a phenomenon confined to the US. Fernald (2015) estimates that productivity growth trends returned to 1973–1995 level pace, after an acceleration in the period 1995–2004, considered as an 'aberration' driven by IT producing and using industries. Therefore, the relevant question is, have advanced economies entered a phase of 'new mediocre' (Dabla-Norris et al., 2015), marked by the exhaustion of the positive influence on productivity generated by IT diffusion? In other words, have "the 'low–hanging fruit' of IT–based innovation had been plucked" (Cette et al., 2016, p.15)?

To begin answering the question above, we now overview the main positions in the debate on the nature of the productivity slowdown. A first position suggests that the productivity slowdown is just the result of a mismeasurement problem (Brynjolfsson and McAfee, 2014; Mokyr, 2014) and that the task of economists is to estimate correctly the gains produced by a new generation of products and services that are profoundly shaping the economy (e.g. digitization–related goods, Internet and social networks).

A second position rejects the previous idea, pointing out that even the larger estimates that take into account the mismeasurement won't cover the observed output gap generated by the productivity slowdown, hence pointing to the possibility that what hides behind the slowdown is a deeper structural transformation. Such a position is coupled with evidence suggesting

a decreasing dynamism (in terms of new ventures formation, workers flow and job creation/destruction) of economic actors in the US, especially for what concerns young firms (Decker et al., 2014), therefore reinforcing the view that the economy is transitioning to a growth plateau.

A third position, pooling together a smaller set of more heterogeneous studies, connects decreasing economic dynamism and productivity slowdown with intensity of governmental regulation (Goldschlag and Tabarrok, 2014) and the level of macro–prices, in particular lower interest rates that allow for the survival of less–efficient firms and, hence, reduce resource reallocation in some European economies (Cette et al., 2016). We direct our attention to the first two positions, as the third set of explanatory variables have been found less cogent and significant and, in any case, it deviates from the focus of this Chapter on structural dynamics.

Addressing the mismeasurement hypothesis, Syverson (2017) suggests that the mismeasurement

> "could take one of two related forms in the data. One would occur if a smaller share of the utility that these products provide is embodied in their prices than was the case for products made before 2004. If this were true, measured output growth would slow even as total surplus growth continued apace. The second form of mismeasurement would occur if the products' price deflators were rising too fast (or falling too slowly) relative to their pre–2004 changes. This would understate quantity growth as backed out from nominal sales." (Syverson, 2017, p.2)

. However, when confronted with the data (for US, in this case), any estimation of the size of the output gap due to the productivity slowdown compensates only for a fraction of the counterfactual missing output that would have been generated if the productivity growth trend would have persisted at the pre–2004 level. Syverson points out further reasons that play down the importance of the mismeasurement hypothesis: first, the trend of productivity slowdown is not confined to the US (see also Cette et al. (2016) for France); second, even the least conservative estimates of Internet–generated

consumer surplus cannot account for the trillion dollars magnitude of the output gap; third, the correction of productivity mismeasurement for ICT–related industries would imply a five–time increase in their revenues (and six–fold increase in their value–added), which is hardly justifiable (Syverson, 2017, p.3); fourth, mismeasurement has always occurred. The last point offers a nice argument in favor of the idea that the productivity slowdown can be related to technology non–linear dynamics and adoption lags; in fact, the mismeasurement hypothesis is, in a sense, the re–statement of the Solow paradox with respect to the last wave of IT–related innovations.

While the mismeasurement hypothesis seems not to provide a robust explanation for the labor productivity growth slowdown, Byrne et al. (2017) show that the mismeasurement of high–tech products prices does play a significant role in explaining the patterns and sectoral distribution of multi–factor productivity (MFP) growth (though, not of aggregate MFP growth). However, this evidence only deepens the productivity puzzle, as the upward corrections of the mismeasurement in high–tech sector MFP growth — suggesting a faster pace of technological development for some types of economic activity — do not reverberate into labor productivity growth, that continues to slow down.

Taking stock from the above discussion, the possibility that current slowdown is rooted in long term tendencies rather than in measurement inaccuracies is not to be played down. This has induced a surge in studies on the more structural nature of the phenomenon, creating a divide between techno–pessimists, broadly claiming that technology advances will not alter the cap to economic growth posed by structural dynamics of advanced economies, and techno–optimists, offering a brighter perspective on the role technological change will have in fostering future growth. In particular, a debate has been fostered by the publication of the book 'The Rise and Fall of American Growth' by Robert Gordon (Gordon, 2016). There, the broad issue under analysis is what has been labeled 'secular stagnation' (Teulings and Baldwin, 2014). With secular stagnation — a term firstly introduced by the Keynesian scholar Halvin Hansen — two phenomena are usually conflated together: one on the demand–side (Summers, 2015), and the other on the supply–side.

The supply–side version of secular stagnation is the one discussed by Gordon, that camps on the more techno–pessimist side. Two dynamics are playing against future economic growth. On the one hand, there is the end of a historical phase of 'great inventions' ranging from the American civil war to the 1970s. This goes in line with the evidence of non–linearity of technological progress. On the other hand, the emergence of several 'headwinds' is threatening future growth prospects. These headwinds are not rooted in technological dynamics, but have to do with increasing inequality, decreasing returns to education, and aging of population. Crafts (2016) and Clark (2016) reinforce this side of Gordon's argument building up on the evidence that it is the growth of total factor productivity, rather than labor productivity in general, that is slowing down. The reason for that has to do with the massive transition of advanced economies to services. The supporters of the headwinds role as growth constraint highlight how

> "... (a) surprising share of modern jobs are the timeless ones of the pre–industrial era — cooking, serving food, cleaning, gardening, selling, monitoring, guarding, imprisoning, personal service, guiding vehicles, carrying packages. Food production and serving, for example, now employs significantly more people (9.1 percent) than do production jobs (6.6 percent). One in ten workers is employed in sales. The information technology revolution to date has left these jobs largely untransformed. Workers in these types of jobs in Europe in 1300, if transplanted to modern America, would need little retraining." (Clark, 2016, p.68)

Hence, technological changes do not necessarily have to be the *prima movens* of current economic trends.

Arguments about the end of the period of great innovations and the increasing burden of headwinds on advanced economies are played down by techno–optimists scholars. They point out that if "... some inventions are more important than others" (Gordon, 2016, p.72), as Gordon claims, nothing prevents new technological revolutions from arriving in the future and 'rejuvenating' technological opportunities and unleashing growth by creating a new techno–economic paradigm (Perez, 2010). In fact, the literature

on general purpose technologies (Cantner and Vannuccini, 2012) pushes further the argument that the arrival of new pervasive and enabling technologies and productivity dynamics are strictly intertwined, as slowdowns and accelerations in the rate of productivity growth follow the adjustment of the economy to such 'macro–inventions' (Mokyr, 1990). Furthermore, the severity of aging as a headwind blowing against growth is questioned. Indeed, Acemoglu and Restrepo (2017) suggest that aging produces no effects on GDP per capita growth due to endogenous responses of technology, where the negative effect on growth of exiting labor force is neutralized by the adoption of robots.

Finally, it is worth mentioning that if the productivity slowdown depends on the transition of advanced economies to services, digitized and immaterial activities, then the problem at stake can be considered a sort of revised version of the classic 'Baumol disease' (Baumol, 2012), according to which productivity improvements in service industries cannot be pushed further indefinitely. The structural transformation of advanced economies, in this sense, should naturally lead to a productivity growth slowdown given the relatively decreasing weight of manufacturing in the overall value–added generation.

To sum–up, the detected trend of productivity growth slowdown is found in the literature not to be caused by the Great Recession of the recent years (while the crisis could have amplified the slowdown effect by uncovering structural weaknesses in the economies), and most likely not to be due to severe mismeasurements of the welfare gains derived from the digitization of the economy. The productivity slowdown can be a transient and short–term phenomenon, the outcome of advanced economies' restructuring, or a symptom of supply–driven secular stagnation. As already mentioned, the main line of reasoning behind the supply–side view of secular stagnation is that innovations are not all alike. Ironically, while recognizing the heterogeneity among technologies, the studies we cite do not fully take on board the heterogeneity among economic activities that can lead to the observed aggregate trends. In fact, the productivity slowdown may be a generalized trend for the whole economy, or a statistical artifact due to the compensation of trends going in different directions at more disaggregated levels of

analysis: stagnation can 'bite' the whole economy, or just some industries, with different magnitudes. For this reason, a structural perspective on the productivity slowdown is necessary to identify the industry–level sources of productivity growth dynamics.

Put differently, our idea is to investigate, in the field of productivity and innovation, what the classical Harberger (1998) study about mushroom– and yeast–like processes investigated for output growth dynamics. Indeed, our task is to understand if the 'data generating process' behind the productivity slowdown has a mushroom (localized) nature — that is, is generated by few industries — or rather a yeast (generalized) one, where all industries contribute uniformly to the slowdown, and if a common pattern exists across countries. Napoletano et al. (2006) already pointed out using a theoretical model how aggregate output growth is a compositional construct deriving from the combination of sectoral output growths, and how this combination depends on the magnitudes of cross–industry demand elasticities. This renders problematic the empirical identification of what they call 'pure general purpose technology processes' (where the aggregate growth is the result of uniform shifts in growth at the industry level) and 'pure idiosyncratic processes' (where aggregate growth is derived from industry–specific shocks whose reverberation intensities are functions of cross–industry elasticities).

In what follows, we take up on the issues raised above by looking at the compositional nature of productivity slowdown.

2.3 The Compositional Nature of Productivity Slowdown

2.3.1 Theory

To rationalize our argument, we theoretically connect our expected skewed distribution of industry–level contributions to aggregate productivity dynamics to retardation theory (Metcalfe, 2003). Retardation has been found to be a stylized empirical fact in the evolution of industries; with retardation

we mean "the systematic tendency for rates of growth of specific entities or their ensemble to decline with the passage of time" determining "secular or long time movements in the volume of economic activity" (Metcalfe, 2003, p.412). In a nutshell, retardation theory suggests that the composition of an aggregate growth rate matters, as the rates of growth of its parts are heterogeneous and vary in time. In our case, the aggregate growth rate is the one of labor productivity, that we decompose in that of industry groups. Interestingly, retardation theory is conceptually connected to population dynamics, where the growth rate of a given characteristic in a population depends on the structure of the population itself in terms of individual heterogeneity with respect to this characteristic. Usually, such population dynamics are modeled using the replicator dynamics model (Metcalfe, 1994) to approximate the working of market selection. Indeed, unlocking the dynamics of an aggregate indicator like productivity is an exercise in understanding economic evolution. Following (Holm, 2014, p.1011), "evolution is the change in the mean characteristic of a population"; in our case, productivity is the characteristic taken into consideration, and its (weighted) mean the evolving indicator of interest.

The economic validity of the replicator dynamics is usually tested by applying indirect methods, such as decomposition techniques (Cantner and Krüger, 2008). Non–parametric decomposition techniques are commonly used in studies of productivity dynamics at the micro (firm) level (see Melitz and Polanec (2015); Cantner et al. (2016) for a review). Usually, either the productivity level is decomposed Olley and Pakes (1992), or the productivity change Foster et al. (2001); Metcalfe and Ramlogan (2006). The latter method, also labeled dynamic decompositions (as opposed to static decompositions of productivity levels) is the one we use in this study. Our novel contribution stands on testing the replicator model through decomposition at the level of analysis of industries.

The theoretical rationale for the decomposition runs as follows. The aggregate growth rate of any relevant economic variable can be expressed as a composite indicator weighting the growth rate of the variable for each component of the aggregate and their respective proportion (share) in the aggregate; this can be expressed as $\sum s_i g_i$, where for each component g_i

of the aggregate the growth rate g is weighted by its proportion s_i. Given that, the dynamics (the change over time) of the aggregate variable can be decomposed to highlight different elements contributing to the change; by taking the time derivative of g we obtain

$$\frac{\partial g}{\partial t} = \sum \frac{\partial s_i}{\partial t} g_i + \sum s_i \frac{\partial g_i}{\partial t} \tag{2.1}$$

In this generic version of the decomposition, we identify two components. The first term on the RHS of the equation captures how aggregate the growth rate changes due to the changing proportion of the elements in the aggregate, that is a how much structural reallocation affect aggregate dynamics. The second term captures how the change in the growth rate of a component contributes to changes growth of the aggregate. While there may be several ways to decompose and aggregate a variable of interest (Metcalfe and Ramlogan, 2006), we consider this distinction between a reallocation force and an idiosyncratic force a first illustrative way to identify the structural properties of a given process.

With the theoretical decomposition in mind, we now outline our empirical decomposition. Following Cantner and Krüger (2008), we use a decomposition formula that identifies three components, the so–called within, between, and covariance (or cross–level) components (or effects). The intuition behind the different effects is the following: at the firm level, the within effect is interpreted as learning/innovation (change in productivity between periods ceteris paribus the firm's market share, corresponding to the second term in 2.1), the between effect is interpreted as a measure of reallocation and market selection (the change in firms' market share ceteris paribus the productivity of the previous period, compared to a benchmark, corresponding to the first term in 2.1), and the covariance (the co–movement between periods of firms' productivity indicator and market share) as a proxy for the regime of returns (increasing, decreasing, constant returns to scale) in a particular market. In our case, having industries instead of firms as units of analysis, the within component can be interpreted as the specific industry contribution to productivity change; the between component indicates how reallocation of economic activity (e.g. workers) between industries — that

is structural change — affect aggregate dynamics and should not be interpreted as a degree of competition between industries (in this sense, it would rely on a ill–posed conceptualization of competition). Firm–level productivity decompositions (Cantner and Krüger, 2008) usually include additional components accounting for entry and exit dynamics. As the time–span of our study is not enough to observe full obsolescence of industries nor fine–grained enough to observe the entry of completely new economic activities, we rule out these components. Holm (2014) extends the framework of decomposition techniques to account for multi–level selection, that is, a separation amongst reallocation due to firm–level dynamics and to industry–level dynamics. Focusing on industry–level data, our analysis does not allow to follow the same path; however, it is worth stressing the importance of accounting for different levels of economic activity.

Studies of productivity slowdown already engage in decompositions; however, they usually decompose labor productivity change into the contribution of Human Capital (education), capital deepening, and TFP (Gordon, 2016, p.73). The result of this kind of decomposition is usually to highlight how the slowdown in productivity growth is depending almost uniquely on the decrease of TFP. The rate of capital deepening is decreasing in the period 2011–2014 due to capital devaluation driven by the economic crisis; however, as pointed out by Fernald (2015), it is TFP growth that slows down starting from 2004. This evidence allowed scholars to infer that the nature of productivity slowdown was technological — as TFP remains, besides 'a measure of our ignorance' (Abramovitz, 1989) — also a proxy for technological efficiency. Another kind of distinction in the literature is the one between ICT producing and ICT using industries (Fernald, 2015), which, however, does not provide an in–depth view on the possibly heterogeneous contributions to productivity growth across the industry structure. With our decomposition, we extend these kinds of analyses by looking at the contribution to productivity change due to industries' improvements (as a proxy to measure the magnitude and direction of technological opportunities) and to structural change.

Table 2.1: Dataset construction

Country	Industries available	From	To	Time span (years)	Data source
Austria	60	1976	2015	40	ISIC 4 SNA08
Czech Republic	62	1994	2015	22	ISIC 4 SNA08
Denmark	63	1970	2015	46	ISIC 4 SNA08
Finland	63	1975	2015	41	ISIC 4 SNA08
Germany	58	1991	2014	24	ISIC 4 SNA08
Italy	58	1992	2014	23	ISIC 4 SNA08
Netherlands	63	1988	2011	24	ISIC 4 SNA93
Norway	55	1975	2014	40	ISIC 4 SNA08
Sweden	50	1993	2014	22	ISIC 4 SNA08
US	30	1987	2010	24	ISIC 4 SNA93
Mean	54	-	-	29.73	-
Minimum	30	1970	2010	20	-
Maximum	63	1994	2015	46	-

2.3.2 Data

We conduct our analysis using data for 10 countries retrieved from two versions of the OECD Structural Analysis (STAN) database (ISIC v.4 SNA08 and ISIC v.4 SNA93). Facing the trade–off between, on the one hand, the number of countries to be included in the sample and, on the other hand, the level of detail of the data and the available time–span, we opted for an analysis of the most recent available data, in order to better shed light on the compositional effects driving productivity dynamics. The use of the most up–to–date data is another original contribution of this Chapter compared, for example, with the similar analysis conducted by Castaldi (2009). We use production (nominal output) and employment tables to determine labor productivity for a selection of industries at the maximum level of disaggregation available. The classification of industries covers a wide spectrum of economic activities from agriculture/natural resources to air and spacecraft, including both manufacturing and services. In order to provide the most comprehensive analysis in terms of periods covered, the time span and set of industries available vary among the countries considered. Our dataset construction is summarized in Table 2.1.

We set our analysis at the level of aggregation of industries because, despite a detected heterogeneity of economic actors at every level of disaggregation (Dosi and Nelson, 2010) —- industry as a level of analysis captures a fair share of variability in economic and innovative behavior, more than firm size and market structure alone do (Cohen, 2010).

In order to ease data interpretation and visualization, we group the industries according to two different taxonomies. First, we follow a novel OECD taxonomy of economic activities based on industries' R&D intensities (Galindo-Rueda and Verger, 2016); this taxonomy results in six groups of industries, that in turn are separated along a manufacturing/non–manufacturing line in order to take care of possible structural differences in behaviors and nature between manufacturing and non–manufacturing economic activities even when ranking similarly in terms of R&D intensity. Formally, in the following these industry groups are numbered from 1 (high R&D intensity) to 6 (low R&D intensity) and with an additional Figure taking the value of 1 for groups of manufacturing industries and 0 for groups of non–manufacturing industries. Thus, for example, group 1.1 stands for high R&D intensity manufacturing industries.

Second, we grouped industries according to the Pavitt taxonomy (Pavitt, 1984). The identified Pavitt groups are, for manufacturing, supplier dominated (SD), scale intensive (SI), science based (SB), and specialized suppliers (SS). For services we distinguish between supplier dominated services (SDS), scale intensive services (in turn divided into physical network (PN) and informational networks (IN)), and knowledge intensive business services (KIBS). In order to attribute the available industries to a particular Pavitt group, we initially followed the assignment proposed in Castaldi (2009), which already introduced the identification of services groups alongside manufacturing ones. However, in practice, as the set of ISIC v.4 industries we use has a different structure and contains additional branches with respect to previous ISIC versions, for the generic industries not included in the work of Castaldi we estimated and assigned the groups to which these industries belong using on the employment shares of their sub–industries as weights. This procedure led to slightly different classifications of industries in Pavitt classes among the countries in our dataset. For instance, being

Table 2.2: Frequencies of R&D-intensity groups

Country	1.0	1.1	2.0	2.1	3.0	3.1	4.0	4.1	5.0	5.1	6.0	6.1	Industries available
Austria	1	3	1	4	0	4	7	10	26	0	4	0	60
Czech Repuplic	1	3	1	4	0	4	8	11	26	0	4	0	62
Denmark	1	3	1	4	0	4	7	8	31	0	4	0	63
Finland	1	3	1	4	0	4	7	10	29	0	4	0	63
Germany	1	3	1	4	0	4	7	8	26	0	4	0	58
Italy	1	3	1	4	0	4	7	8	26	0	4	0	58
Norway	1	2	1	3	0	3	8	7	26	0	4	0	55
Sweden	0	2	1	4	0	4	4	8	23	0	4	0	50
Netherlands	1	3	1	4	0	4	7	8	31	0	4	0	63
US	0	2	0	4	1	3	2	8	9	0	1	0	30

SDS 'D77T82: Administrative and support service activities' industry has the biggest share of employees in 'D69T82: Professional, scientific and technical activities; administrative and support service activities', which allows the former to prevail over the other types and makes the latest also to be considered SDS. This holds true, for example, in Germany, while in Denmark the entry D69T82 is classified as KIBS due to the fact that in this country the entry 'D69T75: Professional, scientific and technical activities' belongs to KIBS and dominates over D77T82 in terms of the number of employees involved. Thus, each of 118 industries in the dataset was assigned to a particular Pavitt group either by definition or by calculating the mean of employment shares estimation if the industry is composed by different sub–industries belonging to different Pavitt groups.

In order to provide a clear picture of how industries are distributed across our two types of classification (OECD R&D–intensity and Pavitt), Tables 2.2 and 2.3 report the frequency of industries falling under each group. Furthermore, Tables 2.16–2.19 provide a summary of our assignment to a R&D–intensity and Pavitt group for each industry potentially available in the dataset (recall, however, that not every entry is used for each country).

2.3.3 Methodology

For the analysis, five years moving averages of the basic variables are computed in order to calculate labor productivity and its growth rate. Ro-

Table 2.3: Frequencies of Pavitt taxonomy groups

Country	SD	SI	SS	SB	SDS	IN	PN	KIBS	Non-market services	Industries available
Austria	12	10	3	1	8	10	8	4	4	60
Czech Republic	12	12	3	1	8	10	8	4	4	62
Denmark	10	10	3	1	12	10	9	4	4	63
Finland	12	10	3	1	12	8	9	4	4	63
Germany	10	10	3	1	8	10	8	4	4	58
Italy	10	10	3	1	8	10	8	4	4	58
Norway	10	8	3	0	8	10	8	4	4	55
Sweden	10	10	3	0	9	8	5	1	4	50
Netherlands	10	10	3	1	12	10	9	4	4	63
US	7	10	3	0	4	3	2	0	1	30

Note: Pavitt categories. SD: supplier dominated; SI: scale intensive; SB: science based; SS specialized suppliers; SDS: supplier dominated services; PN: physical networks; IN informational networks; KIBS: knowledge intensive business services.

bustness checks are performed using one– and three–years moving averages. After the construction of the variables for the analysis, we perform the decomposition exercise; following Cantner and Krüger (2008) and Cantner et al. (2016) we apply the following decomposition formula to aggregate productivity:

$$\Delta \bar{a}_t = \sum_i s_{i,t-\tau} \Delta a_{i,t} + \sum_i \Delta s_{i,t} \left(a_{i,t-\tau} - \bar{a}_{t-\tau} \right) + \sum_i \Delta s_{i,t} \Delta a_{i,t}, \qquad (2.2)$$

Where t is a time index, τ is a generic term capturing the period over which the change in productivity is calculated, and i indexes the different industries. As the decomposition equation holds for all the countries under analysis, we drop the country index in order to ease reading.

Having performed the decomposition, we aggregate the values of the within and between effects using our two grouping rules (R&D intensity and Pavitt) and normalize the data in two ways: (i), we calculate the magnitude of the effects per worker; (ii) we provide a normalization in line with Cantner and Krüger (2008), where the size of the effect becomes a percentage of the of the labor productivity level of the previous period.

Table 2.4: Average yearly percentage point change in labor productivity

Country	Period of analysis	Annual percentage delta LP
US	1991-2010	-0.039
Germany	1995-2014	-0.046
Sweden	1997-2014	-0.154
Norway	1979-2014	-0.049
Netherlands	1992-2011	-0.042
Italy	1996-2014	-0.191
Finland	1979-2015	-0.095
Denmark	1974-2015	-0.003
Czech	1998-2015	-0.115
Austria	1980-2015	-0.052

2.3.4 Analysis and Discussion of the Results

First of all, we give a closer look at the general development of labor productivity growth for the ten countries under consideration. Doing that, we verify that the data used shows the very features of productivity slowdown which is related to a long–run tendency of productivity growth rates to decline — if not even to become negative.

Table 2.4 shows for each country the respective time–span average change in labor productivity growth rates (in percentage points). These numbers corroborate the aggregate evidence of productivity slowdown. However, they are averages taken over all industries, whose contribution to the aggregate dynamics could be heterogeneously distributed, as we claimed in the theory Section.

In fact, the percentage point changes in labor productivity growth are rather different among the different industries (see Table 2.5). Looking at productivity growth trends disaggregated by (Pavitt) groups in Table 2.5, there are commonalities for all countries involved with respect, for example, SD and SI, which show always a declining tendency, nearly also so in SS. In the other classes we find a mixed picture. It appears that in classes related to manufacturing there is a rather general decline whereas in classes related to services we find both directions.

Next, we report the results of analysis for both R&D–intensity and Pavitt–based industry groupings using the Cantner and Krüger (2008) normaliza-

Table 2.5: Average change in labor productivity, percentage points per year. Pavitt taxonomy

Country	Period of analysis	IN	KIBS	NMS	PN	SB	SD	SDS	SI	SS
US	1991-2010	-0.043	-	0.015	-0.079	-	-0.114	-0.066	-0.044	0.051
Germany	1995-2014	-0.180	0.096	-0.041	-0.020	-0.417	-0.130	0.059	-0.059	-0.151
Sweden	1997-2014	0.058	-0.076	-0.099	-0.112	-	-0.277	-0.171	-0.076	-0.331
Norway	1979-2014	0.060	-0.168	0.016	0.054	-	-0.029	0.015	-0.262	-0.016
Netherlands	1992-2011	0.007	-0.145	-0.023	0.005	-0.265	-0.002	-0.132	-0.013	-0.068
Italy	1996-2014	-0.152	-0.356	-0.038	-0.277	0.046	-0.215	-0.243	-0.070	-0.114
Finland	1979-2015	-0.003	-0.055	-0.028	-0.056	-0.050	-0.141	-0.082	-0.098	-0.187
Denmark	1974-2015	-0.017	0.048	0.130	0.030	-0.070	-0.038	-0.049	-0.082	-0.036
Czech	1998-2015	-0.080	0.007	0.005	-0.017	0.145	-0.126	-0.139	-0.249	-0.446
Austria	1980-2015	-0.0618	-0.0752	-0.0355	-0.0256	-0.1394	-0.0248	-0.1004	-0.0741	0.0004

Note: Pavitt categories. SD: supplier dominated; SI: scale intensive; SB: science based; SS specialized suppliers; SDS: supplier dominated services; PN: physical networks; IN informational networks; KIBS: knowledge intensive business services.

tion of the data. Furthermore, we limit our discussion to the within and between effects, as the interpretation of the covariance effect at levels of aggregation above the firm level is not straightforward. Figures 2.1, 2.2 2.3 plot the within and between effects for the Pavitt grouping. Figures 2.4, 2.5 and 2.6 do the same for the R&D–intensity industry classification. Figures 2.7 to 2.10 provide cross–correlations of the effects at the country level to gain some insights on the commonalities of the patterns observed.

First, we focus on the Pavitt taxonomy. We find that the within effect is generally positive (apart from cases in Denmark, Italy, and Norway, especially during the crisis 2007/8 and its aftermath) and it ranges in a common interval (0;0.06) for all the countries under consideration (with the exception of Norway[3]). In order, the stronger magnitudes of the effect are found for SI, IN (excluding Italy and Norway), SDS, SD, SS. As for countries cross–correlations (displayed in Figure 2.7), these are positive and quite strong for Austria, Czech Republic, Denmark, Finland, Germany, Italy and Netherlands, while the USA, Norway and Sweden stand aside and also they are not correlated among each other.

The between effect also ranges across countries in a rather small common interval (-0.06;0.03); it assumes negative values (or close to zero) for SS and fluctuates (Austria, Czech Republic, Denmark, Finland, Germany, Italy) or it is negative (Netherlands, Norway, Sweden, USA) for IN. SI has both

[3]The peculiarity of Norwegian productivity dynamics might be related to the country specialization in oil production; as it is beyond the scope do this Chapter, we do not explore the determinants of Norway within and limit ourselves to track its changes.

Table 2.6: Industry productivity benchmarking with respect to average level

Country	IN	NMS	PN	SI	SB	SDS	SS	KIBS	SD
Germany	1	-1	-1	1	1	-1	1	+1 to -1	-1
Italy	1	-1	-1	1	1	-1	1	+1 to -1	-1 to +1
Netherlands	1	-1	-1	1	1	-1	1	-1	1
Finland	1	-1	-1	1	1	-1	1	+1 to -1	-1 to +1
Czech	1	-1	-1	1	1	+1 to -1	-1 to +1	1	-1
Denmark	1	-1	-1	1	1	+1 to -1	-1 to +1	1	-1 to +1
Sweden	1	-1	-1	1	-	-1	1	-1	1
US	1	-1	-1	1	-	-1	-1 to +1	-	+1 to -1
Norway	1	-1	-1	1	-	+1 to -1	-1 to +1	1	-1 to +1
Austria	1	-1	-1	1	1	+1 to -1	-1 to +1	+1 to -1	-1

Note: Pavitt categories. SD: supplier dominated; SI: scale intensive; SB: science based; SS specialized suppliers; SDS: supplier dominated services; PN: physical networks; IN informational networks; KIBS: knowledge intensive business services. The indicators +1, 1 and −1 show if the Pavitt group's labor productivity is respectively above, equal or below average productivity (aggregate) over all the time period of analysis.

the biggest negative and positive contribution in all countries. Finally, the between effect is usually positive for SD. Other Pavitt groups do not show detectable patterns for the between component, and country cross–correlations (see Figure 2.8) are more scattered. To interpret the between effect, that is given by the changes in sales/employment shares 'weighted' by the deviation of an industries labor productivity from the average labor productivity, the position of an industry or class in this ranking is important. Table 2.6 indicates for each country and each Pavitt class whether the level of labor productivity is below (-1) or above (+1) the average productivity over the full time span of observation. In most cases, these rankings (with respect to the average) are stable over time; for some classes pattern of development form below to above average (-1 to +1) or the other way round (+1 to -1) can be identified. In most of the Pavitt classes, the position towards the average productivity level in the economy is the same over all countries (IN; NMS; PN, SI, SB) or come to close to that (SDS, SS). A mixed picture is found for KIBS and SD. For the SI class, we usually find negative between effects in all countries. Table 2.6 tells that SI always has a productivity level above average. This implies for the within effect that SI experiences a decline in sales/employment share, hence loses economic importance. This, of course, contributes negatively to productivity growth dynamics.

Second, we look at the R&D intensity classification. The within effect is

generally positive (excluding Denmark, Italy and Norway) and with similar magnitudes. Groups 5.0 and 4.1 have biggest effect (only positive since within is generally positive), excluding Sweden, Norway, Italy where 5.0 is generally negative. For the between effect, again 5.0 and 4.1 groups are the sources of both biggest negative and biggest positive effect (except for Norway). However, different from the Pavitt grouping, country–level effects cross–correlations (Figures 2.9 and 2.10) are weak.

From this first decomposition analysis, we summarize the following:

1. productivity slowdown is empirically corroborated, and it displays a compositional nature;
2. within effects prevail in magnitudes over between effects;
3. within effects are usually positive;
4. the negative or small magnitudes of the within effect for SS coupled with positive values for groups of industries that are usually receivers of technological know–how (SD, SDS but also industry groups 4.1 and 5.0) go in the direction suggested in the theoretical part of the study: productivity improvements derive mostly from the 'plucking of the low–hanging fruits' of ICT transformations, while the industries that are 'classic' sources of deep economic transformations are currently contributing less to productivity growth, suggesting either an exhaustion of technological opportunities or a temporary lag due to re-focusing and implementation of new techniques;
5. the latter trend is technology–driven and not country–specific (even though there are specific cases), as a very similar pattern appears to hold across countries;
6. the between effect is negative in some cases; the principle of market selection we applied at the industry level seems not to hold strictly, as a negative between effect means either that less productive industries gain labor shares or that highly productive industries lose market shares. This is true, for example, for SI industries. While a negative between effect would represent a puzzle in studies of firm–level market

shares reallocation, it is an evidence that is much easier to rationalize in our context of structural change. Indeed, industries with higher productivity levels can, on the one hand, contribute less to productivity growth (an evidence captured by the within effect) and, hence, explain the slowdown due to their decreasing returns in the transformation of technology into productivity improvements. On the other hand, having a productive level higher than the Economy's average can result from processes of automation and skill–biased technical change that expel labor and produce structural change. Hence, the slowdown of productivity growth may also be affected by a structural dynamics of industrial transformation where labor moves to less productive industries (e.g. services), an evidence in line with our formulation of the slowdown problem as a novel reading of the Baumol disease;

7. the fact that within effects prevail over between effects allow us to claim that the resulting structural changes are not solely demand–driven; in fact, they are more likely to be innovation–driven, otherwise the within effects, capturing industries' idiosyncratic learning and improvements, would have negligible magnitudes.

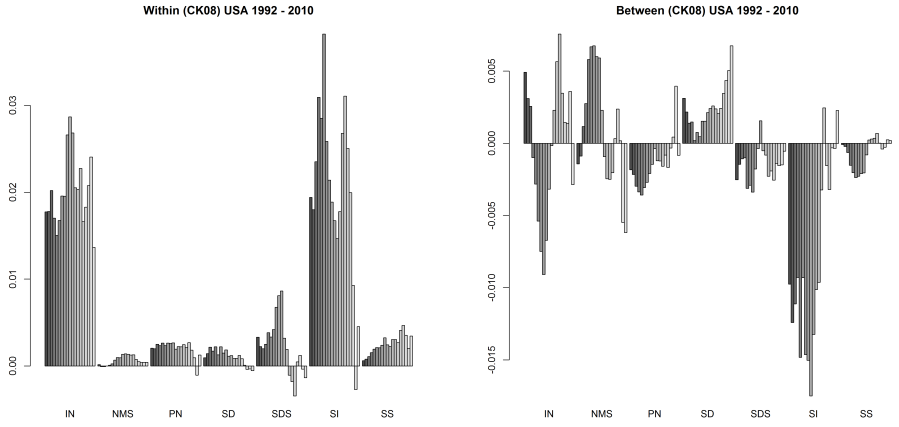


Figure 2.1: Between and Within effects, USA, 5–years moving average

Note: Labels on the x–axis represent Pavitt's industry types.

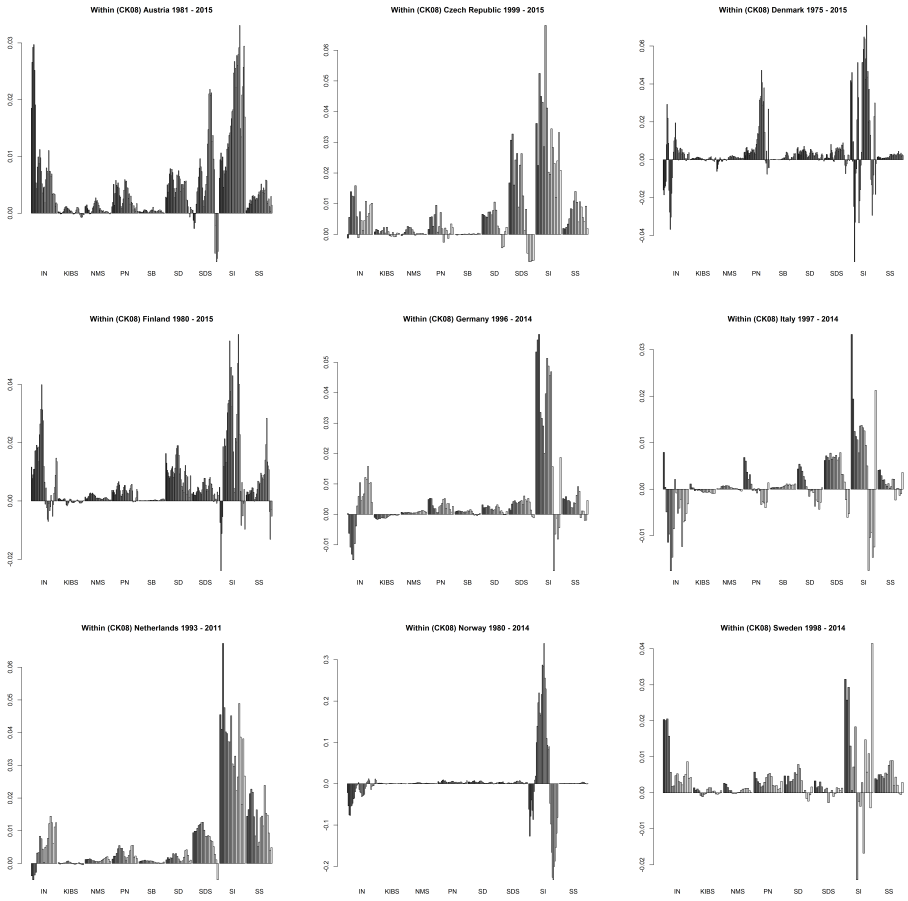


Figure 2.2: Within effect for 9 OECD countries, 5–years moving averages, Pavitt groups

Note: Labels on the x–axis represent Pavitt's industry types.

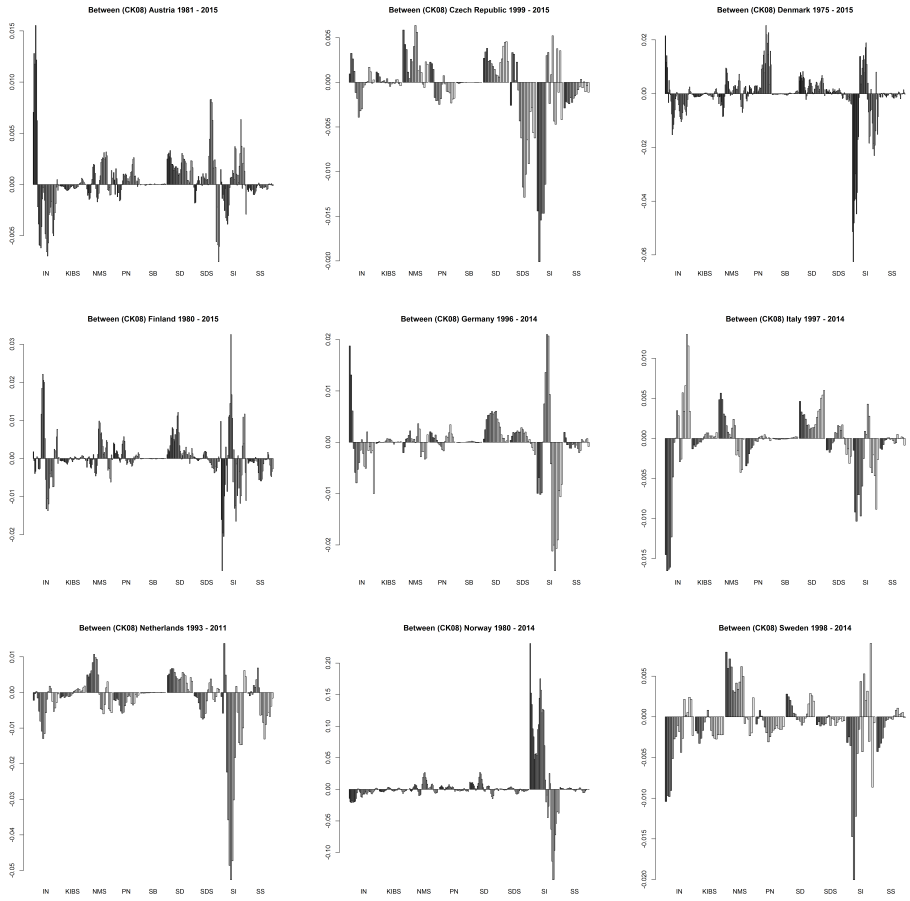


Figure 2.3: Between effect for 9 OECD countries, 5–years moving averages, Pavitt groups

Note: Labels on the x–axis represent Pavitt's industry types.

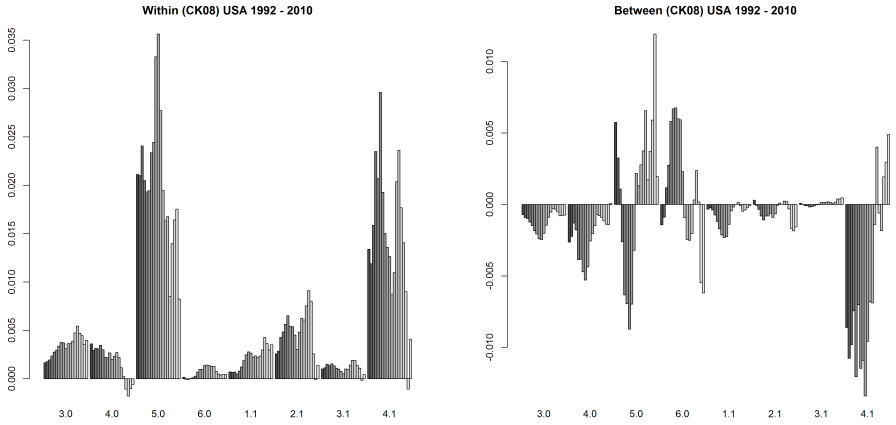


Figure 2.4: Between and Within effects, USA, 5–years moving average, R&D–intensity groups.

Note: Values on the x–axis code the R&D classification of industries, ranging from 1 (highest intensity) to 6 (lowest intensity). Decimal values indicate if the group is belonging to manufacturing (.1) or non–manufacturing (.0). For example, 1.0 indicates the group pooling highest R&D intensity non–manufacturing industries.

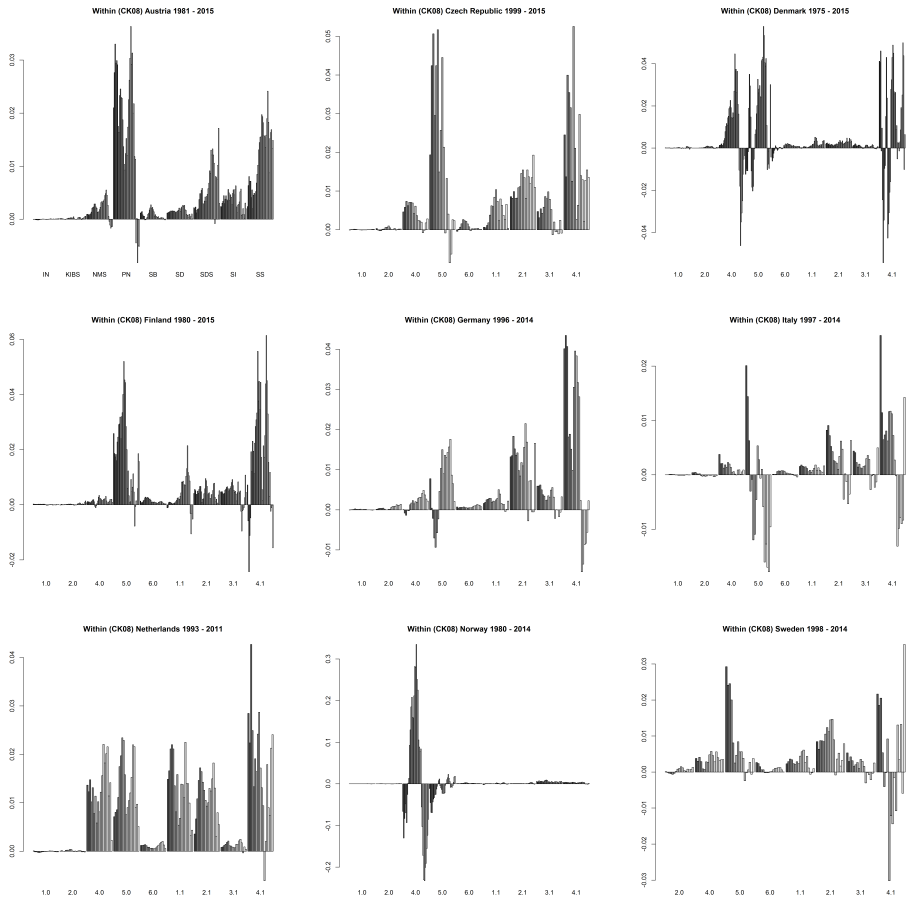


Figure 2.5: Within effect for 9 OECD countries, 5–years moving averages, R&D–intensity groups

Note: Values on the x–axis code the R&D classification of industries, ranging from 1 (highest intensity) to 6 (lowest intensity). Decimal values indicate if the group is belonging to manufacturing (.1) or non–manufacturing (.0). For example, 1.0 indicates the group pooling highest R&D intensity non–manufacturing industries.

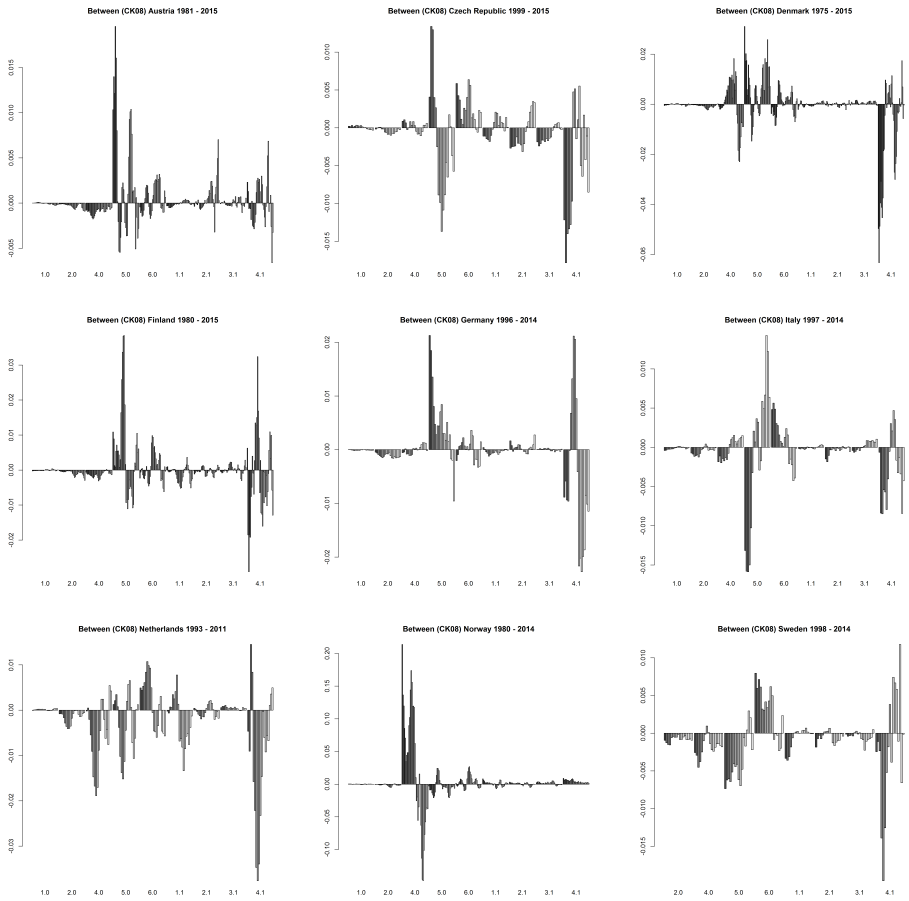


Figure 2.6: Between effect for 9 OECD countries, 5–years moving averages, R&D–intensity groups

Note: Values on the x–axis code the R&D classification of industries, ranging from 1 (highest intensity) to 6 (lowest intensity). Decimal values indicate if the group is belonging to manufacturing (.1) or non–manufacturing (.0). For example, 1.0 indicates the group pooling highest R&D intensity non–manufacturing industries.

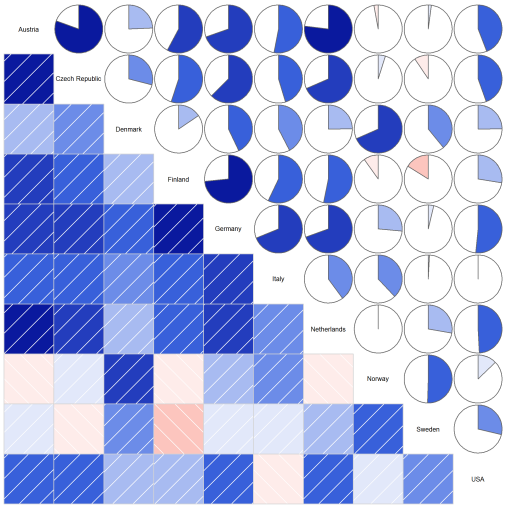


Figure 2.7: Between effect (Pavitt groups), country–level cross–correlations.

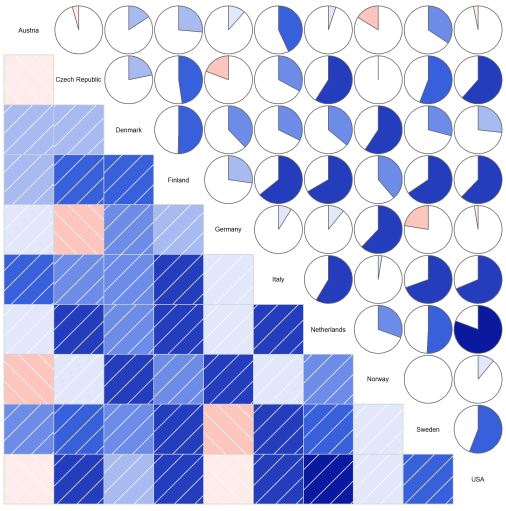


Figure 2.8: Between effect (Pavitt groups), country–level cross–correlations.

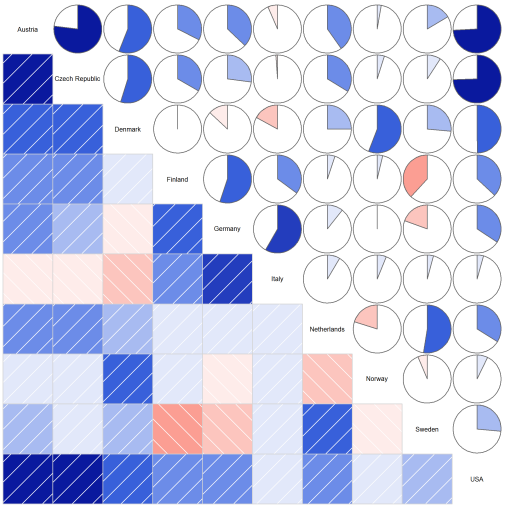


Figure 2.9: Within effect (R&D–intensity groups), country–level cross–correlations.

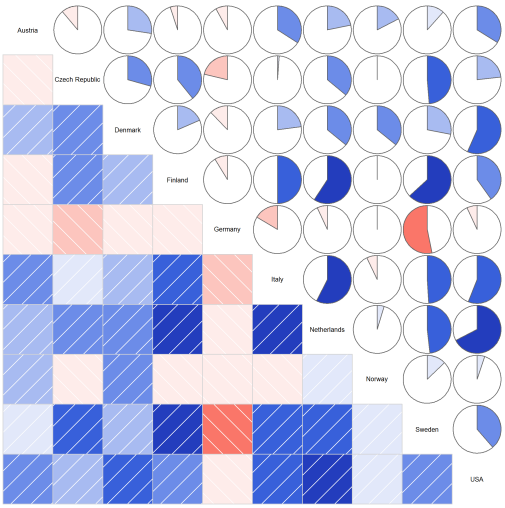


Figure 2.10: Between effect (R&D–intensity groups), country–level cross–correlations.

As a final step of our analysis of the compositional nature of the productivity slowdown, we check the trend of the within and between effects over time, in order to understand the changing contribution of innovation–related and structural–change–related drivers of productivity dynamics for different industry groups. Once again, we rely on the Cantner and Krüger normalization (percentage values of the components out of previous period levels) and show the results for the Pavitt–based industry grouping.

Figures 2.11 and 2.12 display respectively the trends of the within and between effects. To estimate the trend, we plotted the fitted value of a simple time–trend regression, selecting between a linear or a quadratic regression model according to the highest value of the goodness–of–fit. Boxes colored in blue indicate linear and non–linear increasing trends, while the opposite holds for red lines.

In general, even though the within effect is generally positive in terms of the contribution to productivity growth, its trend are mostly negative or inverted–U shaped. Positive trends are concentrated in manufacturing (SI, SS) but most of them follow an inverted–U non–linear trend that could turn negative in the near future. The trend for the SB industries is overall positive, but it applies to a small magnitude of the effect. From the trend analysis of the within effect, we obtain another piece of evidence for the idea that decreasing returns to innovation–based improvements are at currently work, a fact we link to the possibility of an ongoing exhaustion of technological opportunities (or time–lags before the emergence of a new general purpose technology–driven economic dynamism).

The evidence for the between effect is more scattered and display a more equal distribution of positive and negative trends. Roughly, the between effect has positive trends in manufacturing for some countries (like Austria, Denmark, US), and positive trends in services for others (like Italy, Netherlands, Norway). In the first case, industries groups characterized by higher–than–average productivity levels gain from structural change; in the second case, our informed guess about the Baumol disease–style dynamics is showing up, with services (usually displaying lower–than–average productivity levels) gaining labor shares in the economy. We reiterate here a considera-

tion about the sign of the between effect: when analyzing structural change dynamics, the between component should be expected to be negative, as labor–saving technological change in industries that are source of innovation pushes workers to other sectors, less productive in levels. Industries with high productivity lose labor shares (hence, have a negative between component) as structural change and reallocation is function of (relative) productivity levels, while their within component is positive but displays a negative trend, as it is function of the change in productivity, which is in turn affected by technological dynamics.

To summarize the results so far, the aggregate productivity slowdown is a compositional *collage* of innovation–driven and structural change–driven weights varying across industries and over time. We posit that all this results from structural technological mechanisms related to the shifting between techno–economic paradigms. If this holds true, the productivity slowdown might be coupled with a parallel dynamics in the realm of innovative activities that should be detectable. Hence, next Section takes a look at the innovation slowdown.

2.4 The Detection of Innovation Slowdown and Its Impact on Productivity Dynamics

2.4.1 Theory

In this Section, we focus on what we label the innovation slowdown. With innovation slowdown we mean the occurrence of a trend of declining intensity in innovative activities; in what follows, we trace such trend at different level of analysis, look at its compositional nature as we did for the productivity slowdown and later on we relate it to the productivity dynamics discussed in previous sections.

A reduction in the rate of innovative activities can be driven by many factors. First, the evidence of innovation slowdown can be the result of mismeasurement, if the choice of innovation measures used to assess the slowdown is

inappropriate or biased. In our empirical analysis, we build on a measure — the 'idea TFP' or, alternatively, 'research productivity' indicator — used by Bloom et al. (2017). In order to increase the robustness of such measure, we develop alternatives formulations of the same indicator. However, other pieces of evidence on the trends of innovative activities that are built around alternative indicators, such as the declining rate of innovator shares (for Germany, see Cantner (2016)) or pn measures of decreasing firm dynamism (Decker et al., 2014) produce similar results.

Second, starting with Arrow (1962) and following the classic literature on patent races (Reinganum, 1989), different market structure could be more or less conducive of innovative activities, where under the term 'market structure' we summarize here both the static nature of the market in terms of concentration and the dynamic conditions of competition for the market (e.g. appropriability, pre–emption). The debate on the so–called Schumpeter hypotheses (Cohen, 2010) continues until today (see Aghion et al. (2005)) and started to look at the trade determinants of R&D expenditures; for example, Dorn et al. (2016) corroborate the idea that R&D is complementary to manufacturing by showing a declining trend in research expenditures connected with changing patterns of international trade and specialization. Furthermore, very recently and at the very microeconomic level, the rise of markups has been considered a potential driver of macroeconomic consequences like the fall of labor share or the slowdown in output growth (De Loecker and Eeckhout, 2017) and, in the field of micro analysis of productivity, the surge of superstar firms, by increasing the gap in labor productivity (and multi–factor productivity) between the global frontier and the laggards may result in decreasing intensity in innovative activities (Andrews et al., 2016).

Third, innovation can slow down because economic actors redirect resources from exploitation to exploration, in an effort related to the discovery of new innovation 'directions' (Cantner, 2016) that require experimentation, learning, policy support and that, most important, are deeply rooted in uncertainty with respect to the possible outcomes.

Finally, a fourth possibility is that, following Gordon's hypothesis already discussed in Section 2.2, a decreasing importance of capital investments in

TFP growth results in diminishing rates of innovation. In the latter case, it is not the rate of innovative activity (or knowledge creation) itself that slows down; it is, instead, the ability of economic actors to effectively exploit current ideas and technological opportunities that slows down. If this is a permanent phenomenon or a transient one — where innovation slowdown is experienced in the time lag occurring while the needed adjustments (through learning and understanding) take place — is once again a question that divides techno–optimists and techno–pessimists.

In what follows, we assess the structural dynamics generated by these potential 'data generating processes' in order to describe the nature and composition of the innovation slowdown.

2.4.2 Methodology

We claim that the labor productivity growth slowdown discussed earlier on in the Chapter might have roots in learning lags or exhaustion of technological opportunities; in turn, these may result from an increasing difficulty in finding new ideas. In other words, to create new knowledge and keep the current pace of technological development, it is necessary to exploit an increasing amount of resources — decreasing returns 'bite' more and more in the production of new know–how. In order to have an idea of the magnitude of this phenomenon, we adopted a measure suggested by Bloom et al. (2017) labeled idea TFP. The evidence in Bloom et al. (2017) suggests that idea TFP captures the slowing down of research productivity occurred due to a faster growth of employed researchers ('idea input') compared with TFP growth rate ('idea output'). Despite some shortcomings of this statistic and, in general, of the explanatory power of TFP (the measure of our ignorance, to cite Abramovitz (1989)) to capture technological dynamics, we consider idea TFP a useful proxy to capture the innovation slowdown.

To provide a more robust analysis, we calculate other close measures that retain the idea output–to–input ratio concept, along with a precise replication of the idea TFP indicator. A summary of the measures we use can be found in Table 2.7. Indeed, TFP, being a measure of the efficiency of

transformation of input into output, reflects either the efficiency of usage or the quality of the embodied knowledge expressed in applied technology or/and processes. Thus, there are other possible proxies for both idea input and output such as respectively, R&D expenditures and labor productivity (hereinafter LP) growth.

The slowdown in knowledge creation can have as well a compositional nature as we found in Section 2.2 for labor productivity trends. Macro trends at the country level may result either from similar patterns among industries (yeast–like process) or heterogeneous dynamics of industries (mushroom–like process). The more or less evenly–distributed nature of the industries' contribution to aggregate dynamics of a variable of interest (being it productivity of idea productivity) depends on different factors; for example, the accumulated knowledge — the history dependent path of accumulation or the potential for a new GPT emergence (Cantner and Vannuccini, 2012) in a particular industry, the sources, direction and structure of knowledge flows among sectors (Pavitt, 1984), the cross–industry demand elasticities, and so on. In general, such interdependent and connected nature of industrial structures which results in aggregated macro pattern is certainly important to account for, as for instance the pervasiveness of GPTs induces multidimensional and inter–temporal structural changes leading to non–linearities in innovation and economic inducements and reactions. Furthermore, once a technology showing the properties of a GPT appears, its influence spreads through the economy's structure which differs across countries and, thus, producing different speed, directions and overall patterns of structural changes.

Given the discussion above, we conduct the analysis at both the macro (country) and meso (industry) levels of analysis. If the causal relation between knowledge and technological slowdown exists, given the obtained results in Section 2.3.4, it is reasonable to assume similar patterns of idea TFP and LP.

Table 2.7: Indicators of innovation slowdown

Macro level	
Measure	**Description**
Indicator 1	Idea TFP R&D = TFP growth rate (annual)/R&D expenditures
Indicator 2	Idea TFP Researcher = TFP growth rate (annual)/Number of researchers
Indicator 3	Idea LP R&D = Labor productivity growth rate (annual)/R&D expenditures
Indicator 4	Idea LP Researchers = Labor productivity growth rate (annual)/Number of researchers
Meso level	
Measure	**Description**
Indicator 1	Idea LP R&D = Labor productivity growth rate/R&D expenditures. 5 years moving average

2.4.3 Data

As mentioned above, we perform the analysis at two levels of (dis)aggregation: macro and meso. For the macro–level measures based on LP growth rate (Indicators 3 and 4, see Table 2.7) we collected annual data on the same variables used for the productivity slowdown analysis, namely production (nominal output) and employment, retrieved from the OECD Structural Analysis (STAN) database (ISIC v.4 SNA08 and ISIC v.4 SNA93). To construct the indicators, we collected additional data on R&D expenditures (OECD STAN ISIC v.4), number of researchers (OECD Research and Development Statistics database), and multifactor productivity (OECD Productivity Database). Using several ways to measure the innovation slowdown helps to increase the validity and robustness of results against potential mismeasurement problems. Table 2.8 summarizes the sets of countries and time spans used to compute the different indicators.

For what concerns the meso–level indicator, that we employ to assess the compositional nature of the innovation slowdown, we use our previous calculations of LP growth rates (5–years moving averages) to exclude industry–specific shocks potentially distorting the innovation slowdown dynamics. The number of countries for which LP growth rates were available decreases from 10 to 7 because of data gaps and due to different industries classifications of the data on production and employment in OECD STAN ISIC

Table 2.8: Set of countries under analysis

Countries	Indicator 1	Indicator 2	Indicator 3	Indicator 4
Australia	1	1	0	0
Austria	1	0	1	0
Belgium	1	1	1	1
Canada	1	1	0	0
Czech Republic	0	0	1	0
Denmark	0	1	0	1
Finland	1	0	1	0
France	0	1	0	1
Germany	1	1	1	1
Italy	1	1	1	1
Japan	1	1	1	1
Korea	1	1	0	0
Mexico	0	0	1	0
Netherlands	0	1	0	1
Norway	0	0	1	0
Portugal	1	1	1	1
Slovakia	0	0	1	0
Slovenia	0	0	1	0
Spain	1	1	0	0
United Kingdom	0	1	0	0
United States	1	1	1	1
Number of countries	12	14	13	9
From	1997	1987	1997	1987
To	2013	2013	2014	2014

v.4 SNA08 and ISIC v.4 SNA93 (used for idea output) and R&D expenditures in OECD STAN ISIC v.4 (used for idea input) which makes data matching rather cumbersome at this fine–grained level of analysis. A precise description of the data is reported in Table 2.14 and 2.15.

We grouped industries according to the Pavitt taxonomy in order to ease visualization and, most important, to capture a gist of industries' supplier–user structure. By applying this classification we want to understand whether there is a localization of the innovation slowdown and how the structure of knowledge flows shapes the aggregated pattern. As there is a prevalence of manufacturing sector over services in the sample (see Table 2.10), our conclusions from the analysis can be reliably attributed to the manufacturing sector only, while a thorough analysis of services will require additional data.

Table 2.9: Industries dataset description

Country	Number of industries	From	To
Austria	23	2002	2013
Czech Republic	23	2000	2014
Finland	17	1999	2014
Germany	22	1999	2014
Italy	23	1997	2014
Norway	17	1991	2014
USA	24	2002	2011

Table 2.10: Pavitt taxonomy groups

Pavitt	Austria	Czech Republic	Finland	Germany	Italy	Norway	USA
SI	10	8	7	10	10	6	10
SD	7	7	4	5	5	5	4
SB	1	1	0	1	1	0	0
SS	3	3	3	3	3	3	3
PN	1	1	1	1	1	1	2
KIBS	1	1	1	1	1	1	0
IN	0	1	1	1	1	1	3
SDS	0	1	0	0	1	0	2
Total	23	23	17	22	23	17	24

2.4.4 Analysis and Discussion of the Results

First of all, it is important to highlight our evidence at the macro level. We detect the innovation slowdown in a overwhelming majority of countries, irrespective of the indicator used. The evidence is summarized in Table 2.11. This consistent pattern across measures signals that the innovation slowdown exists beyond potential mismeasurement issues.

If we give a closer look at Indicator 2 in Table 2.11, which is a precise replication of (Bloom et al., 2017) idea TFP, one may see that all countries, except for the USA, are experiencing the slowdown in the idea production with respect to employed research effort. However, the USA shows an upward trend mostly because the available data covers the time span from 1987 to 2007 and the lack of more recent data can explain the 'exception'. Furthermore, while the generalized pattern shows an increase (flat positive slope), real annual numbers for Indicator 2 after 2003 sharply went down. For a visual inspection see the bottom rightmost chart in the Figure 2.13

Table 2.11: Summary of results on innovation productivity. Macro level

Country	Indicator 1	Indicator 2	Indicator 3	Indicator 4
Australia	↘	↘	-	-
Austria	↘	-	↘	-
Belgium	↘	↘	↘	↘
Canada	↘	↘	-	-
Czech Republic	-	-	↘	-
Denmark	-	↘	-	↘
Finland	↘	-	↘	-
France	-	↘	-	↘
Germany	↘	↘	↘	↘
Italy	↘	↘	↘	↘
Japan	↗	↘	↘	↘
Korea	↘	↘	-	-
Mexico	-	-	↘	-
Netherlands	-	↘	-	↘
Norway	-	-	↘	-
Portugal	↘	↘	↘	↘
Slovakia	-	-	↘	-
Slovenia	-	-	↘	-
Spain	↗	↘	-	-
United Kingdom	-	↘	-	-
United States	↘	↗	↘	↘
Total	12	14	13	9

Note: The direction of the arrows indicates the sign of the trend–line fitting the indicators series, with ↗ ↘ pointing respectively to an increasing or decreasing trend of the indicator. For cells with (-) the respective indicator could not be calculated with the available data.

plotting the described calculation results on Indicator 2.

Moving to the more disaggregated level of industry, we obtained less consistent results for services, as it was presumed due to the insufficient representativeness of industries from this sector in the sample. For the manufacturing sector, the tendency of slowing–down research productivity is clearer; however, it is rather non–linear in comparison with the country level results as displayed in Figure 2.13. Table 2.12 summarizes the evidence at the industry level.

In sum, the innovation slowdown exists, its detection is robust to an array of different proxies used to capture it, and it emerges both a the macro and at the meso level of analysis. As for the productivity slowdown, also the innovation slowdown is generalized and displays a compositional nature.

Table 2.12: Summary of results on innovation productivity. Meso level

Pavitt	Austria	Czech Republic	Finland	Germany	Italy	Norway	USA
Manufacturing							
SI	↘	↘	→	→ / ↘	↘	→ / ↘	U-inverted
SS	↘	↘	↘	↘	↘	↘	U-inverted
SD	↘	↘	↘	→	↘	↘	↘
SB	↘	↗	-	↘	→	-	-
Services							
PN	-	↘	↘	↗↘	↘	↘	→
IN	-	↘	↗	↘	↘	↘	↘
KIBS	↘	U-shaped	→	↘	→	→	-
SDS	-	→	-	-	↘	-	↘

Note: The direction of the arrows indicates the sign of the trend–line fitting the indicators series, with ↗ → ↘ pointing respectively to an increasing, constant or decreasing trend of the indicator. For cells with (-) the respective indicator could not be calculated with the available data.

A final issue to tackle has to do with the structural relationship between productivity and innovation slowdown. In this sense, our clustering of industries into the Pavitt taxonomy becomes key to identify the possible direction of the relationships and to hypothesize in which cases and to what extent patterns of innovation slowdown may influence patterns of productivity slowdown. For example, a selection of Germany and USA supplier dominated (SD), specialized suppliers (SS) and scale intensive (SI) Pavitt groups in Figures (respectively) 2.14 and 2.15 illustrate co–directed movement of manufacturing industries dynamics among these classes.

Theoretically, if the slowdown in knowledge creation initiates in the specialized suppliers group (innovation producers/sellers), then it may be transferred in form of productivity slowdown to application industries (e.g. SD) that are buyers of SS research outcomes. Such effect of innovation slowdown on productivity dynamics might be less strong if it starts in industry groups that receivers, rather than creators, of new knowledge, as the vertical upstream–to–downstream transmission channel of innovation flows is missing. In a nutshell, being a final link in the knowledge flow chain, if one observes a decreasing research productivity in the group of knowledge–receivers industries, e.g. the SI group, this might have negligible impact on labor productivity of this or other groups as technological improvements (and hence knowledge) are mainly exogenously received and used for 'consumption' within this SI group.

To elaborate further on these insights, however, we need to explore the relationship between the productivity and the innovation slowdown. We provide a first correlation analysis in the following Section.

2.4.5 The Relationship between the Innovation and Productivity Slowdown

In order to assess the relationship between the two structural dynamics we observed in this study, we opt for an exploratory measurement of the intra– and inter–industry correlations among productivity and innovation slowdown. In general, this task can be addressed by adopting a parametric approach in line with classic studies on inter–industry spillovers (Bernstein and Nadiri, 1989) or by capturing interdependencies through using input–ouput methods such as those used in studies on technology flow matrices (Verspagen and De Loo, 1999).

In this Chapter, we limit ourselves to uncovering co–movements in trends between productivity and innovation slowdown across groups of industries. We attempt to capture the structure of connections among industries in order to localize potential sources/origins of innovation slowdown and to identify its transmission channels to productivity slowdown. The non–parametric method applied in this section is a table constructed as a product of, on the one hand, correlation matrices between innovation and productivity measures and, on the other hand, OECD input–output tables describing sales and purchases relationships among industries. The rationale behind the construction of this more elaborated representation is that a co–movement of innovation and productivity measures time series can occur just by chance; by accounting for the input–output relations, we exclude spurious correlations, as input–output coefficients register the existence and the strength of actual trade relations between any two industries. In a nutshell, to test the hypothesis of innovation–productivity nexus, we assume that there is a transmission unit which embodies the result of innovation activity of the selling sector and that, through implementation, may influence productivity of buyer sectors. Thus, here we consider upstream–downstream relations among industries as a premise/precondition for the innovation–productivity

Table 2.13: Summary of significant innovation–productivity correlations, Germany

DEU			Manufacturing				Services			
		Prod	SS	SI	SD	SB	SDS	PN	KIBS	IN
	Inno	N	3	10	5	1	0	1	1	1
Manuf.	SS	3	100%	70%	60%	100%		33%	100%	0%
	SI	10	67%	56%	28%	70%		30%	60%	0%
	SD	5	53%	34%	48%	60%		20%	40%	20%
	SB	1	100%	50%	20%	100%		0%	100%	0%
Services	SDS	0								
	PN	1	67%	40%	40%	100%		100%	0%	100%
	KIBS	1	100%	40%	20%	100%		0%	100%	0%
	IN	1	0%	30%	0%	0%		0%	0%	100%

Note: Color scheme: 50–100% — green, 30–50% — yellow, < 0 —- red, not in the sample — gray

nexus. In light of this interpretation, the Pavitt taxonomy offers a meaningful structure of industrial relations that might be a starting point for the analysis of innovation–productivity nexus.

The procedure we implement runs as follows: on a first stage for each of 7 countries available we constructed correlation tables, where on the vertical axis there is the innovation measure for n industries and on the horizontal axis productivity. Thus, each cell contains a correlation coefficient between innovation measure of row–industry i and productivity measure of column–industry j. Eventually, we have a non–symmetric ($a_{ij} \neq a_{ji}$) matrix of size $n \times n$ where the empty cells indicate statistically insignificant coefficients. The main diagonal shows intra–industry innovation–productivity nexus while off–diagonal values represent inter–industry relations. Inter–industry correlations have to be considered because, as we pointed out in the previous section, given the assumption that some industries are producers of new knowledge–embodied capital while others are buyers of this capital, the consideration of only intra–industry correlations (diagonal values) would ignore the phenomenon we aim at uncovering, namely the structural nature of the innovation–productivity nexus.

In Table 2.13 we report a summary of calculated correlation coefficients for Germany; industries are clustered according to Pavitt taxonomy.

The top left and bottom right quadrants represent intra–sectoral connec-

tions between innovation and productivity measures within manufacturing and services respectively. The top right quadrant shows inter–sectoral connections between innovation measure for manufacturing and productivity measure in services while the bottom left quadrant shows the opposite connection. N is a number of industries assigned to the respective Pavitt group. The values in cells indicate the share of statistically significant coefficients calculated for industries which belong to a pair of Pavitt groups. For example, cell a_{12} indicates that between innovation measure for 3 SS industries and productivity measure for 10 SI industries we obtained 70 percent of significant correlations, which means 21 coefficients out of 30. Colors highlight the possible structure of the connections where shares of significant coefficients are high and either positive (green and yellow) or negative (red). For all 7 countries within the manufacturing sector we found that the correlation between innovation and productivity measures is high, significant for a big share of cross–correlations and in all cases positive. The results within services sector and manufacturing–services interconnections are rather mixed across countries.

However, correlations grasp only co–movement, tell nothing about causality, and can also be spurious. The use of input–output data can resolve these problems by indicating the existence of channels for transmission of innovation slowdown to productivity slowdown and, therefore, allows advancing insights regarding causality. Besides, the magnitude of input–output index shows the potential transmission capacity of the channel. Indeed, even if a correlation coefficient is high and positive — indicating co–movement of innovation and productivity measures for a pair of industries —, if the these industries have a weak seller–buyer relationship it means that this channel barely can transmit slowdown from innovation to productivity.

Given that, in a second step we took OECD Leontief Inverse input–output matrices as a measure of inter–industry connections. The Leontief Inverse contains indices showing rise in output of an industry i due to the unit increase in demand of industry j. Therefore, these matrices reflect the structure and importance (the magnitude) of each connection.

Eventually, to display uncovered structures we construct heatmaps (follow-

ing Acemoglu et al. (2016)) as an insightful visualization tool for the product of correlation and input–output matrices.

The heatmap in Figure 2.16, displaying the structural properties for Germany, shows strong intra–industry connection with high values on the main diagonal. This outcome is a result of high input–output coefficient because of high 'self–demand' share for each industry. The heatmaps for the remaining 6 countries are displayed in Figure 2.17 in the Appendix. The main pattern that the majority of countries share regards evidence of structure within the manufacturing sector; this holds true for all countries apart from Finland and Norway. This suggests that innovation and productivity measures are co–moved and the trade structure among industries allows for the transmission of the slowdown effect. The other regularity, taking place in Germany, Italy, USA and Finland, is that PN industries, for example logistics and warehousing, play a role in affecting the productivity of the manufacturing sector. The same holds true for KIBS industries in Austria, Germany, Italy and Norway.

In sum, after we sorted out potentially spurious correlations by multiplying correlation tables with input–output indices, the structure of connections between innovation and productivity measures remains, especially for the manufacturing sector. This does not allow ruling out the hypothesis about innovation–productivity nexus and keeps it under further investigation.

2.5 Conclusion

In this Chapter, we contribute to the ongoing discussion on labor productivity growth slowdown in a novel way. We assessed the structural properties of productivity dynamics by (i) looking at a more disaggregated pictures — the industry level —, (ii) by applying a non–parametric decomposition technique usually performed to distinguish between learning and reallocation effects at the firm-level, and iii) by interpreting the skewed and compositional nature of the slowdown in an evolutionary manner. We grouped our industry–level analysis using two different classifications, the Pavitt taxonomy and a taxonomy capturing industries' R&D intensities, in order to assess the robustness

of our findings across different logics of aggregation of economics activities. We performed the analysis on a dataset of ten OECD countries, using the most recent available data.

The findings suggest that (i) the compositional nature of the productivity slowdown is a common trend in the countries under analysis, that (ii) within effects prevail over between effects, suggesting that the driver of productivity dynamics have to be searched in the heterogeneous and skewed contributions of different industry groups to aggregate productivity growth rather than to selection (structural change) effects due to reallocation. The latter, however, when present, tend to go again the principles of selection and the replicator dynamics, with labor shares moving to less productive industry groups. All the evidence points to the possibility that the slowdown is driven by the exploitation of established technological opportunities in knowledge–intensive industries coupled with structural shifts of economic activities towards services, in a 'Baumol disease–like' fashion.

With respect to the innovation slowdown, we find a generalized decreasing trend of research productivity across countries and Pavitt industry groups. While the indicator used and suggested by Bloom et al. (2017) to capture the innovation slowdown can be subject to criticism, we derived four alternative specifications of the measure, and all of these follow a comparable dynamics.

Our analysis, though exploratory, is the first to offer a fresh view on the productivity slowdown and to link the tools and concepts of market selection with meso–economic analysis. As mentioned in Section 2.1, we are interested in the structural dynamics lying behind slowdown in productivity and innovative activities rather than in their specific determinants. In this sense, we do not delve into causal explanation but rather offer an 'informed guess' about the nature of the phenomena. Some issues might have therefore received not enough attention. For example, we do not discuss how import–export dynamics (that it the international distribution of production) can affect the PS. Nonetheless, issues such as the effects of international trade on productivity growth falls into our explanation of the trends in terms of structural change — as international competition pushes labor to flow across industries with consequent compositional effects on aggregate productivity

growth.

In sum, the Chapter supports an explanation of productivity slowdown that adds to the long–run threat of growth headwinds highlighted by Gordon, namely the closure of the technological opportunities set available in the current techno–economic paradigm. Such explanation relies on the detected trends of innovation slowdown, but it should not be interpreted as the last word on the long–run trends of productivity developments. Decreasing returns of innovation and discovery due to exhausted technological opportunity are only one side of structural transformations; however, for these transformations to be unleashed new directions of innovative change might be worth exploring. Thus, in conclusion, if the productivity slowdown is a symptom of a deeper technological slowdown, concerns from scholars and policy–makers should be directed to how to open–up new opportunities and, therefore, give a more promising future to economic growth.

Appendix

Table 2.14: Variables description. Country level

Variable	Description
TFP growth rate	Multifactor productivity, annual growth rate
ANBERD	Research and Development (R&D) expenditures are expressed in national current prices, millions
Number of researchers	Measured in full-time equivalent are researchers in the business enterprise sector by industry according to the International Standard Industrial Classification (ISIC) revision 3.1
LP growth rate	$(LP_t - LP_{t-1})/LP_{t-1}$, where $LP = PRDK/EMPN$
PRDK	Production (PRDK) is a volume measure expressed in current price of the reference year 2010, millions
EMPN	Total employment is displayed as thousands of persons (headcounts) engaged

Table 2.15: Variables description. Industry level

Variable	Description
ANBERD	Research and Development (R&D) expenditures are expressed in national current prices. Millions, 5 years moving averages
LP growth rate	$(LP_t - LP_{t-1})/LP_{t-1}$, where $LP = PRDK/EMPN$, 5 years moving averages
PRDK	Production (PRDK) is a volume measure expressed in current price of the reference year 2010, millions
EMPN	Total employment is displayed as thousands of persons (headcounts) engaged

Table 2.16: Assignment of industries to taxonomy groups

Industries	R&D group	RxMy	Pavitt
D01T99: Total	-	-	-
D01T03: Agriculture, forestry and fishing [A]	5	5.0	SD
D01T02: Agriculture, hunting and forestry	5	5.0	SD
D01: Crop and animal production, hunting and related service activities	5	5.0	SD
D02: Forestry and logging	5	5.0	SD
D03: Fishing and aquaculture	5	5.0	SD
D05T09: Mining and quarrying [B]	4	4.0	SI
D05T06: Mining and quarrying of energy producing materials	4	4.0	SI
D07T08: Mining and quarrying except energy producing materials	4	4.0	SI
D09: Mining support service activities	4	4.0	PN
D10T33: Manufacturing [C]	3	3.1	na
D10T12: Food products, beverages and tobacco	4	4.1	SI
D10T11: Food products and beverages	4	4.1	SI
D10: Food products	4	4.1	SI
D11: Beverages	4	4.1	SI
D12: Tobacco products	4	4.1	SI
D13T15: Textiles, wearing apparel, leather and related prodcuts	4	4.1	SD
D13T14: Textiles and wearing apparel	4	4.1	SD
D13: Textiles	4	4.1	SD
D14: Wearing apparel	4	4.1	SD
D15: Leather and related products	4	4.1	SD
D16T18: Wood and paper products, and printing	4	4.1	SD
D16: Wood and products of wood and cork, except furniture	4	4.1	SD
D17: Paper and paper products	4	4.1	SD
D18: Printing and reproduction of recorded media	4	4.1	SD
D19T23: Chemical, rubber, plastics, fuel products and other non-metallic mineral products	3	3.1	SI
D19: Coke and refined petroleum products	4	4.1	SI
D20T21: Chemical and pharmaceutical products	2	2.1	SI
D20: Chemicals and chemical products	2	2.1	SI
D21: Basic pharmaceutical products and pharmaceutical preparations	1	1.1	SB
D22T23: Rubber and plastics products, and other non-metallic mineral products	3	3.1	SI
D22: Rubber and plastics products	3	3.1	SI
D23: Other non-metallic mineral products	3	3.1	SI
D24T25: Basic metals and fabricated metal products, except machinery and equipment	4	4.1	SI
D24: Basic metals	3	3.1	SI
D241T31: Iron and steel	3	3.1	SI
D242T32: Non-ferrous metals	3	3.1	SI
D25: Fabricated metal products, except machinery and equipment	4	4.1	SI

Table 2.17: Assignment of industries to taxonomy groups

Industries	R&D group	RxMy	Pavitt
D26T28: Machinery and equipment	2	2.1	SS
D26T27: Electrical, electronic and optical equipment	2	2.1	SS
D26: Computer, electronic and optical products	1	1.1	SS
D27: Electrical equipment	2	2.1	SS
D28: Machinery and equipment n.e.c.	2	2.1	SS
D29T30: Transport equipment	2	2.1	SI
D29: Motor vehicles, trailers and semi-trailers	2	2.1	SI
D30: Other transport equipment	1	1.1	SI
D301: Building of ships and boats	3	3.1	SI
D303: Air and spacecraft and related machinery	1	1.1	SI
D304: Military fighting vehicles	2	2.1	SI
D302A9: Railroad equipment and transport equipment n.e.c.	2	2.1	SI
D31T33: Furniture; other manufacturing; repair and installation of machinery and equipment	4	4.1	SD
D31T32: Furniture, other manufacturing	4	4.1	SD
D33: Repair and installation of machinery and equipment	3	3.1	SD
D35T39: Electricity, gas and water supply; sewerage, waste management and remediation activities [D-E]	5	5.0	SDS
D35: Electricity, gas, steam and air conditioning supply [D]	5	5.0	SDS
D36T39: Water supply; sewerage, waste management and remediation activities [E]	5	5.0	SDS
D36: Water collection, treatment and supply	5	5.0	SDS
D37T39: Sewerage, waste collection, treatment and disposal activities; materials recovery; remediation activities and other waste management services	5	5.0	SDS
D41T43: Construction [F]	5	5.0	SD
D45T56: Wholesale and retail trade; repair of motor vehicles and motorcycles; transportation and storage; accommodation and food service activities [G-I]	5	5.0	PN
D45T47: Wholesale and retail trade, repair of motor vehicles and motorcycles [G]	5	5.0	PN
D45: Wholesale and retail trade and repair of motor vehicles and motorcycles	5	5.0	PN
D46: Wholesale trade, except of motor vehicles and motorcycles	5	5.0	PN
D47: Retail trade, except of motor vehicles and motorcycles	5	5.0	PN
D49T53: Transportation and storage [H]	5	5.0	PN
D49: Land transport and transport via pipelines	5	5.0	PN
D50: Water transport	5	5.0	PN
D51: Air transport	5	5.0	PN

Table 2.18: Assignment of industries to taxonomy groups

Industries	R&D group	RxMy	Pavitt
D53: Postal and courier activities	5	5.0	IN
D55T56: Accommodation and food service activities [I]	5	5.0	SDS
D58T63: Information and communication [J]	3	3.0	IN
D58T60: Publishing, audiovisual and broadcasting activities	5	5.0	IN
D58: Publishing activities	4	4.0	IN
D59T60: Audiovisual and broadcasting activities	5	5.0	IN
D61: Telecommunications	4	4.0	IN
D62T63: IT and other information services	2	2.0	IN
D62: Computer programming, consultancy and related activities	2	2.0	IN
D63: Information service activities	2	2.0	IN
D64T66: Financial and insurance activities [K]	5	5.0	IN
D64: Financial service activities, except insurance and pension funding	5	5.0	IN
D65: Insurance, reinsurance and pension funding, except compulsory social security	5	5.0	IN
D66: Activities auxiliary to financial service and insurance activities	5	5.0	IN
D68T82: Real estate, renting and business activities [L-N]	5	5.0	SDS/KIBS
D68: Real estate activities [L]	5	5.0	IN
D69T82: Professional, scientific and technical activities; administrative and support service activities [M-N]	4	4.0	SDS/KIBS
D69T75: Professional, scientific and technical activities [M]	4	4.0	KIBS
D69T71: Legal and accounting activities; activities of head offices; management consultancy activities; architecture and engineering activities; technical testing and analysis	4	4.0	IN/KIBS
D69T70: Legal and accounting activities; activities of head offices; management consultancy activities	4	4.0	IN/KIBS
D69: Legal and accounting activities	4	4.0	IN
D70: Activities of head offices; management consultancy activities	4	4.0	KIBS
D71: Architectural and engineering activities; technical testing and analysis	4	4.0	KIBS
D72: Scientific research and development	1	1.0	KIBS
D73T75: Advertising and market research; other professional, scientific and technical activities; veterinary activities	4	4.0	KIBS
D73: Advertising and market research	4	4.0	KIBS
D74T75: Other professional, scientific and technical activities; veterinary activities	4	4.0	KIBS

Table 2.19: Assignment of industries to taxonomy groups

Industries	R&D group	RxMy	Pavitt
D74: Other professional, scientific and technical activities	4	4.0	KIBS
D75: Veterinary activities	4	4.0	SDS
D77T82: Administrative and support service activities [N]	5	5.0	SDS
D77: Rental and leasing activities	5	5.0	SDS
D78: Employment activities	5	5.0	SDS
D79: Travel agency, tour operator, reservation service and related activities	5	5.0	PN
D80T82: Security and investigation activities; services to buildings and landscape activities; office administrative, office support and other business support activities	5	5.0	SDS
D84T99: Community, social and personal services [O-U]	6	6.0	NMS
D84T88: Public administration and defence; compulsory social security; education; human health and social work activities [O-Q]	6	6.0	NMS
D84: Public administration and defence; compulsory social security [O]	6	6.0	NMS
D85: Education [P]	6	6.0	NMS
D86T88: Human health and social work activities [Q]	6	6.0	NMS
D86: Human health activities	6	6.0	NMS
D87T88: Residential care and social work activities	6	6.0	NMS
D90T99: Arts, entertainment, repair of household goods and other services [R-U]	5	5.0	SDS
D90T93: Arts, entertainment and recreation [R]	5	5.0	SDS
D90T92: Creative, arts and entertainment activities; libraries, archives, museums and other cultural activities; gambling and betting activities	5	5.0	SDS
D93: Sports activities and amusement and recreation activities	5	5.0	SDS
D94T96: Other service activities [S]	5	5.0	SDS
D94: Activities of membership organizations	5	5.0	SDS
D95: Repair of computers and personal and household goods	5	5.0	PN
D96: Other personal service activities	5	5.0	SDS
D97T98: Activities of households as employers; undifferentiated activities of households for own use [T]	5	5.0	SDS

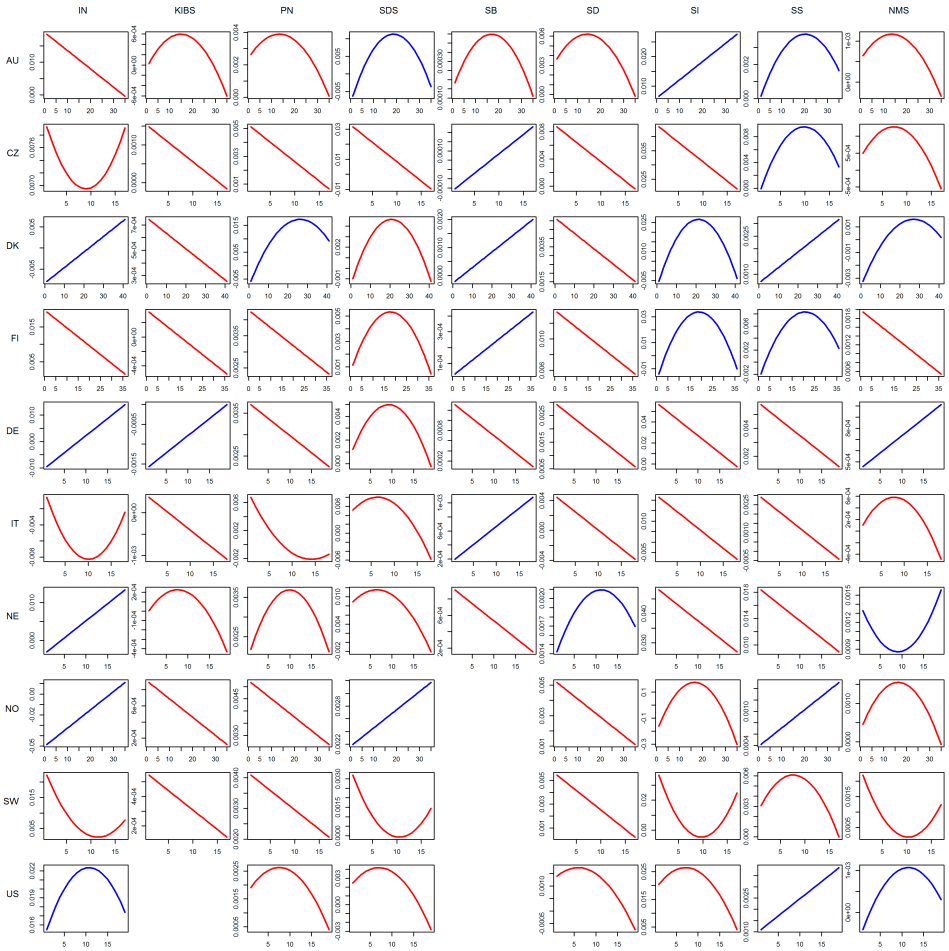


Figure 2.11: Within effect — trends

Note: The trend of the component is obtained from the fitted values of a linear or quadratic regression using the time–series of within effect for each country–group pair. The choice of the regression model is based on the higher goodness–of–fit (F–stat); line colors indicate an increasing (blue) or decreasing (red) trend.

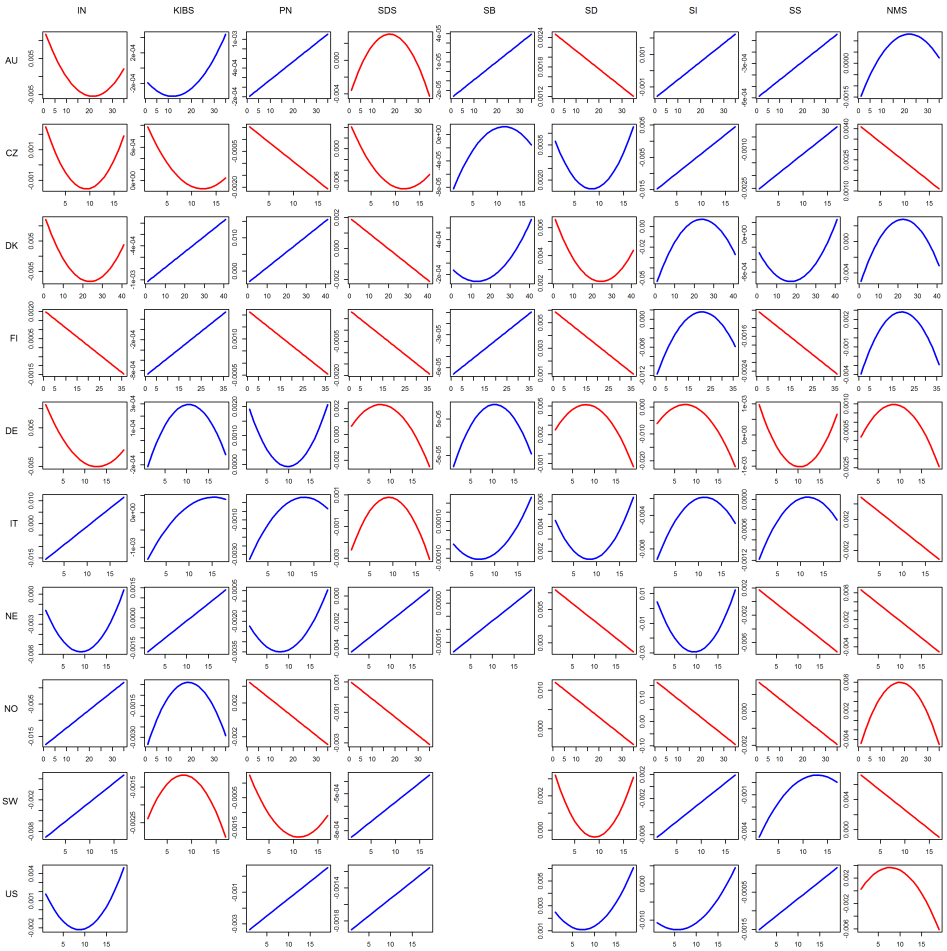


Figure 2.12: Between effect — trends

Note: The trend of the component is obtained from the fitted values of a linear or quadratic regression using the time–series of between effect for each country–group pair. The choice of the regression model is based on the higher goodness–of–fit (F–stat); line colors indicate an increasing (blue) or decreasing (red) trend.

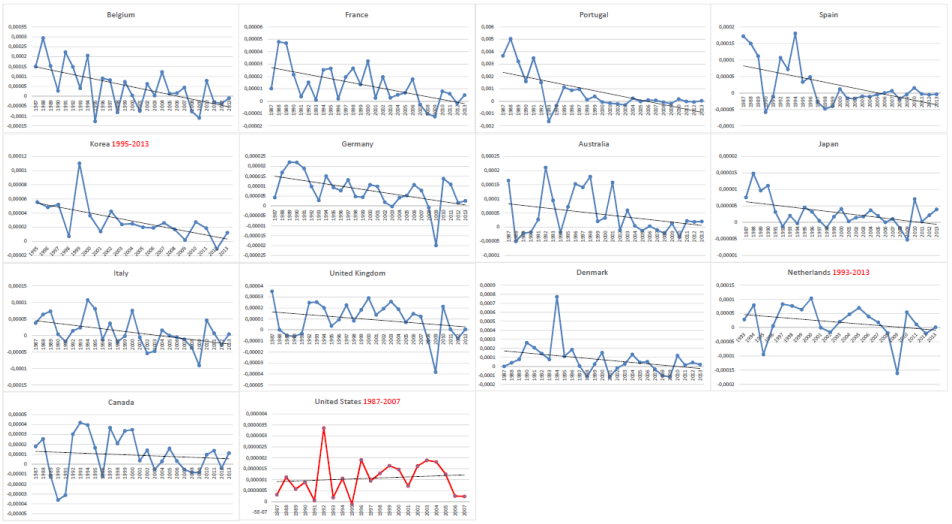


Figure 2.13: Indicator 2 dynamics – whole sample, macro

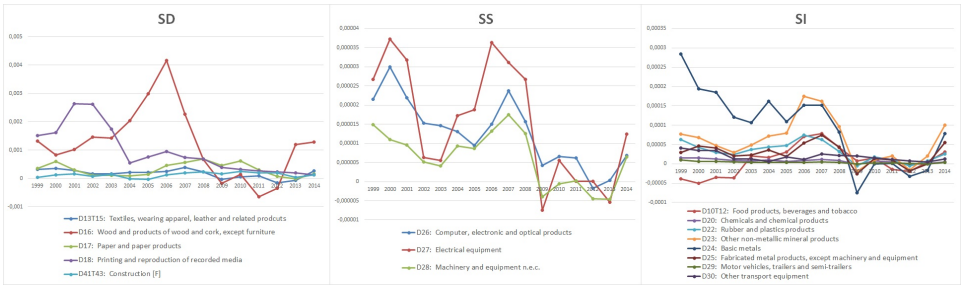


Figure 2.14: Indicator 1 dynamics for Pavitt groups. Germany

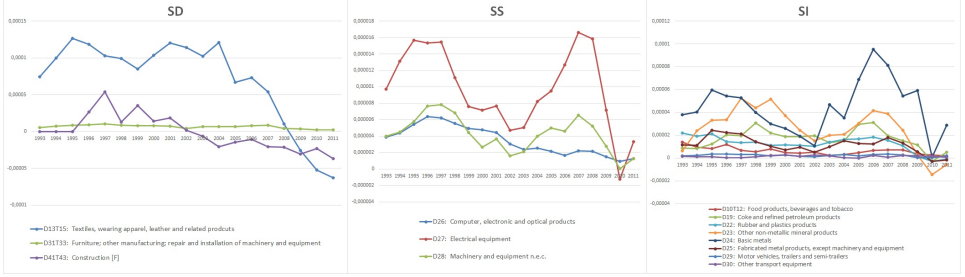


Figure 2.15: Indicator 1 dynamics for Pavitt groups. USA

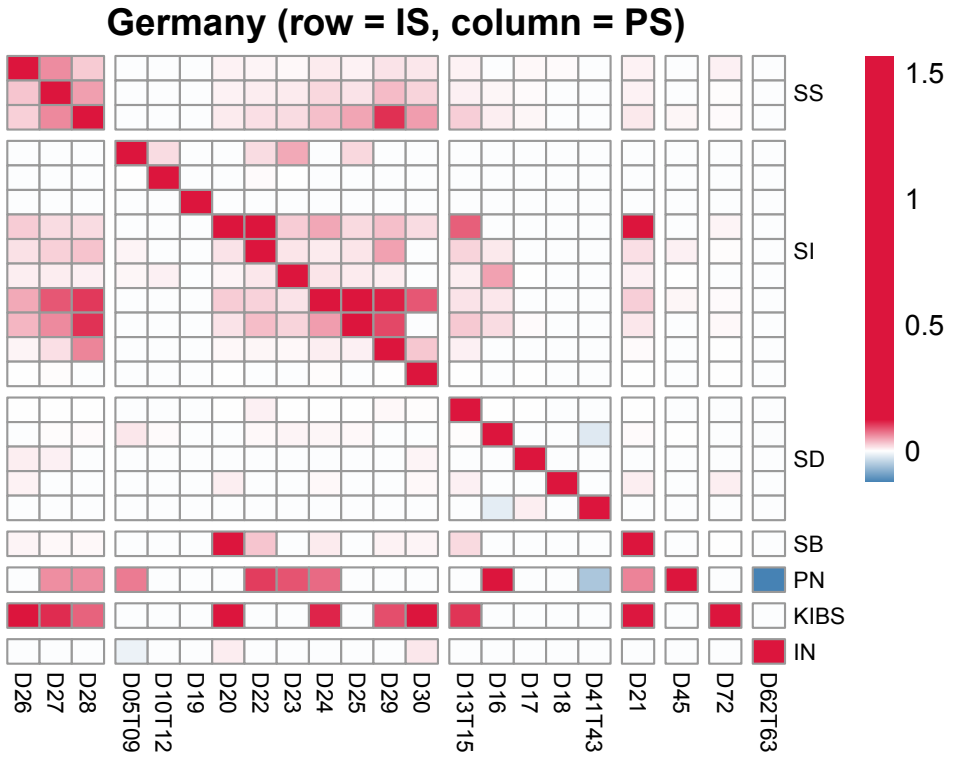


Figure 2.16: Correlation table weighed by input–output coefficients. Germany

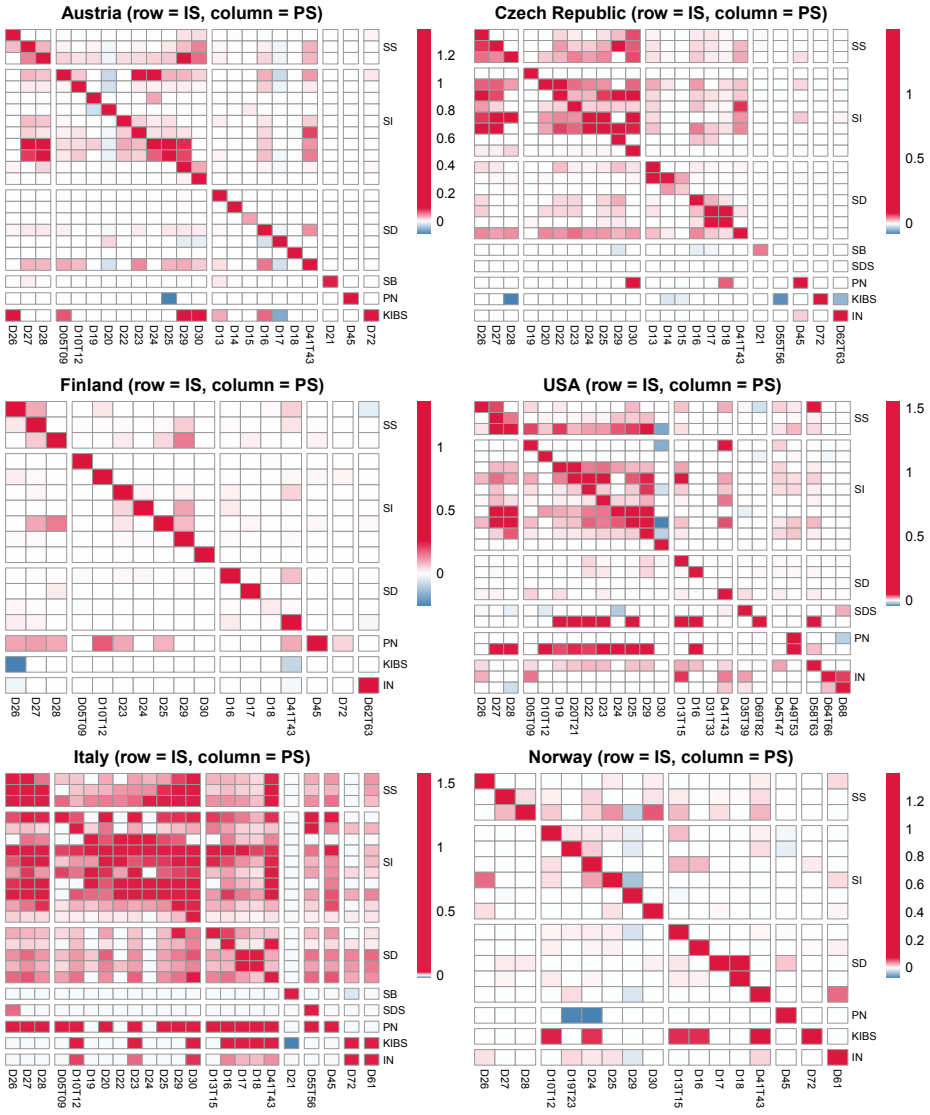


Figure 2.17: Correlation matrices

Note: Rows indicate measures of innovation productivity (growth of labor productivity over R&D input); Columns indicate measures of labor productivity growth.

Chapter 3

ICT's Wide Web: a System–Level Analysis of ICT's Industrial Diffusion with Algorithmic Links

3.1 Introduction

The *system* or *cluster* of technologies labeled Information and Communication Technologies (ICT) came to the attention of economists in the late 1980s due to a mismatch between expected and de facto productivity dynamics and have been studied extensively since then. The intensive investments into ICT fueled the development and production of new IT capital and the expansion of ICT application. Eventually, catering to heterogeneous demand needs, ICT evolved into a complex and interconnected system of technologies that forms an infrastructure with a variety of applications based on it. However, in the literature ICT is very often considered as a monolith at a coarse, aggregate level, while in fact it is composite. This makes it difficult to isolate the effects (both benefits and failures) of ICT diffusion and understand how they are achieved. Firms, entrepreneurs and policy–makers

cannot utilize information on ICT at such aggregated level for strategy and policy design. This leads to coordination failures between ICT supply and demand and makes the development of ICT myopic or haphazard (Bresnahan, 2019b).

In this Chapter, I adopt a systemic approach to ICT by considering industrial diffusion of a set of distinct ICT technologies, each separately as well as in relation to each other within the ICT cluster. More precisely, the analysis estimates the scale and scope of industrial connections for each distinct ICT technology and their dynamics. This reveals directions of development by identifying those ICT technologies that intensify connections with industries by moving closer to the center of the knowledge base and those that experience an exploration phase acquiring new industrial connections. The consideration of the ICT cluster allows putting each ICT technology in context to compare scale and scope of their diffusion not only based on individual growth rates but also in relative terms; this is especially important for the discussion of novel and fast–growing technologies such as Artificial Intelligence (AI). Then, I estimate the relatedness among ICT technologies based on their co–occurrence in the knowledge space and in industries to identify the dimension in which ICT technologies are proximate. The study seeks to identify patterns in the dynamics of industrial penetration by ICT over the period 1977–2020 among 28 EU member states. This helps to estimate the modern state of ICT diffusion and put it in historical perspective.

Methodologically, the ICT cluster is captured combining the new ICT taxonomy by OECD (Inaba and Squicciarini, 2017) and the PATENSCOPE AI index (WIPO, 2019a). The resulting taxonomy used in the Chapter aggregates patents' IPC technological classes into 13 distinct ICT technologies including AI. Economic activities are represented by 74 industries at the 2–digit level from International Standard Industrial Classification of All Economic Activities rev.4 (ISIC). The primary connections between industries and ICT technologies are established through extraction of keywords from ISIC industrial descriptions and their subsequent search in patents' titles and abstracts. Further refinement of these connections with the Algorithmic Links with Probabilities method (Lybbert and Zolas, 2014) produces the final ICT technology–industry matrices, one for each subperiod. As

these matrices are essentially bimodal networks, several network metrics are applied to analyze the structure and dynamics of industry reliance on ICT. Finally, relatedness indicators provide insights into the overlap of knowledge and application bases among ICT technologies.

This Chapter contributes to several literature strands. Given that the analysis is set at the system level and investigates the linkages within the cluster and beyond, this Chapter belongs to the strand of research on technology systems and the industrial connections that they create (Freeman, 1994; Perez, 2010). It fills the gap of empirical studies that operationalize the concept of technology system applied to ICT. Unlike studies that consider ICT as a monolith and estimate its impact on a set of industries (one technology to many industries), for example, through the lens of General Purpose Technologies (GPTs) (Basu and Fernald, 2007; Castellacci, 2010), this Chapter constructs a bimodal technology–industry network connecting a number of well–defined ICT technologies with many industries to uncover the industry portfolio corresponding to each ICT technology and decompose pervasiveness of the whole ICT cluster.

The Chapter builds on studies that focus on the economic impact of ICT (Brynjolfsson and Hitt, 2000; Van Ark et al., 2003; Brynjolfsson et al., 2019) and traces back the source of this impact to particular ICT technologies. By offering an estimation of pervasiveness, this study contributes to the literature on GPTs and their identification (Bekar et al., 2018). Another contribution of this research lies in the field of sectoral patterns of innovation (Malerba, 2002; Castellacci, 2008) by showing the structure of industrial connections through shared technological knowledge base. In this context, the relatedness indicators used in this Chapter represent an instantiation of research on the principle of relatedness (Hidalgo et al., 2018). As the analysis takes a closer look at AI technologies, the study is a contribution to the Economics of AI (Agrawal et al., 2019b). In particular, to the best of the author's knowledge, this work is the first to study together AI diffusion among the industries' knowledge base and the complementarity of AI with other ICT technologies. Finally, considering the methods used, the Chapter is also an application of the "text-as-data" approach (Gentzkow et al., 2019) to technological dynamics.

This study can inform both technological and economic perspectives. From a technology–centered perspective, it offers a fine–grained estimation of ICTs' pervasiveness and diffusion patterns. From an economic perspective, the study pinpoints potential loci of adoption externalities by identifying connected ICT technologies. Establishing influential ICT technologies for each industry provides insights on the technological regimes they induce and consequently on the economic conditions such as innovation opportunities, entry dynamics and market concentration, appropriability, and eventually commercial value and price of products and services. Finally, the industry–technology mapping presented here can help policy–makers in identifying related markets that rely on the same technologies even across industrial boundaries as well as technologies related through the same industries–applications; this is especially relevant for the regulation of merger and acquisition applied to digital markets (see, for example, Federico et al. (2020); Morton and Dinielli (2020)).

The Chapter proceeds as follows. Section 3.2 lays out chronologically the changing focus of ICT research, from macro level productivity dynamics to micro level changes in organisational routines and occupations within firms; from a coarse notion of ICT through multiplex networks of economic complexity to case studies of particular technologies. At the end of this Section, the systemic nature of the ICT cluster emerges as a crucial feature characterizing ICT. This frames the rationale behind the construction of the industry–technology mapping described in Section 3.3, with a further focus on the connections of the ICT cluster: with industries and among the distinct ICT technologies. Section 3.4 discusses the results of the analysis on the within ICT cluster structure, the estimation of the pervasiveness of particular ICT technologies, and their relatedness through knowledge and application bases. Special attention is devoted to AI technology with a deeper dive into the AI's connections with other ICT. Section 3.5 concludes.

3.2 Many Faces of ICT: from Productivity Paradox of the 80s to Modern Technology System

It has been many decades since ICT diffuse inside the economies transforming and creating new markets, business models, and jobs. ICT are enabling technologies engaging in coinvention with application industries to address market demands and organizational supply processes (Bresnahan and Yin, 2017). ICT have induced an encompassing process of digitalization that led to a restructuring of the socio–economic life and redomaining the economy around digital infrastructures.[1] In retrospect, the impact of ICT is immense and visible but complex and often non–linear in how it propagates. The discussion on ICT came across in the context of the first "productivity paradox" or "productivity puzzle" expressed in 1987 by Roach (1987) and Solow (1987) as surging growth of investment in Information Technologies (IT), in particular in computerization, was coupled with retarding growth of productivity. Put simply, the rationale behind attention to ICT is their expected enhancing effect on resource (capital and labor) productivity which at the time was not observed, creating a mismatch between expected increasing and de facto slowing down productivity growth. Eventually, resource productivity and its rate of growth define feasible production volumes hence market expansion and in the end economic growth. Given that productivity dynamics reverberates at economic growth, the first productivity slowdown that lasted until mid–1995 spurred debates on its origins. To investigate this oddity, scholars went from macro trends down to meso and micro data on industries and firms; these studies suggested several potential explanations for observed phenomenon such as an offsetting effect while IT capital substituted for non–IT capital (Dewan and Min, 1997), firm heterogeneity (Brynjolfsson and Hitt, 1995), and mismeasurement (Diewert and Fox, 1999). Indeed, the period from mid–1990s until mid–2000s has been characterized by productivity acceleration in the US, suggesting delayed but substantial contribution of the ICT–producing sectors to aggregate productivity growth and delayed returns on investment in IT capital by ICT–using

[1] For example, in 2018 intangible assets account for 84% of value of S&P500 companies including software code and licenses, data and databases; top–5 largest global companies by market capitalization are digital giants Apple, Alphabet, Microsoft, Amazon, Facebook (Gonzalez and Ponemon, 2019).

sectors due to implementation lags associated with learning, complementary capital accumulation and reorganization (Jorgenson et al., 2003; Basu and Fernald, 2007; Corrado et al., 2007; Bresnahan et al., 2002). Timmer and Van Ark (2005) conduct a comparative study of the EU and the US, regions with not completely synchronized business cycles so that differences related to the time lags are exposed. They find that the contribution to the aggregate productivity by ICT–manufacturing sectors and overall ICT capital deepening are the two factors that explain almost fully the US's lead over the EU in labor productivity growth. While the discussion on the contribution of ICT to economic growth was still ongoing, the second productivity slowdown started around mid–2000s (Fernald, 2015; Syverson, 2017), fueling a new wave of debates between techno–optimists and techno–pessimists. These two research strands differ in predictions for the future of productivity dynamics but they all acknowledge the role ICT played initiating structural changes in production processes, workplace, labor demand, and contributing to overall economic growth (Bresnahan and Yin, 2017).

To capture the profound transformations ICT induced and mechanisms through which these transformations unfold, scholars employ different approaches: (i) inductive inference analyzing the changes at the micro and meso level e.g. demand for skills and income distribution (Autor et al., 1998; Michaels et al., 2014), organizational routines and structures (Brynjolfsson and Hitt, 2000), novel products and services (Bakos and Brynjolfsson, 2001), etc.; (ii) deductive analysis assuming a mechanism that potentially produces a set of observed stylized facts and/or state of the economy e.g. modeling new production factors experimenting with a production function, structural models and simulation and/or more technology–centered framework of General Purpose Technologies (GPT) (Castellacci, 2010; Corrado et al., 2009; Basu et al., 2003; Guerrieri and Padoan, 2007). In both (i) and (ii), ICT is very often considered as a monolith, a coarse notion of all information technologies with differentiation between ICT–using and ICT–producing firms and industries. However, the technology under consideration does make a difference. Given that, another approach to capturing ICT transformations is (iii) studying up close the diffusion of a particular technology considering a larger set of factors including (a) *supply–side* and (b) *technological* ones along with demand–side and economic factors such as preferences and price.

As argued by Rosenberg (1972), (a) "the rate at which new technologies replace old ones will depend upon the speed with which it is possible to overcome an array of supply side problems" and (b) "better understanding of the timing of diffusion is possible by probing more deeply at the technological level itself, where it may be possible to identify factors accounting for both the general slowness as well as wide variations in the rate of diffusion.".

Indeed, at the level of individual technology, the economic value is endogenous to the technological function(s) a particular technology can perform, and how exactly this technology executes the function(s): as infrastructure or network vs fully–fledged component or stand–alone product or service (for example, mobile telecommunication network vs integrated circuit). Moreover, different technologies induce different technological regimes that form around them; in turn, a technological regime defines the environment for innovating agents: opportunity and appropriability conditions, properties and channels of transmission of technological knowledge. These conditions implied by the technological regime reverberate to innovation patterns as well as firm size and entry–exit dynamics at the supply side, and foster industrial dynamics and evolution at various rates and directions (Malerba and Orsenigo, 1997). At the same time, "individual technologies are not introduced in isolation. They enter into a changing context that strongly influences their potential and is already shaped by previous innovations in the system." (Perez, 2010, p.188). These previous innovations might be instantiations of the same technology illustrating path dependency (David, 2007) within one technological trajectory (Dosi, 1982) as well as of another related technology in the *technology system* (Freeman, 1994) capturing dynamic interrelatedness among technologies. Indeed, Freeman (1994) stresses the systemic aspect of technological diffusion. All this applies to ICT as well, hence the ICT cluster can be considered a technology system. Studying ICT as a system of interrelated technologies in connection to industries can provide a better understanding of their diffusion patterns and impact. For example, several ICT–producing industries, being connected via complementary technologies, might engage in synergistic interactions (Steinmueller, 2002) through, for example, (indirect) networks with adoption externality (Church and Gandal, 2005). In the field of ICT, a famous case of strategic exploitation of such positive externalities is the so–called Wintel stan-

dard formed between Microsoft's OS and Intel's processors (Takahashi and Namiki, 2003).

ICT technology system spans over a wide range of industries and hence constitutes a part of their technological knowledge base. ICT carved a slot in the knowledge space becoming a distinct technology system through the tortuous process of upstreaming and technological convergence (Rosenberg, 1963); a complex infrastructure has been built gradually around the function of handling information (Steinmueller, 1996; Greenstein, 2019) performed within commercial as well as military and scientific applications. An increasing number of application industries incorporates ICT in the knowledge base tying their business models, production processes and overall development to the technical progress and manufacturing of ICT goods. On the one hand, this creates an inflow of investments in ICT–producing industries, boosting their growth. On the other hand, numerous application sectors have heterogeneous preferences about the pace of production and performance of ICT capital. Locked inside the so–called dual inducement mechanism (Bresnahan and Trajtenberg, 1995), for example, the semiconductor industry adopted Moore's law as the main roadmap to sustain the demands of its applications and control the pace of development. Nevertheless, preemption strategies and capacity races (Steinmueller, 1992) in producing the next generation of chips have been always present in the industry, contesting the established roadmap. Besides pressures from the heterogeneity of ICT–using industries, a failure to recognize the economic value of ICT application might occur as well; technology–provided opportunities for commercial applications are not always obvious *ex ante* which creates the coordination problem between technical progress in ICT and the technical progress in its applications (Bresnahan, 2019b).

The multiplicity of linkages between economic activities and technologies is highlighted in the literature on regional development and economic complexity (Balland et al., 2019). Such studies represent the path of regional growth through a bimodal network that connects local capabilities[2] and diversity of economic activities in the region (Hidalgo and Hausmann, 2009). Establish-

[2]Capabilities is an umbrella concept that comprises resources, institutional framework, human capital and knowledge (Maskell and Malmberg, 1999).

ing these connections serves two purposes: (i) it uncovers the correspondence between products or industries (economic activities) and the required knowledge or inputs (capabilities) and (ii) reveals related economic activities that require similar capabilities. The latter is measured through various relatedness metrics (Hidalgo et al., 2018). For example, two industries or products can be considered related if they rely on similar technological knowledge (Breschi et al., 2003; Balland et al., 2019), labor skills (Neffke et al., 2011, 2018), or input–output structure (Essletzbichler, 2015). In dynamic perspective, it helps to explain the entry probability of a new economic activity in a spatial unit based on existing local capabilities (industrial change), and the consequent evolution of local capabilities (structural change) in response to the entrance of new industries (Neffke et al., 2011). In sum, in the field of economic complexity, this circular impact between industries and technological knowledge base as an instantiation of capabilities illustrates the underlying mechanism of regional diversification and growth (Frenken et al., 2007). The current study shares this rationale and introduces complexity through the constructed industry–technology mapping with relatedness metrics applied to it. Unlike the economic complexity literature that conventionally focuses on economic activities (industries, products), this Chapter focuses on ICT technologies as part of the technological knowledge base in the EU28 region to study the diffusion of the ICT cluster among industries.

In sum, ICT constitutes a technology system or cluster, has a transformative effect on economic activities that incorporated ICT in their technological knowledge base, and enables industrial synergies. Referring to ICT as a *cluster* or a *system* is crucial because it stresses the composite nature of ICT and exposes the rationale to consider it as a set of interrelated yet distinct technologies. In this Chapter, I adopt this systemic approach to ICT and claim that within the ICT cluster technologies are heterogeneous in their nature, leading to uneven scale and scope of adoption among industries. In other words, not all ICTs are pervasive, not all ICTs are key technologies[3]. A more fine–grained consideration of the ICT cluster can reveal the structure of its pervasiveness by identifying ICT technologies that experience (a) increasing scale of penetration by deepening the connection with industries or

[3] "Key technologies are defined as holding a central position within the knowledge base." (Graf, 2012)

(b) increasing scope of application by creation of new applications/markets or both. The analysis of relatedness is aimed at uncovering the dimensions along which heterogeneous ICT technologies are proximate. Altogether, this draws a more complex and up–to–date picture of the ICT cluster.

3.3 Methodology: Constructing Industry–Technology Mapping

To conduct the analysis, I employ OECD and WIPO patent–based classifications to break down ICT into a set of technologies, and using text mining and probabilistic matching, construct technology–industry nexus tracking its development over time. This nexus is a dynamic mapping between economic activities represented by industries and the set of ICT technologies. Estimating how distinct ICT technologies penetrate industries' knowledge base in dynamics and looking deeper into the origin of these relations provides a more accurate and meaningful view on ICT diffusion. The approach I have employed to construct industry–technology mapping is based on the Algorithmic Links with Probabilities (ALP) method proposed by Lybbert and Zolas (2014). This method allows establishing industry–technology connections in two essential steps: (i) connect industries and patents via the search of keywords extracted from industries' descriptions in patent's abstract and title, and then, based on these links through patents, (ii) connect industries with patents' technological classes that belong to the ICT taxonomy. Further refinement of obtained industry–technology frequency matches implies a transformation of simple cross–tabulation values into Bayesian probabilities.

The ALP method has a number of advantages in comparison with other industry–technology concordances (e.g. Yale Technology Concordance (YTC) (Kortum and Putnam, 1997) and DG Concordance (Schmoch et al., 2003) as it is: (i) modifiable — new keywords, industries, or technologies can be added and linkages easily recalculated without reconstructing the whole mapping from scratch; (ii) dynamic — over time the industry–technology linkages can emerge or disappear; the method is dynamic as it allows con-

struction of the mapping for any defined period; (iii) scalable — technology and industry classifications employed in the study each has several levels e.g. 4–, 3–, 2–digit level; the ALP method can be applied to any combination of classifications' levels; once industry–technology connections are calculated at chosen levels, further aggregation along each classification is possible: for example, 2–digit level (ISIC divisions) can be further aggregated up to 1–digit (ISIC sections) level by simple summation of calculated connections for nested industries. Overall, the ALP method employed for the purposes of this Chapter helps to connect meaningfully industries to large but distinct ICT classes, not to a particular technology confined in a patent.

Industrial classification. The first step of the ALP method requires the extraction of keywords that characterize the economic activity the industry carries out. The description of industries comes from the United Nations' International Standard Industrial Classification of All Economic Activities Revision 4 (ISIC Rev.4) (UN, 2008). It is a suitable choice of meso–level classification that spans over all sectors of the economy from the primary sector with agriculture and raw materials production to the tertiary sector of services such as consultancy, advertising, research, etc. The choice of ISIC taxonomy's depth fell on 2–digit level because this level at the same time allows for a sufficient amount of text description per industry and produces a fine–grained matching between industries and technologies. Overall, there are 74 industries included in the analysis.

Keywords extraction. The purpose of keyword extraction is to create a set of characteristic tokens to represent each industry. The text corpus used for keyword extraction is ISIC rev.4 industrial description with industries at the 2–digit level (division) including their nested levels (3–digit group and 4–digit class) treated as separate documents. The choice to go for the characteristic tokens or phrases, so-called n–grams, is motivated by the need to balance between type I and type II errors: exclusion of useful tokens (false positive) and inclusion of distorting tokens (false negative). The latter means that some words can have multiple and sometimes quite distant meanings while indeed being actively used in a particular economic activity. Thus, the exclusion of such words from the keywords set would harm the representation of an industry while inclusion would confuse/conflate several

industries. Bigrams help to solve this conundrum because they can consist of separately ambiguous words and by combining them create a phrase with a more specific meaning that allows attributing it to a particular industry unequivocally and not losing an important word. Table 3.1 illustrates the described principle with some examples.

Word	Bigram	ISIC
equipment	irrigation equipment	16
	communication equipment	26
	signaling equipment	27
	freezing equipment	28
	dental equipment	32
	optical equipment	33
plant	forage plant	11
	plant propagation	13
	power plant	42
	nuclear plant	71
	sewage plant	81

Table 3.1: Disambiguation with bigrams

Breaking down the text into tokens leads to a large list of single words and bigrams with many of them being redundant. To illustrate the process with an example, one sentence of n words turns into a list with n single words and $(n-1)$ bigrams. Removal of n–grams that contain stop–words, such as articles, forms of the verb *to be*, etc. accounts for only a fraction of cleansing of this list. To select meaningful n–grams after the removal of stop–words, I have applied two techniques (i) Part–of–Speech tagging (PoS) and (ii) calculation of term frequency–inverse document frequency statistic (TF–IDF).

The TF–IDF statistic is a composite indicator that helps to construct a broad representation of what a document is about. Precisely, it estimates how important is an n–gram to a document based on its occurrence frequency within and between documents. The first component, term frequency (TF), is a simple frequency of an n–gram within a document that shows how often the n–gram occurs in the document. The second component, inverse document frequency (IDF), divides the total number of documents by the number of documents that contain the n–gram which reaches its minimum (equals to 1) when the n-gram is found in all documents and its maximum (equals to $log(n)$) when the n–gram belongs to only one doc-

ument. The product of these two parts form the TF–IDF statistic that is high for n–grams that are frequent within one document but is not common for the rest documents. Calculation of TF–IDF is possible for n–gram of any length. In application to ISIC description of industries, the notion of document is equivalent to 2–digit level industry description with all nested 3– and 4–digit level descriptions. More generic words, like already used example of word *equipment*, go down in the ranking because they can occur in many 2–digit industries at the same time.

The PoS tagging is a Natural Language Processing (NLP) technique that helps to identify word's part of speech (noun, verb, adjective, gerund, etc.) given the context of the text. Consideration of the context in identifying word's part of speech is important because of the coincidence between forms of different parts of speech (e.g. *to fish* and *a fish*) which creates ambiguity which tag to attach to a word. Therefore, the PoS method is applied to the raw text to identify contextually the part of speech for every word in the text. As a result, the text is transformed into a lexicon where every word has a corresponding PoS tag. Only then the text is broken down into n–grams which should be sorted in the following way. First, nouns both singular and plural and gerund parts of speech are selected from the words list as potential candidates for keywords. Second, with bigrams the criterion of selection is put on each word separately to create a meaningful combination: the first word can be a noun, gerund or adjective while the second can be still only a gerund or noun.

Stop–words removal and application of these two techniques to the ISIC industrial descriptions provide each type of n–gram (word, bigram) with its PoS tag and TF–IDF statistic. This treatment cleanses the initial list of n–grams to almost purely characteristic phrases. However, the number of key n–grams per industry is not evenly distributed across industries because the original description can consist of few short lines that result in a few key n–grams to extract. To tackle this problem and reduce type I error for industries represented with a small number of tokens, the set of tokens was expanded using two methods: (i) synonyms search and (ii) vocabulary expansion. The first method, expansion through synonyms, uses

PATENTSCOPE's Cross Lingual Expansion.[4] This tool provides synonyms found in patents' texts based on selected technological domains. The same tool can help with finding synonyms for words with many meanings that differ conditional on the context hence it also helps to act upon type II error, the same problem as the usage of bigrams does. To illustrate the mechanism at work, consider a word *skin* whose lexical connotation changes from textile industry to pharmaceutical one. Thus, for textile, the synonyms can be *leather*, *pelt*, and *hide*, while for pharmaceuticals *derma* can be used.

The second method, vocabulary expansion, uses patents selected during the first round of search of the preliminary list of keywords and extracts additional keywords relevant to the economic activity from titles and abstracts of these selected patents. Consider an industry with k key tokens (both words and bigrams). These k tokens are found in n patents' abstracts and titles. Most of these n patents contain only one token, less contain two tokens and so on; in other words, such histogram is skewed resembling some asymmetric distribution like Pareto or exponential. I select a subset of m patents out of these n where 2 or more tokens are found.[5] The titles and abstracts of the subset m patents are broken down into tokens as well; stop words removal and TF–IDF are applied to extract additional characteristic tokens. Given that a patent mostly contains information about technology and only a small share of patent's abstract might describe the application related to the economic activity, tokens extracted in this way were manually revised and selected. The final list of key n–grams contains slightly more than 4800 regularized n–grams for 74 2–digit industries.[6]

Algorithmic Links with Probabilities. The link between an industry and a technology is established through searching a key token extracted from the industry's description in the patent's title and abstract. Thus, technology areas that a patent belongs to according to International Patent Classification (IPC) serve as an approximation for technologies that industry might rely on. The sample of patents to construct the mapping is limited

[4] https://patentscope.wipo.int/search/en/clir/clir.jsf.

[5] i.e. the full set of patents found with k key tokens is $f(1 \leq x \leq k) = n$ while a subset of patents that contain at least two key tokes out of k is $f(2 \leq x \leq k) = m$ where $m \leq n$

[6] The quantiles for the number of key tokens per industry: $Q25 = 34, Q50 = 50, Q75 = 74$

to original European patents where either inventor or an applicant is located in one of 28 European countries (EU28) based on the data from the OECD REGPAT 2020 database.[7] The data on abstracts and titles of EU28 patents is retrieved from the EPO Worldwide Patent Statistical Database (PATSTAT) Spring 2020.

The ALP method is extensively described in Lybbert and Zolas (2014), hence here I will outline only its essence and some important details for this study. Once a key token for the industry is found in the title or abstract of a patent, it also creates a connection between the industry and the IPC classes of the patent.[8] The IPC classes represent technologies. Therefore, the outcome of the first step of the ALP method is a matrix with a simple count of matching between industries and IPC classes. To transform the raw count into the industry–to–technology ALP concordances, Bayes rule is applied. The resulting ALP concordances are Bayesian probabilities adjusted to account for some technological fields that can be naturally very prolific in patenting hence by size compress shares of other, less prone to patenting, technological fields in industry's recipe. Thus, the share W_{ij}^{H} of technology j (IPC_j) in industry's i ($ISIC_i$) technological recipe is calculated according to the following formula:

$$W_{ij}^{H} = \frac{Pr(ISIC_i|IPC_j) \times W_{ij}^{R}/J}{Pr(ISIC_i|IPC_1) \times W_{i1}^{R}/J + ... + Pr(ISIC_i|IPC_J) \times W_{iJ}^{R}/J} \quad j = \overline{1, J} \tag{3.1}$$

$$W_{ij}^{R} = \frac{Pr(ISIC_i|IPC_j) \times Pr(IPC_j)}{Pr(ISIC_i|IPC_1) \times Pr(IPC_1) + ... + Pr(ISIC_i|IPC_J) \times Pr(IPC_J)} \tag{3.2}$$

where J is a number of IPC classes at the 4–digit level. Inevitably, some rubbish connections can show up due to a large number of abstracts and titles and various key n–grams. To eliminate such connections the cut–off threshold is set at 2% meaning shares W_{ij}^{H} lower than 0.02 are set to zero and the remaining shares are renormalized to sum up to 1. In sum, the

[7] Consideration of EU28 countries that include the UK is motivated by the UK's participation in the EC and later in the EU since 1973. Given that the time period in the study covers years from 1977 till 2020, it justifies the usage of the data on EU28.

[8] In this study the 4–digit level of IPC classes is taken for the mapping construction.

ALP concordance matrix W^H for each period consists of 74 2–digit ISIC industries and 638 technologies at 4–digit level of IPC classes (IPC4). The choice of IPC4 level is motivated by construction of the ICT taxonomy so that it allows application of the ALP method; the next paragraphs discuss it in details.

ICT cluster. The constructed mapping is used to track the diffusion of the ICT cluster over time. The representation of the ICT cluster expressed in patents' IPC classes is constructed for this study combining two taxonomies: (i) the new ICT taxonomy of OECD (Inaba and Squicciarini, 2017) and (ii) PATENTSCOPE AI Index (WIPO, 2019a). The new ICT taxonomy by OECD provides concordance between 13 ICT classes such as High speed network and Mobile communication and various–level IPC classes. This taxonomy is taken as the main structure of the ICT cluster. The modification of the OECD taxonomy concerns its class Cognition and meaning understanding which is turned into a class of Artificial Intelligence (AI). The rationale behind is that original class Cognition and meaning understanding represents a subset of AI while the latter gains place inside the ICT cluster as a distinct class of technologies by entering the active commercial phase and experiencing intensive development and experimentation. First, I have identified that IPC classes which represent AI techniques and AI functions from WIPO's AI Index are found only in three ICT groups of OECD taxonomy: (i) a substantial overlap with class Cognition and meaning understanding, (ii) Imaging and sound technology and (iii) Others. Then, to create AI class on the basis of Cognition and meaning understanding, all WIPO's AI–related IPC classes were excluded from (ii) and (iii) and transferred to (i) so the 13 ICT classes remain mutually exclusive. Table 3.2 contains the resulting ICT taxonomy.

Overall, 55 4–digit IPC classes constitute the ICT cluster. Limiting the constructed mapping to these 55 IPC classes and their subsequent aggregation into 13 ICT classes focuses the attention on the ICT cluster. However, according to the ICT taxonomy, only a fraction of 4–digit IPC class might be related to the ICT cluster. Moreover, the distinction between one or the other ICT group can occur at a deeper level of IPC class than 4–digit; an ICT group can consist of various–level IPC classes at the same time such as

ICT class	Description
cl1	High speed network
cl2	Mobile communication
cl3	Security
cl4	Sensor and device network
cl5	High speed computing
cl6	Large–capacity and high speed storage
cl7	Large–capacity information analysis
cl8	Artificial Intelligence*
cl9	Human–interface
cl10	Imaging and sound technology
cl11	Information communication device
cl12	Electronic measurement
cl13	Others

*Created on the basis of class Cognition and meaning understanding by merging with PATENTSCOPE AI Index taxonomy

Table 3.2: The modified new ICT taxonomy by OECD

4–digit along with fine–grained 8–digit. In other words, the ICT taxonomy is constructed on varying IPC level.

For example, in the upper panel of Table 3.3 distribution of IPC4 class G06K among 13 ICT groups is shown. The division of IPC G06K between ICT classes occurs on a 5– and 6–digit levels and one subclass, G06K21, doesn't belong to ICT at all. Therefore, counting every patent with G06K IPC class as 1 for each ICT group connected with G06K — ICT classes 3, 6, 8, 9, 13 — would inflate the size of each ICT class as not all patents with G06K IPC4 class belong to ICT class 6 or 13. In general, the ALP method uses a single level of IPC to construct concordance, e.g. 4–digit, and cannot use different levels at the same time. For example, a concordance matrix constructed at 5–digit level can be aggregated to 4–digit level, but it cannot contain 4– and 5–digit levels simultaneously. Thus, to keep the ALP concordances on the chosen 4–digit level with respect to the technologies and avoid double counting, the following shares are calculated based on the *sample*. First, the between share, the de facto share of an IPC4 class that belongs to the whole ICT cluster, without division into 13 groups. According to the lower panel of Table 3.3, in the third period, 99.9% of patents with IPC4 class G06K belong to the ICT cluster i.e. patents' IPC classes (4–digit and longer) match with IPC classes listed in the ICT taxonomy. This means that in the EU28 sample, there is only 0.1% of patents that belong to non–ICT subclass G06K21 in the third period. Second, the within share, the de facto

share of an IPC4 class that relates to one of the 13 ICT classes. In other words, the between share is now further decomposed into 13 shares, each for one ICT class, that are renormalized to sum up to 1. In the example of G06K class, the 99.9% between share is decomposed into within ICT cluster shares shown in the lower panel of Table 3.3. For instance, the 29.8% within share of class 3 Security means that inside the 99.9% between share of IPC4 class G06K there are 29.8% of patents that belong to longer IPC subclass G06K19.

Decomposition of G06K 4–digit class among ICT groups based on the OECD taxonomy														
G06K	cl1	cl2	cl3	cl4	cl5	cl6	cl7	cl8	cl9	cl10	cl11	cl12	cl13	non-ICT
			G06K19			G06K1		G06K9	G06K11				G06K15	G06K21
						G06K3							G06K17	
						G06K5								
						G06K7								
						G06K13								
Within and between shares based on actual patent count in the sample														
G06K	cl1	cl2	cl3	cl4	cl5	cl6	cl7	cl8	cl9	cl10	cl11	cl12	cl13	Between ICT
Within ICT (% of between)	0.0	0.0	29.8	0.0	0.0	19.1	0.0	0.0	48.0	0.0	0.0	0.0	3.1	99.9

Table 3.3: The between and within shares of G06K class in the ICT cluster in the third period 2005—2020

In sum, the IPC4 level for the study is a convenient choice as 4–digit level is the highest among IPC classes used in the ICT taxonomy because combined with the between and within ICT shares, it allows (i) working with a single level of IPC classes to apply the ALP method and (ii) avoid double the counting by tracking precisely the share of the whole ICT cluster in the technological composition of industries and the shares of its 13 classes separately.

Figure 3.1 shows the stepwise procedure that leads to obtaining industry–ICT class matrices for each period. Starting from extraction of preliminary keywords, the subsequent synonym and vocabulary expansion result in the final list of 4.8 thousand tokens that characterise 74 industries. Frequency matches between industries and IPC4 patent classes that represent all technologies are further transformed into Bayesian probabilities. The modified ICT taxonomy created based on merged OECD and WIPO classifications is applied to subset ICT technologies (55 IPC4 classes out of 638) obtaining industry–ICT matrices W^H for each period. Finally, the correction with between and within shares is applied to avoid inflation of ICT groups' sizes.

Network analysis and relatedness. As a next step, the industry–ICT

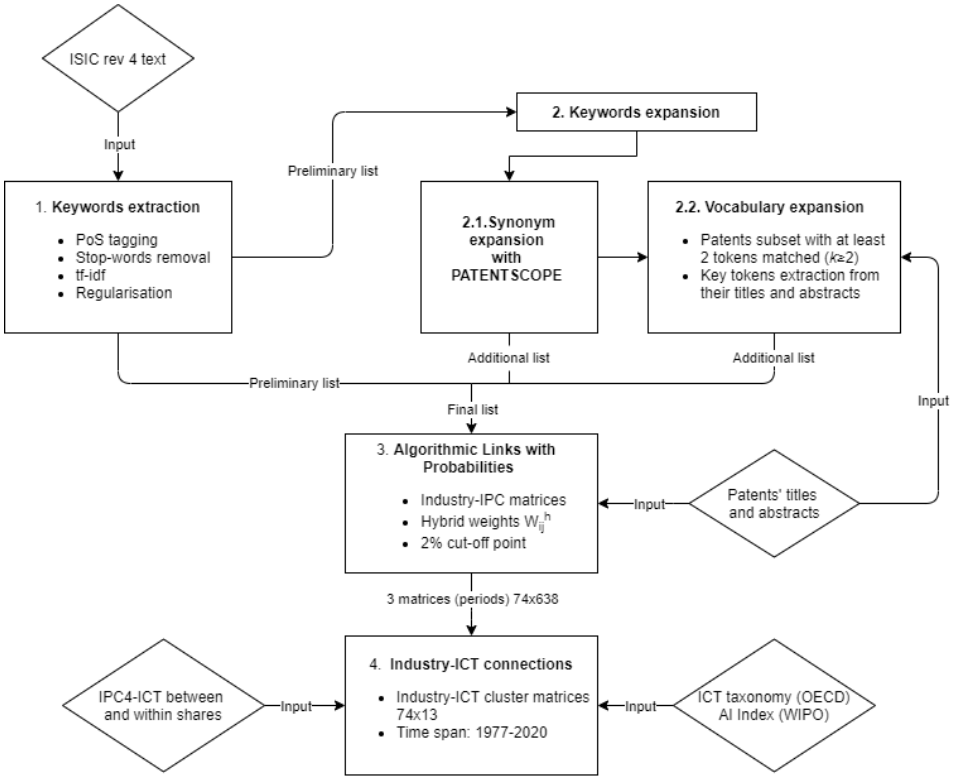


Figure 3.1: Procedure of obtaining industry–ICT concordance with used techniques and data

class matrices are transformed into bipartite networks. In this study, ICT technologies represent a subset of technological knowledge base i.e. proxy for a part of local capabilities. The set of industries reflects diversity of economic activities in the EU28 region related to ICT technologies. This part of analysis is focused on ICT technologies unlike studies on economic complexity that conventionally focus on economic activities. However, the Chapter shares the methodological approach and the stance on capability–activity dynamic interrelatdness as an important mechanism of growth.

Network analysis provides a variety of tools to uncover structural properties of a network. For example, flow betweenness centrality (FBC) is used in this Chapter to rank ICT classes according to their mediation role in connecting industries. It is worth noting that industry–ICT class connections are shares (W_{ij}^H from Equation 3.1) or probabilities hence the higher the share of a particular ICT class in a technological recipe of an industry, the stronger

the connection. This explains the rationale behind the choice of FBC, as it accounts for the strength of connections between an ICT class and all incident industries, and thus shows *a weighted scale of industry reliance on each ICT class*. Nevertheless, in the context of industries–ICT relationship, a node's degree carries useful information about *the scope of industry reliance on each ICT class*. These two indicators provide further details for a more fine–grained consideration of otherwise homogeneous ICT cluster.

Another dimension to study ICT technologies is their relatedness. As pointed out earlier in the Chapter, very often ICT technologies are pooled together under an umbrella label of generic ICT or digital technologies. However, not only ICT cluster comprises a wide array of distinct technologies, their alleged similarity can be tested. To do so, I calculate two metrics, technological and application relatedness, to measure the proximity of ICT technologies with regard to the underlying knowledge and application bases respectively.

The *technological relatedness* applied to ICT technologies shows which of them are rooted in similar knowledge base i.e. related through shared knowledge areas. This is measured as a standardized frequency of co–occurrence of ICT classes in patents. If, according to the modified OECD taxonomy, IPC classes of a patent belong to two different ICT classes this represents an instance of co–occurrence; the total number of co–occurrences of two ICT classes is an absolute frequency. However, the absolute frequency of co–occurrence can be a misleading indicator of technological relatedness for two reasons. First, a larger ICT class has a higher potential to co–occur with other ICT classes even by randomness. Second, in absolute terms frequency of co–occurrence of two small ICT classes, for example, i and j, might look negligible in comparison with co–occurrence of two large ICT classes, p and k, and yet represent significant technological relatedness of i and j. Therefore, a test for the randomness that accounts for the size of the ICT classes (number of patents assigned to each of them) must be conducted. Following Breschi et al. (2003), I assume that the frequency of co–occurrence of any two ICT classes, O_{ij}, is a hypergeometric random variable X_{ij}. Thus, ICT classes i and j co–occur in exactly x patents with the following probability, mean and variance:

$$P[X_{ij} = x] = \frac{\binom{R_i}{x}\binom{T - R_i}{R_j - x}}{\binom{T}{R_j}} \quad \mu_{ij} = \frac{R_i R_j}{T}; \quad \sigma_{ij}^2 = \mu_{ij}\left(1 - \frac{R_i}{T}\right)\left(\frac{T}{T-1}\right) \tag{3.3}$$

where T is the number of patents in a particular period, R_i and R_j patents belong to ICT class i and j respectively. The standardized frequency of co–occurrence (with zero mean and unit variance) is equal to:

$$\tau_{ij} = \frac{O_{ij} - \mu_{ij}}{\sigma_{ij}} \tag{3.4}$$

The τ_{ij} statistic takes into account the size of the ICT classes and compares the actual co–occurrence with the expected one as if patents were assigned to ICT classes randomly. If the statistic is significant the null hypothesis of a random relationship between two ICT classes is rejected; the positive statistic $\tau_{ij} > 0$ suggests the presence of technological relatedness of two ICT classes while the negative one $\tau_{ij} < 0$ means two ICT classes occur together even less often than by random assignment.

The *application relatedness* is a correlation that measures how (dis)similar the structure of industry reliance for each pair of ICT classes. There are two significant differences of this novel indicator in comparison with other relatedness metrics used in the literature on economic complexity. First, as the focus is placed on technologies and not on industries, the application relatedness indicates proximity of two technologies through linkages to shared industries i.e. shared application base for the two technologies. Thus, the logic of the application relatedness is reversed with regard to co–occurrence based industry relatedness. For example, Neffke et al. (2011) derive relatedness of industries "from the co–occurrence of products that belong to different industries in the portfolios of manufacturing plants". Technically, Neffke et al. (2011) calculate product relatedness, and the industry relatedness derived from it is called *revealed relatedness* because it is measured through the intermediate layer of products. However, the overarching logic of industrial relatedness as co–occurrence of industries in plants stands. Re-

versing this logic, I define application relatedness as co–occurrence of ICT technologies in industries using industry–technology matrices W^H.

To calculate co–occurrence based relatedness for a set of technologies, for example, a technology–to–technology adjacency matrix is used. Each cell of such matrix contains the co–occurrence frequency of two respective technologies in all relevant industries pooled together and each counted as one. This might lead to an incorrect estimation of relatedness between technologies. For example, in a conventional relatedness metric, if two technologies are jointly present in a particular industry, this would add one to the overall co–occurrence count. Such binary count and summation to obtain frequencies do not include useful information respectively on (i) the strength of industry–technology connections and on (ii) the set of particular industries relevant to each of the two technologies. Together (i) and (ii) constitute *a distribution or a structure of industrial connections of each technology.* Thus, the second difference of the suggested application relatedness is that it captures the similarity of distribution of industrial connections for a pair of technologies. By accounting for both (i) and (ii), the metric can capture relatedness of technologies that both have weak connections with industries but the distribution of their connections is similar for the two technologies.

The application relatedness calculates the correlation between two columns that represent two ICT technologies of the industry–technology matrix W^H.[9] In each period there are m industries (indexed $i = \overline{1,m}$) that rely on at least one of the 13 ICT classes (indexed $j = \overline{1,n}$ where $n = 13$) (see Appendix 3.7) for some share or weight W^H_{ij}. For example, for ICT class $j = 1$ there are $k < m$ industrial connections[10] $W^H_{\overline{i=1,k},j=1} = (W^H_{1,j=1}, W^H_{2,j=1}, ..., W^H_{k,j=1})$. This vector represents the structure of industry reliance on ICT class $j = 1$. The same vector exists for ICT class $j = 2$ (and any other class in the ICT cluster). Therefore, the application relatedness between class 1 and 2 is

[9] Undoubtedly, this is a coarse, linear approximation of the relationship among ICT classes through industries. One can test for non–linear relations using polynomial models or various link functions under the Generalized Linear Model (GLM) framework.

[10] The length of all vectors is m as the remaining $m - k$ connections that do not exist between an ICT class and an industry are set to 0.

calculated as:

$$r_{1,2} = corr(W^H_{i,j=1}, W^H_{i,j=2}) \quad \forall \quad i = \overline{1,m} \tag{3.5}$$

Together the two indicators, application and technological relatedness, create four combinations characterizing similarity among ICT technologies. There are shown in the quadrants of Table 3.4. This scheme provides a useful framework to identify potential loci of adoption externalities and drivers of development for each pair of ICT technologies. As implied by the notions of dynamic interrelatedness and dual inducement mechanism, the development of a particular ICT technology is linked to and influenced by both its related technologies and industries. For example, Bresnahan (2019b) distinguishes between two types of ICT, namely scientific and engineering ICT and commercial and enterprise ICT, and argues that innovation processes occurring in each domain are subject to different factors. In a nutshell, the invention of a scientific and engineering ICT application follows "purely technical requirement" while "[t]he invention of the applications of ICT in much of commercial and enterprise uses necessarily takes the analysis outside "purely technological level"" and follows "visibility" or obviousness of application. Thus, for example, in the first and the forth quadrant, where the technologies are proximate in terms of underlying knowledge, it is easier to find a common technical ground for two ICT technologies and innovate based on available technical possibilities. While in the second quadrant, where ICT technologies exhibit similar strength of connections across shared set of industries but not proximate in knowledge space, the invention might indeed follow "visibility" of commercial value. In general, if a pair of ICT technologies exhibits high application relatedness, one might observe bundling of products and services embodying these two technologies inside a shared set of industries as it is motivated by commercial value. If technologies are not only co–present in an industry but also complementary, that will produce adoption externalities which imply even bigger commercial value but also larger risks (for concrete examples see Simcoe and Watson (2019)). An even further degree of integration between two application–related technologies is mergers and acquisitions among firms producing products and services that have commercial value in bundling and/or exhibit adoption externalities. Therefore, the estimation of application relatedness can inform regula-

tors and policy–makers by identifying related markets even across industrial boundaries connected by technologies that themselves are not related in the knowledge space.

		Technological relatedness			
		low	**high**		
Application relatedness	**high**	Different knowledge, similar industries	Similar knowledge, similar industries	**similar**	**Application base**
	low	Different knowledge, different industries	Similar knowledge, different industries	**different**	
		different	**similar**		
		Knowledge base			

Table 3.4: Relatedness space with four distinct quadrants

In sum, the rationale behind the construction and usage of the mapping with the ALP method is the dynamic nature of constructed concordances as that allows capturing the changing reliance of industries on a set of ICT technologies; some industry–technology connection might decline and some might emerge. On the one hand, the invention and patenting process can last for years and it also takes time for a new technology to enter the knowledge base of an industry, hence it makes sense to use longer time periods to construct ALP concordances. On the other hand, considering too long periods would extinguish the dynamic nature of the concordances. Therefore, I split the whole time period into three subperiods approaching 15 years length which also creates a nearly even distribution of patent sample over these three periods: (i) 14 years: 1977—1990; (ii) 14 years: 1991—2004; (iii) 16 years[11]: 2005—2020. Finally, the chosen industry and technology levels for which ALP matrices are constructed can be changed by aggregation of matrices' values instead of their recalculation. For example, the industrial level can be aggregated going from 2– to 1–digit ISIC codes without chang-

[11]The last period from 2005 till 2020 includes more years because the patenting activity of the last couple of years is still ongoing and technically they are not fully represented yet

ing the patents' IPC level or layering up another industry–IPC concordance for the novel combination of levels.

3.4 Results and Discussion: Inside the ICT Technology System

The application of the methods outlined in Section 3.3 allows considering the ICT cluster not as a monolith and look behind the common notion of "pervasive ICT". First, I will provide initial findings based on a bird's eye view of the cluster. Then I proceed to a more fine–grained level and analyze the ICT cluster as a collection of distinguished technological classes.

Industrial diffusion of the ICT cluster. In this study, the whole economy of EU28 is represented by 74 industries. The construction of ALP concordances reveals that over the time span from 1977 to 2020 the number of industries relying on the ICT cluster increases from 26 to 36 (see Appendix 3.7), covering almost half of the industries. This means that one third of all economic activities in the first period and nearly half in the last systematically incorporates technologies from the ICT cluster into its knowledge base.[12] Thus, the scope of ICT application grows over time. Nevertheless, the intensity (scale) of reliance on the ICT cluster is distributed unevenly across industries as it is shown in the left panel of Figure 3.2.

The left panel of Figure 3.2 plots the share of the whole ICT cluster in the technological recipe of industries. The industries that exhibit the strongest connection with the ICT cluster are not surprising: programming and broadcasting (code 60), information services (63) and telecommunications (61), motion picture, video and television programme production, sound recording and music publishing activities (59), computer programming and consultancy (62), manufacture of computer, electronic and optical products (26). The right panel of Figure 3.2 displays the first differences of the ICT cluster share in the total technological composition of an industry with the first period as a baseline; it shows that for the majority of industries the connection

[12] It is worth noting that this estimation is a lower bound because it is based solely on patent data.

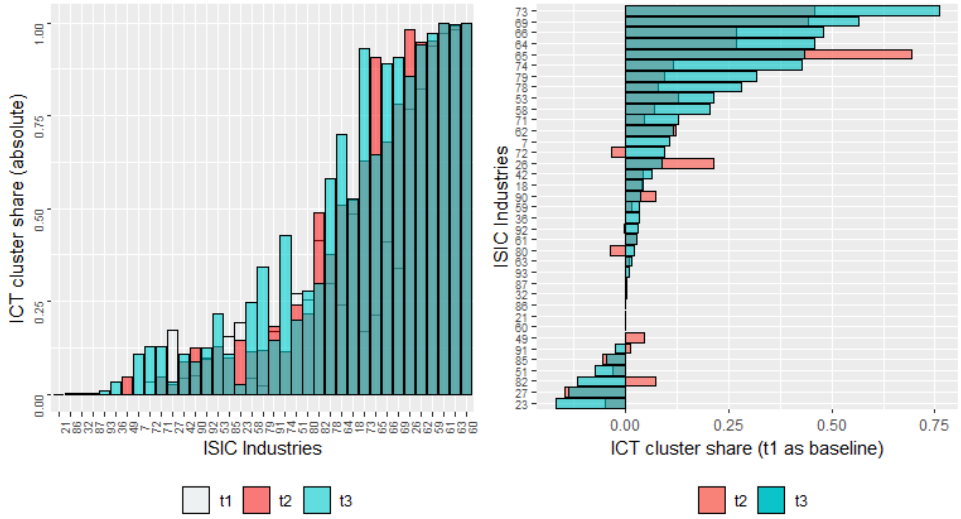


Figure 3.2: The share of the ICT cluster in technological recipes of industries

Note: Color-period correspondence: period 1 – white, period 2 – red, period 3 – blue. The bars are overlaid with transparency parameter allowing to see all three periods at once. In the left panel, for example, for industry 65 the red bar being the tallest means that in period 2 this industry had the highest share of ICT cluster in its technological composition in comparison with two other periods. In the right panel, period 1 is taken as a baseline by subtracting the share of the first period from the second and third; all bars to the right side of the vertical axis indicate increasing share of the ICT cluster in industries' recipes while the left ones — decreasing.

with the ICT cluster is strengthening. The top–3 industries that experience the highest growth are 73–Advertising and market research (in accord with Anderson (2012)), and 69–Legal and accounting activities followed by the whole Section K: Financial and insurance activities (divisions 64–66). Overall, the majority of industries that exhibit at least some reliance on the ICT cluster belong to services starting from Section H: Transportation and storage onward.[13]

Inner structure of the ICT cluster. The modified OECD taxonomy (see Table 3.2) allows looking deeper into the ICT cluster distinguishing 13 classes of ICT technologies. Figure 3.3 helps to conduct an inspection of the ICT cluster composition based on the sample of EU28 patents.

Notably, the leaders of the ranking in the left panel of Figure 3.3 by absolute share of an ICT class in the ICT cluster — classes High speed network, Information communication device and Large-capacity and high speed storage,

[13]The share of industries related to the ICT cluster in Sections H, J, K, M, N, P–R on ICT ranges between 60–100%.

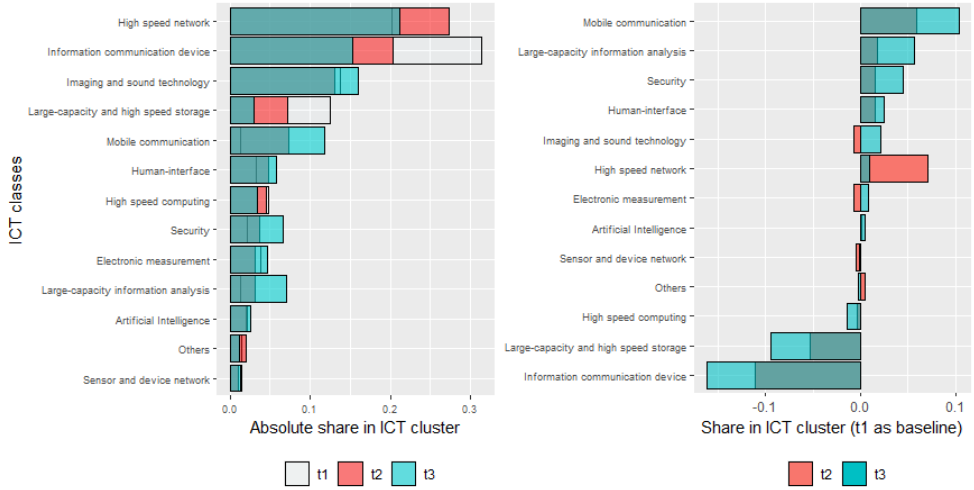


Figure 3.3: Shares of ICT classes in the cluster

Note: Color-period correspondence: period 1 – white, period 2 – red, period 3 – blue. The bars are overlaid with transparency parameter allowing to see all three periods at once. The left panel plots shares in all three periods while the right panel takes the first period as a baseline subtracting it from the second and third periods. For example, class Mobile communication exhibits its highest share in the ICT cluster in the third period as the blue bar is the tallest. Transparency allows seeing the height of the red bar that indicates share of Mobile communication class in the ICT cluster in the second period.

— experience a decrease of their shares in the ICT cluster though keeping strong positions. Instead, a subset of technologies such as Mobile communication, Large–capacity information analysis, Security, and Human interface steadily increase their presence in the ICT cluster indicated by their top positions in the right panel of Figure 3.3 (first differences with the first period as a baseline). This can be viewed as a sign of structural change experienced by the ICT cluster: from building bulk elements of the infrastructure to transmit data, for example, IDSN and fiber optic networks (Greenstein and Spiller, 1996), Next Generation Networks (Fitchard, 2003), to working on the functionality for numerous applications to make the infrastructure more agile, scalable, secure and affable. The latter can be exemplified with broadband cellular networks such as 4G and 5G, cloud computing, cybersecurity such as HTTPS protocol, virtual assistants, and proliferation of frontend GUIs and VUIs (Graphical and Voice User Interface). In other words, soon after the turn of millennia the completion of the physical infrastructure and its operation processes were mostly over, creating a coherent platform for applications' deployment; the new vector of ICT development is oriented at the improvement of specific aspects of the constructed platform for both businesses and end users. Probably witnessing the end of the

first phase of ICT, Nicholas Carr wrote: "While no one can say precisely when the buildout of an infrastructural technology has concluded, there are many signs that the IT buildout is much closer to its end than its beginning" (Carr, 2003, p.10). The first and the third period exemplify the two outlined phases with the second period being transitional where continued construction of bulk elements like Content Delivery Networks (CDNs) and datacenters (though already improving the capability to distribute workloads dynamically) (Greenstein, 2019) started sharing the spotlight with mobility of access and application–oriented development of the ICT, for instance, the rise of Application Programming Interface (API) at the turn of millennia. The phase of establishment of the bulk part of ICT is captured in the literature on Large Technical Systems (LTS) (Mayntz and Hughes, 1988) since an early prominent instantiation of ICT was telecommunication infrastructure (Davies, 1996). Later developments of ICT as control systems rooted in IT (Nightingale et al., 2003) already have a flavor of application–oriented mode. Beyond a systemic perspective of LTS theory, some studies consider specific ICT industries and mechanisms that grew atop the ICT infrastructure — Bresnahan et al. (2014) for mobile applications, Moore and Anderson (2012) for internet security, Jian et al. (2012) for the user–contributed production model of information goods — just to name few. In sum, in line with other studies, the results of the analysis indicate that the evolution of the ICT cluster is taking place and seems to have a direction towards agile, scalable, and omnipresent (mobile access) configuration with myriads of applications and devices. Thus, using the language of Helpman and Trajtenberg (1994), the next cycle of reaping the benefits of ICT development is on its way with the "new fruit" of ICT yet to be plucked.[14]

Moreover, if this path is to continue even further, more layers of functionality will be created making every next "frontend" layer more distant from the previous, lower level layers and providing more opportunities for forking. This calls for an inclusive process of integration, standardization, and compatibility preserving conditions for fair competition and accounting for societal welfare. The problem of achieving an inclusive consensus is already prominent in markets for ICT goods and services that are characterized by

[14] As opposed to the view of techno–pessimists such as Gordon (2016) and Cette et al. (2016)

strong network effects (Shy, 2011) and grows larger due to the rise of platform business models (Belleflamme and Peitz, 2018); each firm has clear incentives to lock the network effects on itself, devising various strategies to cut-off, outpace or acquire competitors (Simcoe and Watson, 2019; Park et al., 2018; Cabral, 2018). Digital platforms and mega–apps like WeChat created by Tencent, GAFAM[15], Baidu, Alibaba tech giants are examples, on the one hand, of success of these strategies but severely damage competition on the other hand (Prat and Valletti, 2019; Laitenberger, 2017). Thus, the current challenge for ICT development is to resolve the tension between the push for monopolization and pull for integration and compatibility (Gandal, 2002; Doganoglu and Wright, 2006) to achieve effective functioning within and among ICT applications and devices and yet preserve fair market conditions.

Pervasiveness of ICT classes. The estimation of the scope and scale of the economy's reliance on a particular technology plays a role in the estimation of the pervasiveness of the technology under consideration. In turn, the question of pervasiveness or general applicability is one of the cornerstones of the General Purpose Technology (GPT) theory (Bresnahan and Trajtenberg, 1995). There are ongoing debates on the nature of pervasiveness and its measurement: using patent data (Hall and Trajtenberg, 2006; Feldman and Yoon, 2012; Graham and Iacopetta, 2014) or industrial diffusion patterns (Jovanovic and Rousseau, 2005; Castellacci, 2010). In this Chapter, I offer an alternative measure of pervasiveness based on the estimated scale and scope of reliance on ICT classes among industries derived from patent data. This goes in line with the view of Bekar et al. (2018) who suggest (i) to define GPTs "according to their micro–technological characteristics, not their macro–economic effects", and (ii) to assess pervasiveness as one of such characteristics by distinguishing between cases when technology is (a) *widely used* and/or has (b) *many uses*. The former means that most of the economy relies on technology *at scale* even if the technology has a single application, while the latter implies distinguished and multiple ways of using this technology or, in other words, a big *scope* of application. Thus, the calculated indicators of the scale and scope mirror this perspective on pervasiveness, and can contribute to the thread of literature on the empiri-

[15]Google, Amazon, Facebook, Apple, and Microsoft

cal testing of technological pervasiveness by applying them to ICT classes. It is worth noting that in this Chapter I do not test whether either of the ICT classes is a GPT; instead, I estimate their pervasiveness in a novel way following the theoretical reasoning of Bekar et al. (2018).

As already mentioned, the industry–ICT matrices are in essence bipartite networks with industries and ICT classes being two different types of nodes. As explained in Section 3.3, the flow betweenness centrality (FBC) is used to proxy the scale of industry reliance on each ICT class. In application to the industry–ICT network, unweighted betweenness centrality would produce a similar picture compared to the calculation of the degree metric reflecting the number of incident industries. Instead, the FBC indicator takes the strength of the connection (i.e. weight of an edge) into account in the construction of the shortest path. Thus, FBC captures ICT classes that might be less frequently connected (in terms of the number of incident industries) but compensate for that by being intensely connected. This convenient property of the FBC indicator allows comparing it against the degree metric to derive conclusions about the scope and scale of reliance on each ICT class. In Figure 3.4, each observation is an ICT class (see Table 3.2) with the change of the degree and FBC indicators between the first and the third period as coordinates; the size of observations is defined by the magnitude of FBC of this ICT class in the third period.

A comparison of Figure 3.4 and Figure 3.3 exposes the fact that the most central and connected ICT classes are not necessarily the most represented ones in the ICT cluster. Significantly smaller classes like Large–capacity information analysis (cl7) and Security (cl3) forged ahead of top–3 large classes High speed network (cl1), Information communication device (cl11), and Imaging and sound technology (cl10) in terms of increase in scope and scale, with class 7 acquiring a leading position with respect to both indicators and class 3 overcoming in scope and getting to parity in scale (see Appendix 3.8). In general, the biggest ICT classes have a moderate number of incident industries but with stronger connections; by contrast, many smaller classes have numerous but weaker connections. The exceptions from this pattern are the two named classes 7 and 3.

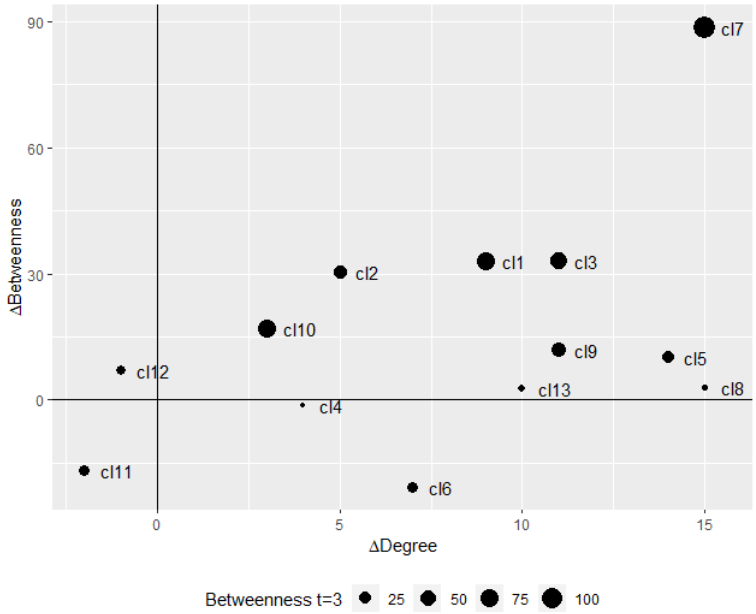


Figure 3.4: The change of scope (Δ degree) and scale (Δ FBC)

Note: Δ Degree (x–axis) and Δ Betweenness (y–axis) are the first differences (change) of the respective metrics between the first and third periods. The size of the observations represents the absolute magnitude of the Flow Betweenness Centrality metric in the third period.

Combining empirical estimations of scale and scope of diffusion and theoretical notions on pervasiveness, class 7 appears as the fittest candidate to be called pervasive by having many uses and being widely used. Classes 1, 3, and 10 are the closest competitors, though they significantly lag behind with respect to either scope or scale. An interesting dynamic unfolds for class 8 Artificial Intelligence (AI): it shows the largest increase in scope (the same as class 7) yet negligible growth in scale displaying an overall small absolute magnitude of the latter. Thus, class 8 AI has many applications but each of the applications doesn't rely at scale on AI. This finding of multiple but yet "shallow" applications is consistent with observations on AI diffusion (for example, as noted by Bresnahan (2019a) and Brynjolfsson et al. (2019)). A reasonable question arises: is AI pervasive? Turning again to Bekar et al. (2018) as the source of the proposed definition of the pervasiveness, the authors suggest that pervasive (in a GPT sense) can be called a technology that is rather widely used (at scale) than the one that has many uses (big scope); AI's pervasiveness can be overestimated because it is conflated with the pervasiveness of the whole ICT cluster it belongs to.

Altogether, this conclusion raises a valid point of caution and can contribute to the discussion on whether AI is really a GPT (Vannuccini and Prytkova, 2020).[16]

In sum, the whole ICT cluster increases scope and scale of diffusion over four decades with discernible within–cluster differences among the constituting ICT classes.

Technological vs application relatedness. In this part I will discuss (i) obtained results on relatedness of knowledge and application bases among ICT technologies and (ii) analyse the change in industrial mix related to ICT.

Figure 3.5 is an empirical expression of the framework presented in Table 3.4; it plots technological and application relatedness as abscissa and ordinate respectively for each pair of ICT classes and contains two periods 1977–1990 ($t = 1$) and 2005–2020 ($t = 3$) to expose the dynamics of the indicators. Both indicators must be significant in either period for an ICT pair to be displayed in the graph.[17] For 13 ICT classes, there are 78 pairs (excluding combination with itself) however only 22 have indicators that are significant, and, hence, present in Figure 3.5. Most of these pairs are located in the second quadrant, which implies similarity of application base among ICT technologies but specific knowledge base underlying each technology in the cluster. For example, based on Figure 3.5 one can observe that the following pairs of ICT technologies increase significantly the similarity of their application base: Information analysis and Security (pair 7–3), Human Interface and High Speed Computing (pair 9–5), High Speed Computing and Security (pair 5–3), and Imaging and Sound Technology and AI (pair 10–8). As suggested by the framework in Table 3.4, bundling and M&A patterns among firms, products and services embodying one of the listed ICT pair might be more pronounced in the industries related to these ICT pairs. To test the validity of the framework, further research is required to estimate bundling and M&A patterns in the industries related to ICT pairs from different quadrants and investigate their differences.

[16] Proponent literature to this statement is represented by Brynjolfsson et al. (2019), Trajtenberg (2019), Agrawal et al. (2019a)

[17] Appendix 3.6 shows all 78 pairs.

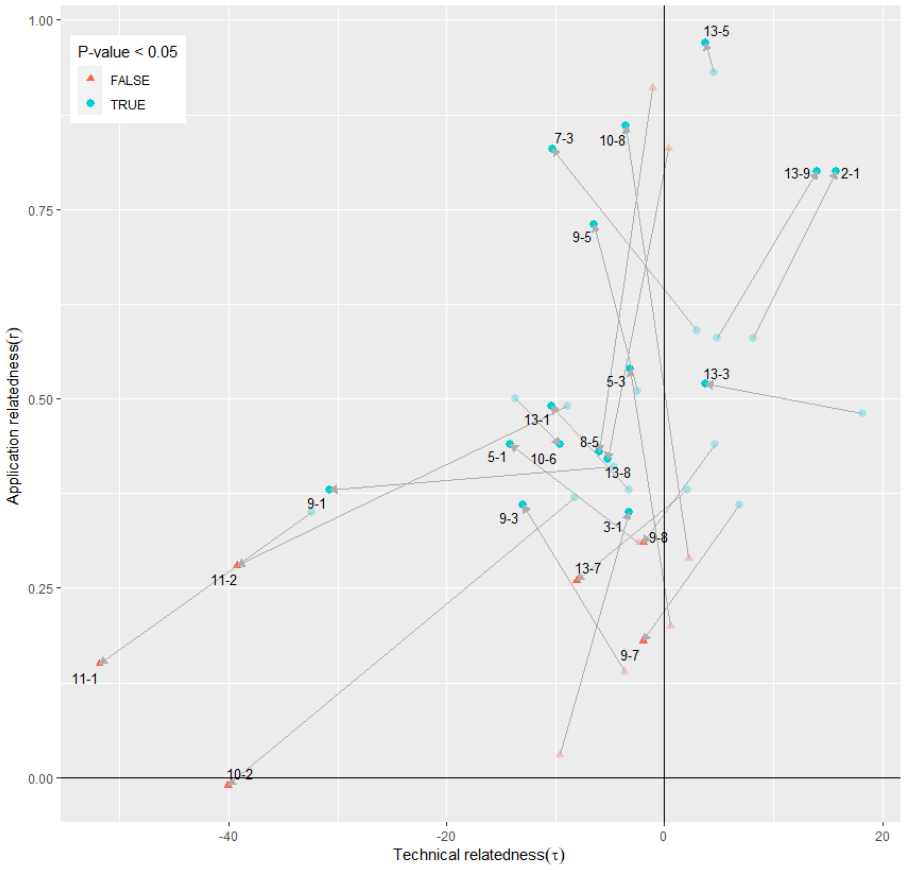


Figure 3.5: Movement of ICT pairs in relatedness space between periods 1 and 3

Note: Two connected observations represent the same ICT pair in the first and the third period. Arrows originate in $t = 1$ position of the ICT pair going to its $t = 3$ position.

One exception from the overall picture is the pair formed by class 1 High speed network and class 2 Mobile communication that lies in the first quadrant. This means the pair not only exhibits similarity with respect to both relatedness indicators but strengthens its similarities over time. Overall, only 4 out of the 22 represented pairs increase their technological relatedness between the first and third period, while 12 out of 22 pairs experience growth of application relatedness. This indicates that already in the first period ICT technologies shared the application base and that tends to grow. As for the knowledge base, it remains delineated along the ICT taxonomy classes. Perhaps, probing deeper levels of ICT classification, might reveal further proximity of knowledge base among more fine–grained, smaller ICT classes.

Completing the analytical exercise, in the next paragraphs I will focus on the AI ICT class to provide insights in its relatedness with other ICT classes and the change in application base AI has experienced. In addition to the estimations of pervasiveness, the position in the relatedness space and the structure of industry reliance on AI contributes to a better understanding of AI's development and to the construction of the broader context in which AI is considered.

AI and High speed computing. As can be seen in Figure 3.5, the two technologies (the pair of classes 8 and 5) exhibit a decrease in both indicators of relatedness. The pair remains in the second quadrant indicating similarity of the industrial base but distinct knowledge behind each technology. Between the two periods the relatedness of knowledge base decreases marginally, and the pair moves only slightly to the left. In contrast, application relatedness has clear dynamics plummeting from a strong to moderate level. The next paragraphs provide insights into this trend.

At the inception in the 1950s, AI was inseparable from *contemporary* powerful (high speed) computing machines. It was not yet a commercial technology but a scientific experiment, an attempt to simulate the higher functions of the human brain such as speaking and understanding language, sensory perception, reasoning, self–improvement or learning, forming abstractions, creative thinking, etc (McCarthy et al., 1955).[18] This implies that in the past AI instantiations could exist if attached to an actor that has sufficiently powerful computing capacity and labor with programming skills. Within the 1977–1990 period, both resources were in their gestation and growth phases so initially scarce but growing rapidly. According to Beckhusen (2016), at the beginning of the 1970s in the US, "computers were large, expensive mainframes mostly used by governments, research laboratories, and manufacturing firms" while in 1990 already penetrated households and businesses. This also fueled the demand for IT workers whose number grew from 0.45 to 1.5 million between 1970 and 1990. These numbers include all types of IT workers with computer programmers alongside database administrators

[18] Mohamed et al. (2020) summarize the evolution of AI from its onset until now claiming that "AI has seen itself elevated from an obscure domain of computer science into technological artefacts embedded within and scrutinised by governments, industry and civil society".

and computer network architects, hence only a fraction of which could be relevant to AI. For example, the enrollment in the introductory AI and Machine Learning courses in several US universities since 1990 is in the range of 200–300 up until approximately 2010 when the growth took off (see Appendix 3.7). As suggested by Timmer and Van Ark (2005), investments in ICT capital and production of ICT goods in the EU lagged behind the US, therefore a similar dynamics could unfold in the EU with the whole timeline shifted in time.

In fact, the relation between AI production and the possession of resources for that was the opposite: at the pre–commercial phase of AI the possession of computing capacity and programming skills didn't stem from the incentive to develop AI but eventually allowed doing so, as opposed to the current substantially commercial phase when the goal of AI development and/or usage consequently drives the decision to acquire the necessary resources. With the invention and gradual growth of commercial AI forms, the markets for hardware and labor with programming skills started experiencing a positive shock. This triggered an ameliorative loop among markets for AI–related labor, hardware and AI solutions (software). The form of provision of labor and hardware resources to developers and adopters of AI solutions is one of these two: production in–house or purchase (make–or–buy dilemma). In particular, over the decades computing power became not only more affordable (Flamm, 2019) but also more accessible facilitated by cloud computing (Byrne et al., 2018). The growing quality and availability of both AI solutions and AI–related resources facilitates AI adoption in industries where the commercial value is either not obvious (experimentation) or requires additional adjustments (implementation lags) of production processes, capital, business models, etc. Thus, those novel industries that adopt AI are likely not the ones that relied before on high speed computing for their tasks and functions. Moreover, high speed computing is bound to AI in industries that build functionality and maintain AI systems in–house and have not purchased access to it from AI provider with subsequent servicing.

Looking at the empirical results, the most recent industries that appeared in Table 3.6 such as pharmaceutical (e.g. drug discovery, medical imaging), scientific R&D (AI as Invention of Method of Invention (IMI) (Cockburn

et al., 2019)), employment (e.g. HR algorithms), security and investigation (e.g. predictive policing, suspect identification with visual recognition) are rather industries buying AI solutions for their data processing. As noted earlier, the first production of AI systems emerged atop of the possession of computing and programming labor resources hence producers of AI must be the long–standing holders of both resources. In terms of the current analysis, potential AI–producing industries are likely to be (i) listed among industries related to both AI and High speed computing since the first period (incumbent industries) and (ii) ranked high in terms of strength of connection with both ICT classes (by the sum of ranks over three periods).

	ISIC rev. 4	Class 5: HSC	Class 8: AI
59	Motion picture, video, sound recording, music publishing activities	-	1
62	Computer programming, consultancy and related activities	1	2
63	Information service activities	2	3
26	Manufacturing of computer, electronic and optical products	3	-

Table 3.5: Top–3 positions in the intertemporal ranking of industries connected to AI and HSC

Not surprisingly, two industries that fulfill both criteria are 62–Computer programming, consultancy and related activities and 63–Information service activities. As shown in Table 3.5, these two incumbent industries occupy the second and third positions in the intertemporal ranking for the strength of connection with the AI class and first and second for the High speed computing. The production process of digital products and services in these activities involves precisely the two mentioned production factors — labor with programming skills and computing machinery as capital. An interesting finding is which industry completes the top–3 ranking for each of the ICT classes. Concerning AI, the first position in the ranking is held by the endemic to AI industry 59–Motion picture, video, sound recording, music publishing activities. As for High speed computing, the third position in the ranking belongs to industry 26–Manufacturing of computer, electronic and optical products. Altogether, this breaks down AI technology into its basic components: perception through sensory data, information–processing algorithms, and computing machinery, echoing the composition suggested by Taddy (2019).

In sum, the number of AI–related industries grew apparently due to the increase in the number of AI–using industries. These industries adopt AI

solutions developed and managed by an AI–provider who bundles the provision of AI algorithms with computing power required to support the application. The joint presence of AI and High Speed Computing becomes localized and limited to a small number of AI–producing industries (i.e. 62 and 63), while AI–using industries show connection to AI only. This results in a decreasing application relatedness between class 8 AI and class 5 High speed computing.

1977-1990 t=1	1991-2004 t=2	2005-2020 t=3	ISIC rev.4 Code	Description
			18	Printing and reproduction of recorded media
			21	Manufacture of basic pharmaceutical products and pharmaceutical preparations
			26	Manufacture of computer, electronic and optical products
			32	Other manufacturing
			51	Air transport
			58	Publishing activities
			59	Motion picture, video, sound recording, music publishing activities
			61	Telecommunications
			62	Computer programming, consultancy and related activities
			63	Information service activities
			64	Financial service activities, except insurance and pension funding
			65	Insurance, reinsurance and pension funding, except compulsory social security
			66	Activities auxiliary to financial service and insurance activities
			69	Legal and accounting activities
			71	Architectural and engineering activities; technical testing and analysis
			72	Scientific research and development
			73	Advertising and market research
			74	Other professional, scientific and technical activities
			78	Employment activities
			80	Security and investigation activities
			82	Office administrative, office support and other business support activities
			86	Human health activities
			87	Residential care activities
			90	Creative, arts and entertainment activities
			91	Libraries, archives, museums and other cultural activities

Table 3.6: Dynamics of industry reliance on AI

AI and Imaging and sound technology. The pair (class 8 and class 10 respectively) migrates from the first to the second quadrant by decreasing its technological relatedness. Similar to the previous pair of AI with High speed computing, this pair exhibits a significant change in the application relatedness but in this case it surges to a strong level.

As it has been pointed out in the previous paragraphs, one of the onset goals of AI is to replicate the complex function of sensory perception in different modalities (visual, audio, tactile). Understanding the algorithms behind the processing of *unstructured*, raw sensory data that results in structured information (pattern recognition) has been a challenge due to algorithms' non–deterministic nature; only after studying the same processes in living organisms the idea of artificial neurons and their "assembly formation", i.e

neural networks capable of calculation have been formulated.[19] Even after proof of concept in the late 1950s and 1960s, it took almost five decades to overcome initial limitations and find the way to scale up a simple perceptron to a modern version of Artificial Neural Networks (ANN) that surpass human performance in a set of tasks (see Eckersley et al. (2017)). It turned out that pattern recognition capabilities of ANNs that could harness raw sensory data without preprocessing can also work as well with structured, e.g. trading data, demographic data, medical records, and fuzzy–structured, e.g. language, clickstreams, data. Thus, starting in the domain of pure logic and deterministic rules with the symbolic approach, eventually, AI has evolved eventually into a non–deterministic, highly perceptive instantiation, representing the connectionist approach. The ability to transform vast amounts of structured and especially unstructured data into information let connectionist AI in many different industries where such information (in a form of inference and/or prediction) has a value (Agrawal et al., 2019c). Thus, Image and sound technology became an ultimate tool of data collection for further processing by AI systems, which is reflected in the surging growth of application relatedness between these two ICT classes.

3.5 Conclusion

The importance of ICT for the functioning of any economic system cannot be underestimated. However, studies on the impact of ICT often considered this cluster as a monolith block. In this Chapter, I distinguish a set of ICT technologies employing the new ICT taxonomy from OECD and PATENTSCOPE AI Index and estimate ICTs' connections with industries using the Algorithmic Links with Probabilities method. The construction of a fine–grained industry–technology map allows assessing the structure and evolution of industry reliance on ICT. This required the application of several text analysis techniques to break down industrial descriptions into sets of keywords to match them with patents' abstracts and titles. The subsequent application of network analysis and relatedness indicators helps to uncover patterns and regularities in the structure and dynamics of the

[19] Ground work by McCulloch and Pitts (1943) and Hebb (2005)

constructed ICT technology–industry network.

The results indicate that the ICT cluster shows signs of a "phase transition", passing the phase of building bulk elements of the infrastructure and around the 2000s entering the phase of working on the functionality for business applications deployment and users' convenience. More application–oriented technologies like mobile communication, information analysis, security and human interface show significant and persistent growth of their shares in the ICT cluster in the EU28 region. In contrast, more mature technologies that represent major physical components of the infrastructure such as high speed network and information communication device recede though keeping a strong presence in the cluster. The inclusion of the industries into consideration allows looking into the structure of connections between ICT technologies and industries and its dynamics. Despite being the largest in the ICT system, high speed network, information communication device and imaging and sound technology are not the most central and connected ones. Instead, information analysis is forging far ahead with regard to both scale and scope, penetrating an increasing number of industries and strengthening its industrial connections. Said differently, information analysis moves rapidly towards the center of the knowledge base of the ICT–related industries compared to other ICT technologies in the cluster. Security technology occupies the second position after information analysis by overcoming the largest ICT classes in scope and getting to parity in scale. Overall, the biggest ICT classes have a moderate number of incident industries but with stronger connections; by contrast, many smaller classes (except for information analysis and security) have numerous but weaker connections.

According to the framework represented in Table 3.4, the position of the overwhelming majority of ICT pairs the second quadrant indicates shared industrial base though distinct knowledge underlying each ICT technology. Some pairs strengthen their position in the second quadrant such as: Information analysis and Security, Human Interface and High Speed Computing, High Speed Computing and Security, and Imaging and Sound Technology and AI. Following the suggested framework in Table 3.4, for the pairs in the second quadrant the visibility of commercial value might be a factor that navigates the innovation process. Pursuit of commercial value might

lead to more pronounced and systematic bundling of products and services and/or M&A among firms producing goods embodying these ICT technologies than for ICT located elsewhere in the relatedness space. Overall, among ICT pairs, application relatedness tends to increase over time (12 pairs out of 22) while technological relatedness appears rather stable (4 pairs out of 22).

A special focus of the analysis is placed on AI technology among the ICT cluster. On the one hand, AI is a novel, fast–growing technology that enters the commercial phase and is subject to intensive development and experimentation. Multiple applications, unprecedented potential for automation and billions in generated revenues make AI a fruitful topic to study. On the other hand, the research on AI might benefit from putting this technology into context, and study AI in relation and in comparison with other technologies. In this Chapter, ICT classes serve simultaneously as potential complementors and as benchmark, building a framework for AI's evaluation. An interesting finding concerning AI is that it shows the largest increase in scope yet negligible growth in scale and its small absolute magnitude. This points at multiple but yet "shallow" connections between AI and industries going in line with AI's gestation and early growth phase. In the technical literature, it is reported that both AI algorithms and hardware represent ad hoc solutions that lack flexibility (Sze et al., 2020; Hooker, 2020), and producers are only at the beginning of addressing this issue. In particular, the identified connection between High Speed Computing and AI technologies shows decreasing but significant application relatedness perhaps due to the acquisition of AI–using industries that largely rely on computing power in the cloud or adopt pretrained AI models and themselves do not employ powerful hardware. This reflects the actual dilemma between fragmentation into specialized hardware and integration of broad functionality under a platform chip that the semiconductor industry is currently facing (Prytkova and Vannuccini, 2020). If the semiconductor industry will decide in favor of specialized hardware (appealing to AI producers) the trend of purchasing access to AI solutions run by AI producers among industries is likely to continue; the application relatedness between High Speed Computing and AI might experience a further decrease. This would also mean the completion of AI upstreaming, with AI becoming a fully–fledged, distinct industry. In

addition to the described challenges to be resolved at the supply side, the reorganization of production and business models on the demand side and adoption lags are likely to delay AI deployment at scale as well.

A further disaggregation of the ICT cluster into more fine–grained technologies might improve the precision of estimation of technological relatedness among the technologies in the cluster. The framework in Table 3.4 has to be tested, hence further research is required to estimate bundling and M&A patterns in the industries related to ICT pairs from different quadrants and investigate their differences. An investigation into the identified connections between ICT technologies akin to the ones involving AI discussed in this Chapter can be not only an interesting exercise for historians of technology but can also inform economists studying technological diffusion, system products and network externalities, and policy–makers in identifying related markets and technologies even across industrial boundaries. As for the estimation of the industry–technology connections, the inherent shortcoming lies in the organization of industrial classification based on the predominant activity; hence in the description the related activities can be underrepresented. However, related activities can be picked up at least partially if their keywords are mentioned together with keywords of the predominant activities in a patent document. Thus, the keyword–based matching partially tackles the problem of false positives by placing no restriction on the matching. The same advantage might turn into a disadvantage when it creates non–existing connections (false negatives). The bigrams and keywords refinement (expansion or/end replacement) with synonyms has been used to reduce the number of false negatives.

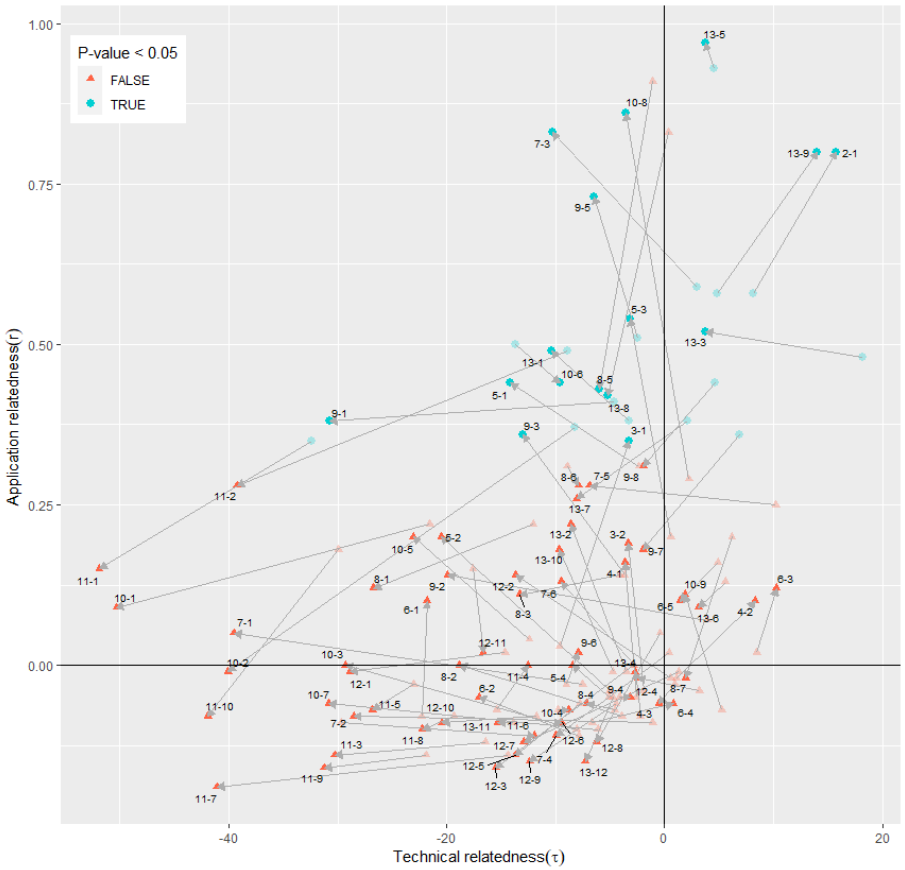
Appendix

Period		1977-1990	1991-2004	2005-2020
			Degree	
ICT	n	13	13	13
	mean	8.62	13.92	16.38
	median	9.00	14.00	20.00
	min	4	1	3
	max	14	22	25
ISIC	m	26	32	36
	mean	4.31	5.66	5.92
	median	3.00	6.50	7.50
	min	1	1	1
	max	13	12	12
			Density	
Bipartite		0.33	0.44	0.46

Table 3.7: Descriptive statistics of industry–ICT bipartite network

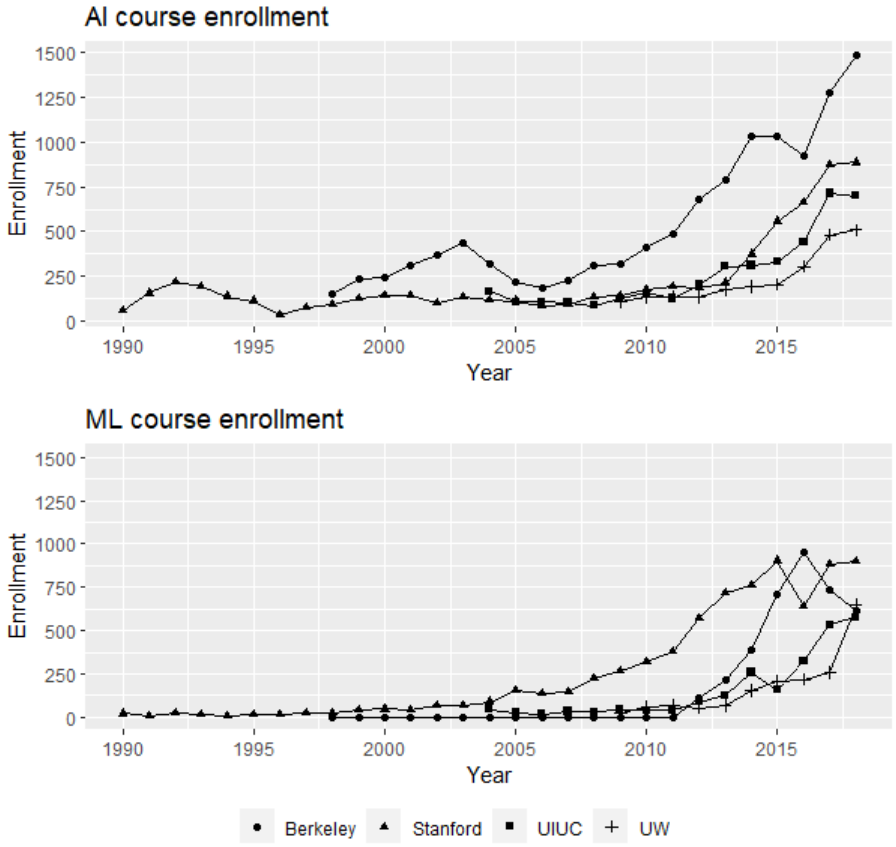
Period	**t=1**		**t=2**		**t=3**	
ICT class	**Degree**	**FBC**	**Degree**	**FBC**	**Degree**	**FBC**
cl1	7	33.71	14	70.88	16	66.78
cl2	4	3.27	8	20.19	9	33.60
cl3	14	22.48	21	44.70	25	55.46
cl4	6	4.82	9	2.41	10	3.50
cl5	6	12.92	14	20.54	20	22.95
cl6	13	37.70	16	27.98	20	17.04
cl7	9	20.82	20	76.17	24	109.37
cl8	9	1.99	20	5.82	24	4.84
cl9	13	24.89	22	30.18	24	36.84
cl10	9	53.07	13	61.24	12	70.15
cl11	8	33.58	6	17.64	6	16.82
cl12	4	7.61	1	0.00	3	14.49
cl13	10	4.30	17	7.44	20	6.96

Table 3.8: Flow betweenness centrality and degree indicators



Note: Arrows originate from $t = 1$ going to $t = 3$ position

Figure 3.6: Technological and application relatedness



Note: Reproduced from the AI Index report 2019 (Perrault et al., 2019)

Figure 3.7: Enrollment in AI and ML courses, US universities

Chapter 4

Artificial Intelligence's New Clothes? From General Purpose Technology to Large Technical System

4.1 Introduction

Technological breakthroughs always acquire a life in their own right, as they beget novel perspectives or anticipate possible futures; doing so, they reshape our expectations and knowledge about potential states of the world. Given that, it is not surprising the increasing attention that social scientists have devoted to recent breakthroughs in Artificial Intelligence (AI). AI technologies are not an absolute novelty; rather, they experienced cyclical phases of hype and oblivion, often with scientific advances arriving with uneven time intervals in between, sometimes distant enough to create an impression of stagnation. The recent focus on AI grows out of the idea that 'this time is different' not only with respect to the previous phases of AI itself, but also in comparison with other technologies that are considered potential candidates to be at the core of technological revolutions. Regard-

less whether this time is really different, the idea that AI is a revolution is producing real–world effects. In fact, likely induced by the warmth of the current AI Spring that is reviving the curiosity and concerns frozen during AI Winters, many studies now focus on the technological features and socio–economic implications of AI.

The premise of human–level performance in many tasks creates expectations of AI's pervasive diffusion and, as a corollary, a key idea has been advanced: that AI is a so–called General Purpose Technology (GPT) (Goldfarb et al., 2019). Along with the literature on the Economics of AI, we think that AI should be assigned a special status among technologies. However, while acknowledging the transformative potential of AI, one can agree that tough every GPT is an influential technology, but not every influential technology is a GPT. So, does AI lie within or beyond the GPT definition? In the Chapter, we try to answer this question by mapping modern AI technology onto micro–characteristics and macro–effects assumed by the GPT framework. The result of this exercise suggests that despite AI having some touchpoints with a GPT, such as technological dynamism and innovation complementarities, equating AI and GPT is currently premature and, eventually, is likely to turn out as an incorrect definition of AI. The primary reason for this is that AI is qualitatively different from a stand–alone technology such as a GPT and instead resembles a system or infrastructural technology, approximating a *Large Technical System* (LTS) *in the making*.

LTS is one of the approaches to the study of technology that stresses the systemic — and thus structural and relational — properties of certain technologies. In fact, the perspective offered by LTS well suits the task of mapping system technologies that display wide reach.[1] The value of LTS theory for the Economics of AI lies in providing a 'thick description' of the technology: the vocabulary of notions it introduces is useful to unpack and understand the complementarities and tensions characterising infrastructural technologies and their development. This allows looking at AI from a novel angle offering a more insightful assessment of AI value, advancements and downsides, risks and benefits, room and tools for governance, opportunities and

[1] An example of this approach at work for the case of telecommunication technologies is offered in Davies (1996).

realistic trajectories of AI development.

Following the logic of the GPT part of the analysis, we introduce the LTS framework and map AI onto its building blocks, with a subsequent discussion of its goodness of fit. It should not be surprising that the LTS and GPT frameworks share or mirror some characteristics of each other, as both describe influential technologies, hence some similarity only speaks for the relevance of the comparison. As the differences in describing AI between the frameworks are of bigger interest, while evaluating AI as LTS we provide recollections of AI as a GPT for comparison and conclusion. As AI has just exited scientific laboratories and broke into the wild of commercial markets, it is yet in the making, and the mature form it will take is still to come. At such a key moment of AI development there are strong winds blowing in the direction of AI–infrastructure akin to the Internet, such as high returns on AI–based system–level substitution, concentrated market power among AI–producers, high costs of setting the system, etc. Neglecting these forces and treating AI in isolation from the rest of the system might lead to misplaced investments and dead–weight losses. Thus, the results of our analysis can be useful to researchers in the field of Economics of technological change and innovation as well as to policy makers, which might take–home from this study a better–suited, overarching framework to deal with AI.

The Chapter proceeds as follows. Section 4.2 places the first brick of the edifice by assessing whether it is correct to label AI as a GPT. Section 4.3 identifies which features of current AI map onto the LTS concepts. Section 4.4 derives implications for policy and strategy. Section 4.5 concludes.

4.2 Artificial Intelligence is a General Purpose Technology. Is it, really?

4.2.1 The 'next big thing': Artificial Intelligence

Scholars already consider AI the latest GPT. However, the search for a new GPT is not a novel endeavour. Over time, the title of general–purpose has

been assigned to a non–negligible set of both narrow and broad technologies: electric dynamos (David and Wright, 2003), ICTs (Steinmueller, 2007; Strohmaier and Rainer, 2016), different computer platforms (Bresnahan and Yin, 2010), control technologies (Thoma, 2009), the Boyer–Cohen recombinant DNA technique (Feldman and Yoon, 2012), carbon nanotubes and nanotechnologies (Kreuchauff and Teichert, 2014; Graham and Iacopetta, 2014), and additive manufacturing techniques based on 3D printing (Choi, 2018). The rationale behind this race for GPT identification is one, common to all the studies: to find the 'next big thing' (Trajtenberg, 2018) that can induce profound socio–economic transformations while generating sizeable economic returns (Strohmaier et al., 2019). Going further back in the past, the search for the 'hopeful monsters' or macroinventions (Mokyr, 1990) that start new industries and pervasively transform the economy is a key feature of the literature on long waves, technological revolutions, and techno–economic paradigms (Perez, 2010). What these different approaches to technological change share is the idea that a certain technology might be the *primum movens* of long–term growth, development, and fluctuations (Kurz et al., 2018).

GPT and its diffusion is an ideal–type of a particular kind of technological change: one leading to the adoption of a radical innovation, used as a core component, across very diverse economic activities — in the limit, to ubiquitous adoption. Thus, to understand whether AI is a GPT is important, as that would provide insights into the future developments of the technology, the types of impact we can expect from it, and guidance for the design of policies to govern it.

A short overview of AI. As a first step of the analysis, we outline the framework through which we consider AI. This is necessary as AI is a 'suitcase word' (Mitchell, 2019) that densely packs an array of different meanings and interpretations. Mohamed et al. (2020) stress the dual nature of AI as *object* and *subject*: as object, AI is a set of technological artefacts; as subject, it is a 'portmanteau' of networks and institutions. Our analysis builds on this dual nature. In this section, we deal with AI as object and proceed through progressive approximations: from the philosophy of the technology to its particular instantiations. It is a useful exercise because the domain

is dynamic and especially at the moment, when a handful of actors have entered the field with new products and new visions. In further sections, we move to AI as subject, building up the AI LTS from the core to its outskirts.

Philosophy. AI, being a technological mirror of ourselves, is inevitably compared to natural intelligence. The seemingly philosophical question of whether or not AI possesses a 'true' intelligence has very tangible technological implications in terms of, for example, engineering and programming. *Cognition and meaning understanding*, just to name two, are in fact the criteria and fields of ongoing research (see the new ICT taxonomy by Inaba and Squicciarini (2017)) that separate the so–called *weak AI* from *strong AI*. The distinction is based on the fact that the former only emulates intelligent behaviour, while the latter aims at re–creating it. While the emulation of intelligence is achieved using either rules of logic, heuristics, statistical learning techniques, or combinations of them, the question of how to re–create intelligence to reach true understanding by algorithms remains yet unanswered. Hence, the *current state of AI* belongs to the weak type. A relevant practical issue is that weak AI's reliance on statistical learning techniques entails risks for the deployment and usage of AITs. Incapable of general understanding, weak AI systems "have proven to be data hungry, shallow, brittle, and limited in their ability to generalize" (Marcus, 2020). Furthermore, neural architectures obtained through training can get obsolete, or can perpetuate biases that exist in the society (the 'garbage in, garbage out' principle). Such systems are vulnerable to (adversarial) attacks aimed at distorting or 'polluting' statistical (co–)occurrences in the data, teaching the system to behave oddly.[2] Moreover, several contingencies of the world with no clear (incomplete) ranking or dominance among alternatives remain challenging for weak AI to deal with (see for example the Moral Machine experiment (Awad et al., 2018)). Tweaking the algorithms in order to avoid these problems — correct for biases of data or society, create a decision–making routine for situations with no dominant strategy — and then retraining them entails high costs in terms of time, programming effort, computing power and energy, new tests, and environmental toll (Strubell et al., 2019)). In sum, current AI belongs to the weak AI domain

[2] A famous example illustrating such case is Microsoft's chatbot Tay: https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/.

and it has direct and tangible technical and societal implications with respect to which uses can be made of the technology and which risks it entails, and how to regulate the industry emerging around it.

Approach. We already mentioned statistical learning (including all: supervised, unsupervised or reinforcement) as a method to emulate intelligence. In general, there are two main approaches to AI: rule–based or symbolic approach (or good–old–fashioned AI, GOFAI) and statistical (data–driven) or connectionism.[3] In the symbolic approach an algorithm's search for a solution is driven by formal logic and explicit rules to deal with a given task, while connectionism uses statistical learning to infer implicit regularities from (un)structured data in order to perform a task. Currently, the latter is the prevailing approach to AI; it earned its fame due to the capability to learn from raw data without any predefined rules. This makes the connectionist approach autonomous, more flexible and effective in pattern recognition when compared to more rigid and bulky symbolic AI systems. However, connectionist AI algorithms, such as Artificial Neural Networks (ANNs), are tied to the task they perform (for instance visual recognition, language processing, or games — with only rare and partial exceptions, such as DeepMind's line of algorithms Alpha), so *they are function–dedicated virtual machines* (Boden, 2016).

Technological constituents. Regardless of the methodological approach, any AI technology necessarily consists of the following domains: (i) *algorithms or virtual machines*, (ii) *computing power* (and related physical devices delivering it), (iii) *data* and (iv) *domain structure*.[4] Domain structure indicates the problem environment and search space of actions an AI system is working with. Current AI technology applies an algorithm capable of learning to data in a given domain structure, requiring a certain amount of computation (which might vary substantially depending on the size of data and complexity of the algorithm and learning technique). In the context of this Chapter, listing these domains already makes clear that the set of actors involved in building AI systems should be extended from only

[3]Streams of research in AI are also partly moving towards hybrid approaches (Domingos and Lowd, 2019). We called our focus the 'current' understanding of AI as this is the one dominating research and societal attention at the moment.

[4]This description builds on Taddy (2019), but introduces some variations.

algorithm–producing actors to a broader set including at least hardware producers, data providers, regulators, vendors, and buyers. Each domain is characterised by its own market structure, business models, regulations and pace of development. For instance, the semiconductor industry is a well–defined, consolidated manufacturing industry with a set of big players and high entry barriers, fuelled by capacity races and economies of scale (Steinmueller, 1992) and, from the technological side (until recently), by the Dennard scaling principle.[5] Differently, data is increasingly being considered as a commodity in recent years and debates on privacy and how to treat data are still ongoing (Savona, 2019). Despite the growing availability of data, compared to other domains, data markets remain opaque (Koutroumpis et al., 2020a).

Functions and applications. In order to complete the picture drawn in previous paragraphs, here we name some instantiation of current AI systems.[6] AI must not be conflated with either automation or robotisation, as AI plays only a specific role in each of these processes. Concerning automation, AI represents its high–end but is not a necessary condition for automation to take place; hence, AI is a subset of technologies associated with automation processes. With robotisation AI is rather in a parity relationship as robots, being mostly 'hardware', may or may not embody AI as a control method for physical work.

A last close up look at AI brings us to its practical functions and actual applications. Different functions like perception (visual and speech recognition), predictive analysis, communication (machine translation, information extraction) or control (robotics, facility–managing systems) might be involved separately or in combination in a variety of industries, from agriculture to advertising, Fintech, and even satellite communication (Vázquez et al., 2020). There are pioneer industries that deployed AI due to either the presence of specific functions that AI was capable of performing effectively or because the whole industry could come to existence thanks to AI; according to WIPO (2019b), based on patent data, the top AI user–industries are

[5] The well–known Moore's law is a resulting expression of Dennard scaling as a physical principle.

[6] A good summary of AI techniques, functions and their applications fields is provided in WIPO Technology Trends Report 2019 on AI (WIPO, 2019b).

transportation, telecommunication and life and medical sciences.

Taking stock, in this Chapter AI as object belongs to the *weak and connectionist AI*. In the next sections, we proceed by building around this core the economic, societal and institutional framework that constitutes AI as subject, and check this new systemic view of AI against the concepts of GPT and LTS.

4.2.2 Assessing the GPT nature of AI

General Purpose Technologies. Usually, to label a certain technology a GPT, characteristics of such technology are checked against the definitional criteria that describe what a GPT is. The most used definition is that of Bresnahan and Trajtenberg (1995), according to which a GPT is a technology that displays (i) *general applicability*, (ii) *technological dynamism*, and (iii) *innovational complementarities*.[7] General applicability captures the pervasiveness feature of GPTs. A GPT can be used as input or core component by a wide array of downstream industries or economic activities. Technological dynamism suggests that a GPT should display a steep learning curve in performance and/or efficiency — namely a fast ongoing improvement pulled by the enlarging downstream expenditure in the technology. Innovational complementarities, or innovation spawning, is instead the property that characterises GPTs from the perspective of innovative activities, rather than diffusion: GPTs are considered enabling mechanisms, as they induce or reinforce innovation incentives in the industries that use them as an input. Taken together, these characteristics illustrate a peculiar mechanism revolving around linked incentives: GPT producer and user sectors play a coordination game in which their optimal choices on technology are supermodular — the technological level of user sectors depends on GPT producers' product quality, and vice versa. A key feature highlighted by the

[7] The research programme on GPTs is characterised by a lack of coordination between different approaches, which has led to unresolved controversies (Bekar et al., 2018; Cantner and Vannuccini, 2012). For example, Lehrer et al. (2016) distinguish between nested 'mega GPTs' and 'anchor technologies', where the former are broad technological areas (such as ICT or nanotechnologies) and the latter are identifiable technologies (like semiconductor chips, or ERP software). From this descends the difficulty to draw a consensus definition of this family of technologies.

GPT literature is that the impact of such supermodularity can have either a positive or a negative sign: due to the externalities coded into the GPT coordination game, both virtuous and vicious reinforcement cycles can occur, leading GPT development to feature multiple equilibria.

In what follows, we focus on the three core definitional characteristics of a GPT — general applicability, innovational complementarities, and technological dynamism, — together with two additional features, uniqueness and implementation lags, and assess whether they fit to describe the current AITs outlined in the previous subsection.

General applicability of AI as a GPT. The most characteristic feature of a GPT is its generality, or *pervasiveness*. A GPT becomes embodied in most of the technologies and used at scale within the majority economic activities. This feature supports the claim that electrification or digitalisation are GPT–related processes: every device or product can be powered by forms of stored or non–stored electrical power, and most of the functions or activities conducted in an analog–mechanical manner (executed by relying on continuous input such as for example force or heat) can either be transitioned to digital (analog signals are replaced by discrete series of bits) or controlled digitally. However, the concept of pervasiveness carries a fundamental ambiguity, highlighted by Bekar et al. (2018) when they distinguish between *many uses* made of a particular technology, and technologies that are *widely used* across the economy. A technology with many uses is general purpose in nature, but that does not imply that it is also adopted at scale in the majority of economic activities; hence, the overall proportion of the economy that uses this technology might be small. In contrast, a single purpose technology can be an essential component in one or few industries. A GPT, in order to be pervasive, should permeate the economy in scale and scope — being widely used (at scale) in many uses (in scope).

It is undeniable that AITs are increasingly used in a disparate set of economic activities. What is remarkable of AITs is that they create *ex novo* activities in which they can be deployed — they kick–start new sectors and enable new products, e.g. autonomous vehicles. However, apart from some ex novo activity, one might argue that the majority of economic activities

has only a limited reliance on AI. The fact that AITs' implementation at scale is localised in a few economic activities can be measured with respect to the following dimensions: penetration of (i) production processes, (ii) tasks within occupations, and (ii) overall adoption at the industry and firm level.

Looking at production processes, AI executes tasks that were already executed by capital, in particular ICT capital. The adoption of AI occurs through a replacement of existing software technologies with more sophisticated ones, those based on AI algorithms. This implies that AITs do not induce a substitution between production factors (capital for labor), and therefore the scale of task replacement is limited. Indeed, Bresnahan (2019a) suggests that AITs generate *system–level substitution*. System–level substitution occurs between production systems — for example online retail replaces brick–and–mortar one, automated user support or algorithmic fraud check replace the computer–aided but human–controlled version. Therefore, this process has to do with the introduction of new, more capital intensive 'production technology'; this includes the infrastructure underlying a firm's activities as well as its business model. In fact, AI–driven system–level substitution occurs in production processes that are already capital intensive pre–AI: these are a narrow set of economic activities or functions, oriented to consumer applications. Limited system–level substitution for AI contrasts the diffusion path at scale of the first wave of ICT (computers), and resembles the adoption dynamics of more recent ICT technologies, such as web and mobile applications: a targeted process of capital deepening in some activities (e.g. recommendation engines) leading to wide use by end–users and high returns but that leaves the rest of the economy substantially untouched.

Focusing on the occupation level, AI fails to permeate jobs at scale. Table 4.1 displays the share of AI–related jobs in the total jobs posted in the US for the year 2019, by NAICS 2–digit sector.[8] AI jobs are posted across a wide array of sectors (from information to construction), indicating a spanning applicability of the workers' skills that are complementary to AITs. However, AI is not widely used: the share in the top–posting sector does

[8]The data is retrieved from the 2019 edition of the AI Index Report (Perrault et al., 2019) at the following link: shorturl.at/oAMU9

not exceed 2.4% of total job posting, and is limited to values below 0.5% for half of the sectors considered. In line with this evidence, Acemoglu et al. (2020), find that while AI–related job postings accelerate, there is "no discernible impact of AI exposure on employment or wages at the occupation or industry level, implying that AI is currently substituting for humans in a subset of tasks but it is not yet having detectable aggregate labor market consequences". Exposure to AI affects some specific tasks within jobs, but not the occupational structure.

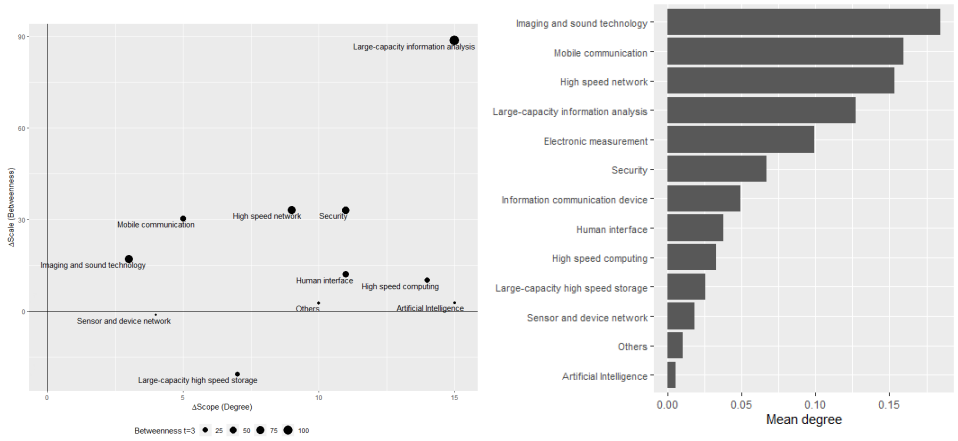
Industry	**Share of AI jobs, %**
Information	2.4
Professional, Scientific, and Technical Services	2.1
Finance and Insurance	1.3
Administrative and Support and Waste Management and Remediation Services	1.1
Manufacturing	1.1
Management of Companies and Enterprises	0.7
Mining, Quarrying, and Oil and Gas Extraction	0.6
Agriculture, Forestry, Fishing and Hunting	0.6
Wholesale Trade	0.5
Educational Services	0.5
Public Administration	0.5
Retail Trade	0.4
Utilities	0.4
Health Care and Social Assistance	0.2
Real Estate and Rental and Leasing	0.2
Transportation and Warehousing	0.2
Other Services (except Public Administration)	0.2
Arts, Entertainment, and Recreation	0.1
Accommodation and Food Services	0.1
Construction	0.1

Table 4.1: Share of AI jobs posted (out of the total) by Industry, United States, 2019

Source: Perrault et al. (2019)

For what concerns industrial connections, there are pieces of evidence that AI's diffusion among industries has a peculiar structure: despite being linked with many industries, these connections are shallow in the majority of cases. Using the expression of Bekar et al. (2018), AI has many uses, but is not widely used. For example, Prytkova (2021) considers the whole ICT system and estimates the scale and scope of industrial adoption of each distinct

technology that constitutes the system, including AI. Figure 4.1 combines the results of Prytkova (2021)'s empirical analysis to illustrate industries' shallow reliance on AI. Figure 4.1a plots the change of scope (x–axis), i.e. number of AI's industrial linkages, versus the change in scale (y–axis), i.e. network centrality measure of AI as a technology connecting industries, between two periods — 1977–1990 and 2005–2020; the size of observations is the absolute value of the scale measure for the respective technology in the latest period. The reading of the figure indicates that AI acquired the largest number of industries between the two periods, but it is nowhere near to be adopted at scale. To reinforce the evidence, Figure 4.1b plots the average strength of industrial connections for each technology in the system; compared to other ICT technologies, AI ranks last.



(a) Dynamics of AI's scope and scale

(b) Average strength of industrial connections

Figure 4.1: Industrial connections of AI (Prytkova, 2021)

In line with previous arguments and findings, Bresnahan (2019a) concludes that despite many uses, AI has gained a truly global reach mainly through one single economic activity: online retail and marketing (personalised advertisement and recommendation algorithms); even within this activity, diffusion is concentrated in and driven by a few, large and dominant actors — excluding them would further reduce the scale of AI adoption. AI applications that are not online retail or marketing sum up to a small fraction of the economy or to specialised and narrow user bases.

In terms of technology adoption at the firm level, Table 4.2 shows responses

to the US Census Bureau Annual Business Survey 2018 — Digital Technology module, which provides information related to the year 2017 for all 3–digit NAICS sectors.[9] New business technologies use includes many AITs, such as machine learning, machine vision, natural language processing, and voice recognition (AITs are highlighted in bold). The table reports the share of surveyed firms that test, use or do not use these technologies; the evidence suggests that only a minority of firms report some use of AITs. In line with our argument, here we can interpret the many uses of AI as reported usage of several different AITs within firms, though not at scale — the share of firms testing or using AITs never goes above 3% of the total sample.

- In sum, across different dimensions of analysis — factors use and production technology, occupation structure and tasks, adoption at industry and firm level — AI achieved a limited penetration: a few economic activities rely on it at scale, while with the rest of the economy it developed a shallow connection. This means that AI is not pervasive in a GPT sense. However, can AI *become* pervasive as a GPT? Is the situation we outlined a temporary one, due to the infancy state of commercial AI? AI adoption is a moving frontier: theoretically, as more traditional operations are re–framed as prediction tasks, more activities could potentially be executed by AI. However, the way in which AI gains scale is as an infrastructure, hence a measure like pervasiveness that is developed for stand–alone technological artefacts such as GPTs does not square well with AI producing little insights into the technology.

Innovational complementarities of AI as a GPT. Given its enabling nature, a GPT is expected to positively influence the *rate* of innovation in the GPT–user industries adopting it. The mechanism behind a GPT spawning innovation in downstream sectors is the so–called 'dual inducement' (Bresnahan and Trajtenberg, 1995). A dual inducement would occur when increasing the 'quality' of the GPT raises the curve of innovation returns for user industries; in turn, this raises the returns for the GPT sector to invest

[9] The data can be found here: https://www.census.gov/data/tables/2018/econ/abs/2018-abs-digital-technology-module.html.

	(1) No use	(2) Testing but not using in production or service	(3) In use for less than 5% of production or service	(4) In use for between 5%-25% of production or service	(5) In use for more than 25% of production or service	(6) Don't know	Total share of use (including testing) (2)+(3)+(4)+(5)
Augmented reality	80.0	0.3	0.3	0.2	0.2	19.0	1.0
Automated Guided Vehicles or AGV Systems	81.7	0.2	0.2	0.2	0.3	17.4	0.9
Automated Storage and Retrieval Systems	76.4	0.3	0.8	0.9	2.5	19.0	4.5
Machine Learning	79.3	0.5	0.8	0.7	0.8	17.8	2.8
Machine Vision Software	80.6	0.3	0.5	0.4	0.6	17.6	1.8
Natural Language Processing	81.1	0.3	0.4	0.3	0.4	17.5	1.4
Radio-frequency Identification Inventory System	81.8	0.3	0.3	0.2	0.3	17.1	1.1
Robotics	82.1	0.2	0.4	0.3	0.4	16.6	1.3
Touchscreens/ kiosks for Customer Interface	77.8	0.7	1.3	1.2	2.3	16.6	5.5
Voice Recognition Software	80.8	0.6	1.0	0.6	0.5	16.6	2.7

Table 4.2: Business Technology use in US firms (AITs highlighted)

Source: United States Census Bureau Annual Business Survey — Digital Technology Module 2018 (Table 3A: Business Technologies by 3–Digit NAICS for the United States and States)
Note: reference year 2017; numbers are totals for all sectors; number of firms surveyed: 4,618,795.

in GPT improvements. Dual inducement is typical of one–to–many architectures of technologies and industries resembling the broadcasting principle.

AI is certainly inducing higher rates of innovation: better AI algorithms are enabling more innovation in AI–using sectors, and the achieved positive results feedback on the incentives of AI–producing sectors to invest in further development of AITs. This description resembles a one–to–many (star) network, with pairwise connections between AI on one side and downstream sectors on the other side — as the stylised dual inducement suggests. In reality, for AITs the feedback is a systemic many–to–many process, with

the whole collection of AI 'sibling' domains (hardware, software, data) connected to downstream sectors. AI evolves as a system, with innovation being 'pulled' by different downstream sectors; each sector calls for improvements in one or several AI domains that hinder its development. For instance, design of autonomous vehicles craves equally for more accurate algorithms because of their high stake loss function, faster processing and less energy consuming chips because of cars' battery capacitance, while more static applications like virtual assistants prioritise heterogeneity of computing and scalability. Even within the hardware domain, the established technological trajectory of semiconductors is being de–railed because of misaligned preferences among an increasing number of downstream sectors (Prytkova and Vannuccini, 2020). Another downstream sector of AI, the pharmaceutical and health industries, exert pressure on AI's development in two domains at the same time: algorithms and data. As for algorithms, the industry demands more explainable and at the same time better performing algorithms, that are usually associated with higher complexity and less explainability. As for data, the problem of availability of medical data to train and test algorithms' performance is tied to the debates on data privacy.

The role AI plays in innovation is broader than the one captured by GPTs' innovational complementarity. A GPT is a component that affects passively the innovation incentives of downstream sectors. Instead, AI actively participates in invention and innovation processes by creating information input: it can handle complexity ('needle–in–a–haystack' problems (Agrawal et al., 2018b)) and explore knowledge combinations in an automated manner, lowering search costs. While a GPT sets in motion a mechanism that raises the returns to innovation, AI directly helps innovating. From this perspective, AITs are invention machines (Koutroumpis et al., 2020b), and, thus, are closer to a so–called invention of a method of inventing (IMI; Griliches (1957)) than to a GPT. AI algorithms bruteforce the knowledge space (for example, corpora of annotated medical text) in order to identify potentially valuable associations and guide exploration. This has practical applications in business and in science. In business, AITs can intervene in product design and prototyping. In science, AI is increasingly used to aid the discovery of new drugs, materials, or biological structures such as the folding of proteins (Senior et al., 2020).

Despite the potential direct role in invention and innovation, AI is not displacing labour nor is used at scale even in this context. Bianchini et al. (2020) show that — at least for the Deep Learning technique and the case of health sciences — AITs do not yet work as a discovery 'autopilot' to explore and exploit the knowledge space. Rather, they remain an auxiliary research tool complementing existing scientific structures and practices.

- For AI, innovational complementarities have a networked, many–to–many nature: the inducement of innovation occurs among the (upstream) domains constituting AI as well as with (downstream) application sectors adopting AI. Moreover, AITs play a broader role than GPTs in inventive and innovative activities: rather than just influencing the rate of innovation, they are invention machines that actively participate in the process by automating the search for useful knowledge combinations and, thus, creating novel information input.

Technological dynamism of AI as a GPT. AI seems to display technological dynamism. The performance of AITs compared to different benchmarks is improving quickly, so much to achieve above–human scores in some specific tasks (Eckersley et al., 2017). However, these tasks are yet the same that AI pioneers envisioned in the 1950s and 1960s, namely, tasks belonging to those fields in which human intelligence can be *emulated*, rather than generated *in silico*: pattern recognition, some perception (vision, speech), learning, bruteforcing of search through combinatorial spaces (Minsky, 1961). AI's technological dynamism is overestimated: the current improvement in performance is achieved through joint advances in different domains, as *a joint effort* taking place in the fields of computing, data and AI algorithms. This depends on the many–to–many structure of AI system, as we highlighted for the case of innovational complementarities. Thus, for AI, the feature of technological dynamism should be re–considered into what we call a *systemic technological dynamism*.

The point of systemic technological dynamism is that improvement is not isolated in one domain of AI, but involves and requires changes in many domains. Therefore, it is hard to trace where exactly the change started i.e. which domain 'decided' to improve. For illustrative purpose, we pick

the algorithm domain of AI and unpack the system aspect of its dynamism. However, the reader must keep in mind that a networked system constantly circulates many changes from domain to domain, hence improvements occur as a result of purposeful action as well as 'fall from the sky'. In the algorithm domain, technological dynamism can be expressed through improvements in algorithm design, and is subject to a rather multifaceted dynamics. Algorithms are information goods; therefore, their production and consumption is shaped by incentives that are different compared to physical products (Shapiro et al., 1998). The kind of improvements of algorithm design are strongly driven by all sorts of efficiencies — computation per time, improvement in prediction per data batch, and trade–offs like bias–variance, complexity–explainability and other factors such as the ease of scaling up and costs of replication rather than by classic physical economies of learning (Nagy et al., 2013). Even at this fine–grained level of analysis, the constraints for improvement of algorithm design have a systemic nature: for example, if the possibility to improve the performance of algorithms depends on the programming environment used or (and) the hardware chosen, then algorithms evolution is function of strategic choices of the actors that control the programming environment and hardware production.

- In sum, technological dynamism in AI is not as rapid and linear as expected and occurs in bursts. The *systemic* nature of AI's technological dynamism manifests itself in the joint involvement of AI–related domains to achieve improvements.

Uniqueness of AI as GPT. Bekar et al. (2018) add another key criterion to the identification of a GPT. To be general–purpose, a technology should have "no close substitutes: (a) is unique — no other combination of technologies can produce an application; (b) without it the whole system (GPT and its application) would not work." The feature of uniqueness is helpful to discriminate between GPTs and what could be vaguely defined as radical innovations and important technological discontinuities. A GPT is identified as such because it does not have close alternatives: it is the sole and essential option to execute certain functions in user sectors. Uniqueness does not match well with current AI: in many economic activities, AI *is not unique*,

but rather working autonomously, (can be) more efficient or/and precise at executing certain tasks compared to humans or to alternative techniques. For example, automated and algorithmic way of performing HR or business analytics can have a significant impact on firms' efficiency and economic returns, but this is not the only way to run these activities. Even though current AI has made its way into new applications and improved end–user experience, the whole system would not fall apart if AITs were to be rolled back. As illustrated earlier on, AI induces system–level substitution through capital deepening — in particular replacing older software technology with newer, AI–powered one, but the before–AI way of performing a task remains a close substitute, and in many cases yet a more reliable and precise one.

- In sum, AI has close substitutes for the functions it provides: it is not unique and rarely essential for the functioning of user sectors.

Implementation lags of AI as GPT. The diffusion of a GPT is expected to generate non–linear impacts on economic outcomes, in particular productivity (Jovanovic and Rousseau, 2005). GPTs do not necessarily produce these macroeconomic effects (Bekar et al., 2018). However, given their novelty and appeal to a variety of uses, it is possible that GPTs display implementation lags. The reason for it is that in order to exploit in full the pervasive potential of a GPT, resources already employed in productive uses need to be temporarily foregone and allocated to develop complementary assets (Brynjolfsson et al., 2021). For GPTs, implementation lags are demand–driven: in order to adopt it, GPT–users need to incur adjustment costs, among which those for organisational changes, capital investments, and development of skills to handle the new technology. In the case of AI, implementation lags are not necessarily driven by the same mechanism. The bottlenecks delaying AI implementation are mostly supply–driven: AI–producers need to obtain required inputs (data, hardware, and skills), set up production process and deliver a minimum viable product. For example, the collection of datasets for training AI models can take time and postpone the launch of AI products. AI producers can shorten the implementation lags by acquiring data on data marketplaces, exploiting cross–product data feedback loops, training their models using pre–trained models (teacher–student) or by "faking until they make it" using AI 'impersonators' (Tubaro

et al., 2020) to buy time while training data is collected. These strategies are viable only in some cases and for some AI companies: data trade and access can be regulated; data feedback loops can be exploited almost exclusively by multi–product firms; the pre–trained models must be available, trustworthy and provide sufficient quality. Notwithstanding the potential remedies, and in contrast with the case of GPTs, bottlenecks for AI implementation remain a supply issue.

Taking stock. Is AI a GPT? Not exactly. AI is not pervasive in a GPT sense. It reaches adoption at scale only in a handful of industries, and even there diffusion is concentrated in and driven by a few large lead actors. Similarly to the Internet, AI provides an additional layer of functionalities to end–users rather than penetrating the economic structure: it is superimposed on existing systems. At the same time, its innovational inducement does not have a simple one–to–many nature and goes beyond a mere stimulation of the innovation rate in application sectors because AITs *participate* in and help navigating invention and innovation processes. AI technological dynamism is systemic and results from advances in interlocked sibling technological domains. Finally, AI has close substitutes, and its implementation can be subject to lags, but compared to a GPT the source of the lags for AI lie in the supply rather than demand side. Using an econometric metaphor, *GPT is a misspecified model of AI*. The GPT misspecification originates from a potentially incorrect use of the included variables (functional misspecification) and, most importantly, due to omitted variables. The latter has two implications: first, it under– or overestimates of the importance of the included factors and, second, it misses a number of dimensions to represent AI adequately. Incorrectly specifying AI as GPT boils down AI to a poorly fitted, flat representation of what is instead a multidimensional complex phenomenon. Misspecifying an infrastructural technology as a single component will lead to incorrect inference and is likely to produce misleading predictions. It is possible to find a scheme that suits better the nature of AI. In the next section, we follow this route and try to look beyond AI–as–GPT.

4.3 Artificial Intelligence as a Large Technical System

4.3.1 Large Technical Systems

Large technical systems (LTS) are "spatially extended and functionally integrated socio–technical networks" (Mayntz and Hughes 1988). The notion belongs to the fields of sociology and history of technology, and science and technology studies. Compared to specific and isolated artefacts or technologies, LTS are 'system artefacts' or system technologies. Recognised examples of LTS are, among others, telecommunications, railways, energy supply and distribution systems. The prevalence of physical infrastructures among the mentioned examples of LTS does not exclude system technologies characterised by a higher degree of intangibility to be classified as LTS. In fact, Ewertsson and Ingelstam (2004) identify information–based LTS that contain both 'hard' and 'soft' components, such as radio and television distribution networks. Since the very introduction of the notion (Hughes, 1983; Hughes et al., 1987), the literature on LTS has investigated an array of issues characterising these system technologies, from definitional issues to the exploration of their dynamics and key actors. For the aim of this Chapter, the value added of the LTS theory lies in two dimensions: first, the outline of the different phases an LTS will experience from birth to maturity. Second, the identification of specific building blocks and driving forces that contributes to the formation and development of an LTS. These two dimensions are related, as different driving forces play a different role and have different relevance along the phases of LTS evolution.

The LTS phases originally singled out by Hughes et al. (1987) are (i) *invention*, (ii) *development*, (iii) *innovation*, (iv) *growth, competition and consolidation*, and (v) *technology transfer*. The latter is characteristic of LTS: technology transfer occurs when an LTS developed in a given context is replicated in other environments, and can happen in parallel to other phases. More recent work added new phases experienced by mature LTS, such a *stagnation, reconfiguration* and *decline* (Sovacool et al., 2018). Furthermore, Gökalp (1992) stresses how LTS develop by layering up over existing sys-

tems, creating a *superposition of systems* that shape an LTS configuration. The superposition of systems is characteristic of infrastructural projects and is an important feature to detect in an LTS. Complementary to the development in phases, a given LTS can be described as the result of a series of driving forces playing out to shape the infrastructural technology: *system builders, reverse salients, load factor, technological style*, and *momentum*.

System builders. System builders are the actors that strive to extend the reach of the system and perform the sociotechnical integration necessary to its deployment (van der Vleuten, 2009). These can be inventors–entrepreneurs or manager with engineering capabilities, individual actors or large firms. In different phases, system builders align the interests and objectives of the different actors involved, allowing an LTS to grow and achieve its goal(s).

Reverse salients. Reverse salients "are components in the system that have fallen behind or are out of phase with the others. Because it suggests uneven and complex change, this metaphor is more appropriate for systems than the rigid visual concept of a bottleneck. Reverse salients are comparable to other concepts used in describing those components in an expanding system in need of attention, such as drag, limits to potential, emergent friction, and systemic efficiency" (Hughes et al., 1987). Reverse salients, emerging from the uneven development of the system's components, are sources of critical problems and, given that problems are typically focusing devices (Rosenberg, 1969) to allocate innovative efforts, they are also potential loci of innovation.

Load factor. Load factor is "the ratio of average output to the maximum output during a specified period" (Hughes et al., 1987) and it is an indicator of performance, here meant as use or deployment of the technology at full potential over time. The distribution of load factor indicates when and where the system is under stress. Knowing that can guide investments in capacity expansions or adjustments, as well as policy interventions.

Technological style. As for the common use of the word, style indicates a type of fashion: the specific design of a particular LTS that descends from choices regarding which features are emphasised, and in which way. An

LTS technological style emerges from the particular choice and combination of its elements, given their relative importance and the specific role they play in the whole system. LTS executing the same function and aiming at the same goal can differ in style in different contexts. For example, the organisation and control structure of energy distribution systems can change across countries while the fundamental function and goal they pursue are comparable.

Momentum. Momentum, or dynamic inertia, is the degree of autonomy the LTS acquires once it reaches a certain stage of development and a 'mass' in terms of relevance for the economic system. Systems with high momentum are less sensitive to pressures for change — they continue their 'motion' undisturbed.

The concept of system builder has mostly a social aspect, while reverse salient and load factor are dimensions of purely technological nature. Many of these concepts have closely related siblings in the field of economics of technological change. For example, reverse salients approximate bottlenecks; momentum approximates path dependence and cumulative change. However, their engineering or social flavour makes them more sophisticated categories to label complex phenomena, enriches the economics perspective and makes them useful to capture the features of system technologies that are uniquely embedded in specific epistemic communities, regulatory settings, and cultural contexts. A system builder can be an entrepreneurial actor, but also a carrier of a rare combination of technical and social skills (and, potentially, power). Momentum is close to path dependence, but path dependence is a process that emerges from chance and choices, while momentum is a later–phase property of a system that keeps existing and functioning due to 'mass' and acquired autonomy, thus refusing any role to chance.

The LTS categories are useful to guide the analysis of a given system technology. For example, one might want to know: where is the 'locus of control' in the system? Which actors store and hold the relevant technological (and market) knowledge to 'produce' the system technology? Who advances and builds the system out of its components? Who has power on the factors constraining the development of the LTS? Which elements of the systems and

related actors can facilitate the process of convergence around standards and protocols in order to improve communication and control at large? What happens if the LTS becomes so large to be unmanageable? Joerges (1988) quotes Aristotle, reminding us that when things get too small or too large "they either wholly lose their nature, or are spoiled". A very timely point, when endless accounts of misuses, biases, discriminatory and malicious deployments suggest that we might be already spoiling AI.

4.3.2 Recognising features of LTS in AI

In Section 4.2.2, we checked some features of current AITs against GPT definitional criteria. The resulting picture suggests that AI substantially differs from a GPT, due to its rather infrastructural, distributed and heterogeneous nature. An alternative view on AI needs to encompass the whole circuit of actors and interconnections involved in its production and diffusion, their distinctive push and pull exert on the whole system, and a representation of how dispersed but linked activities influence the momentum of AI. We claim that LTS well–approximates the infrastructural nature of AI. To support this claim, we now identify element by element the LTS features in AI.

AI is large. LTS draws its specificity from the use of the attribute large. Following Joerges (1988) and Gökalp (1992), large can be considered in terms of territorial or user coverage, involving large–scale actors in the production of technology, or generating far–reaching socio–economic and/or environmental impacts. In this sense, large is used to label a technology that is encompassing, infrastructural, impactful, costly or global, or a combination of these properties. The attribute large is partially overlapping with GPT's pervasiveness, but it is broader, easier to measure, and fit for the purpose of describing an infrastructural technology. For example, a technology adopted only in one sector but at scale can be considered large. A different technology used marginally across a wide range of economic activities could be large as well, as overall it amounts to a wide reach.

The way in which large is measured in the LTS framework well–suits the

measurement of the largeness of AI. Regardless of the few industries in which it originates, current AI spreads large in user base and territorial coverage given the widespread accessibility of its end–user applications. AI is large also because it is developed and promoted by a large community of actors (developers and vendors). The frontier of this large community is represented by large actors — large companies (the tech giants), national and supranational institutions, industrial consortia, global networks of Universities and organisations (and dedicated conferences). This creates a situation in which a large actor invests substantially into AITs and provides access to it to a large consumer base for sharing. For example, this is the case of sharing computing facilities and storage via cloud, or AI–powered software–based services (AI–as–a–service) such as visual recognition systems for airports. The economies of sharing (Shapiro et al., 1998) at work with AITs make the latter similar to classic LTS such as transport and energy supply systems. Finally, the societal traction of AI is large: "AI has seen itself elevated from an obscure domain of computer science into technological artefacts embedded within and scrutinised by governments, industry and civil society" (Mohamed et al., 2020); whole public opinions debate the changes AI will bring to contemporary societies, from its effects on employment, development and inclusiveness, its impact on minorities, and its environmental toll.

- In sum, AI is large according to various criteria identified by the LTS framework. This characteristic is better defined, inclusive and, hence, more convenient for both the identification of LTS and its empirical analysis.

AI is a technical system. AI is already implicitly considered a system from its very essential representation. The view of AI outlined in Section 4.2 helps to shed light on three constituent domains or subsystems that are key for the development of AI as LTS. First, the domain of AI algorithms that, in terms of actors involved and specific system builders engaged, is a subset of the software industry. Second, the domain of computation, in practice constituting a subset of the hardware industry. Third, the domain of data generation, collection, storage, analysis and transaction: data is

collected and organised by public and private actors, globally and locally. As in a Venn diagram, at the intersection of these three domains one can find the state–of–the–art AI. These three domains are large in their own right according to the criteria we used early on: they are widespread (even if often invisible) in physical space, they contain numerous and large actors, and they are interwoven with and impacting socio–economic activities.

When discussing technological systems, Hughes et al. (1987) posits that they "contain *messy, complex, problem solving components*. They are both *socially constructed* and *society shaping*" (italics added). We unpack this statement to show how it tailor–fits to AI.

The AI LTS is messy. AI is still characterised by the turbulence typical of nascent industries, and uncertainties prevail with respect to its technological trajectories, its overall design, and its impacts. In the overall design of the LTS, one can devise alternative scenarios. As corner solutions, an AI LTS can be established either with a few large system builders dominating all the parts of the system or with an ecosystem of small actors scattered across domains. Intermediate settings, in turn depending on the direction taken by the regulation and governance of the LTS, can have large actors taking over some domains while leaving others untouched. Here, the relevant issue is to balance or align the societal and private interests of system builders and to identify important forking points in the path dependent process of AI development before the system gains so much momentum to become resilient to corrections. The very direction of evolution of AITs depends on the step–wise resolution of the current 'messiness'.

The components of the AI LTS are complex and directed at problem solving. Each of the domains of the AI LTS fits into the above statement. The case of AI chips production well captures the complexity of the hardware and computation domain of AI. Prytkova and Vannuccini (2020) summarise the trilateral frontier chipmakers address when developing their products: resolving a technical trade–off among delivering processing speed, energy efficiency, and heterogeneous computing. The data domain of the AI LTS is also complex: its current configuration is shaped by actors' competition to settle regimes of ownership and appropriation of data (Koutroumpis et al.,

2020a). Spiekermann (2019) illustrates the structure of an *ideal–type* data marketplace that includes data buyers and sellers, the data marketplace (exchange) owner, and third–party service providers. AITs might be just tangent to the main goal of such data marketplaces (trading data), but perform an auxiliary function within this mechanism. From an AI–as–LTS perspective, the complexity in this domain arises from the fact that AI–producing and using companies can adopt different configurations: they can act as third parties only (AI–services providers), they can merge the role of third–party service provider and data buyer (e.g. using AI as a complementary technology to improve advertisement), and even layer–up the role of 'data exchange'. The latter is currently the case of Google, which owns a data exchange, uses AI to improve its products offer, and provides AI–based services (Srinivasan, 2019).

The AI LTS is shaping society and it is socially constructed. Harmful AI uses become increasingly evident the more AITs are implemented and turn into commercial and administrative tools. Concerns grow over specific applications of AI (e.g. face recognition), the ethics of algorithmic decision making, the safety of AI systems (e.g. to adversarial attacks), and the 'data colonialism' (Couldry and Mejias, 2019) premises on which these technologies are built, leading to a social pushback against harmful AI (Crawford et al., 2019). The acceptance or resistance to AI developments determines the social construction of this LTS (Mohamed et al., 2020). At the same time, the deployment of these technologies shapes society, in terms of perceptions (regarding, for example, the fears of AI–driven technological unemployment and widespread surveillance coexisting with the techno–optimism of grand opportunities on the brink of a fourth industrial revolution) and tangible implications. For example, companies started optimising their language and format in reports and disclosures, anticipating that these will be analysed by AI algorithms (Cao et al., 2020).

Finally, AI development clearly displays a *superposition of systems* (Gökalp, 1992). Current AI layers up on other LTS such as telecommunication and Internet infrastructure. Within public and private organisations, elements of their information systems are upgraded through AI–based capital deepening, while remaining integrated in existing organisational routines. For example,

AITs integrate with existing database technologies, a trend already started decades ago (Brodie, 1989).

- In sum, computation (hardware), algorithms (software), and data domains are complex constituents of the AI system. The latter is yet messy, and hence contains the potential for different technological trajectories to unfold, leading to different designs of the system (e.g. distributed vs centralised, specialised vs general). This dynamic convergence to a more mature AI LTS happens through a simultaneous impact on and by the society.

Now that we have recognised the nature of LTS in AI, we can proceed with a more fine–grained analysis of the AI LTS by identifying the building blocks of LTS presented in subsection 4.3.1 in AI.

AI system builders. AI and its constituting domains are constructed by a variety of actors that actively initiate, support and shape developments of the system. The system builders in AI LTS are *AI–producing and AI–using companies, dedicated regulatory bodies, industrial consortia, non–profit research organisations*. Every system builder exerts efforts to influence the selection of their priorities and problems for the system to implement and address. This can be done by trying to weave in a particular way the network of elements in the systems (an example are tech giants hiring AI pioneers and leading figures to lead their AI programmes) or by forcing the very system to converge on new standards, protocols and shared practices. The latter can be achieved by making obsolete or ineffective the status quo through, for instance, forking decisions (Simcoe and Watson, 2019).

Consider AI–producing and AI–using companies. The latter are mostly recipients of the novel technology and are experimenting with ways to integrate AI in their 'runtime' (Iansiti and Lakhani, 2020), while the former are proper system builders. AI–producing companies have the power to design the system and to decide which bridges between actors and subdomains to build or cut–off. AI–producing companies are an ecosystem of firms that conduct AI research, develop AI solutions, and participate in the

AI value chain (Tubaro et al., 2020).[10] Among them, key system builders are the already mentioned tech giants, and established software and hardware companies[11], vendors, startups and platforms[12]. In terms of main line of business activity and industrial classification (NAICS), the majority falls under the codes 'software publishers' (49%) and 'Computer Systems Design and Related Services' (17%).[13]

One example that illustrates how system builders exert their power as such is the ongoing issue revolving around handling harmful AI. In this context, current commercial system builders have supported the establishment of ethical boards and voluntary guidelines rather than regulation. While regulation would impose common and accountable rules on the development of the system, commitment–based solutions can be considered forms of strategic concessions[14] to other AI LTS stakeholders (in particular regulators, consumers of AI services, and the society at large). In practice, these initiatives represent a 'seductive diversion' that allow AI–producing companies to show engagement while retaining full power over the design of the system. Another leverage AI system builders acquire with their role is their 'knowledge holding' (Steinmueller, 2006). In the production of AITs, novel know–how is created and knowledge about it settles in the hands of AI–producers. Through this process, AI system builders become knowledge gatekeepers that can facilitate the diffusion of knowledge and expertise through co–invention activities as well as strategically withhold it. For example, machine learning platforms — open or closed source (Isdahl and Gundersen, 2019) — might improve access to and reproducibility of AI solutions, however at the cost of product development knowledge being held for a larger share by the platform owner.

System builders in AI are changing over time, and their variety is increasing.

[10] For an attempt at identifying AI–related companies, see https://cset.georgetown.edu/research/identifying-ai-related-companies/

[11] These companies take the lion's share in AI patenting. According to Van Roy et al. (2020) top AI patenting firms (in the period 2010–2016) have their main activity in 'Manufacturing of electronic equipment' and 'Information and communication'.

[12] https://www.venturescanner.com/category/artificial-intelligence/

[13] https://cset.georgetown.edu/wp-content/uploads/CSET-Privately-Held-AI-Companies-by-Sector.pdf

[14] https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

AI companies superseded individual AI pioneers and universities' computer science departments; at the same time, they are accompanied by governments and 'lateral' organisations. The latter are, for example, advocating to make the system more inclusive and less harmful (e.g. AI Now), pursuing technical advancements through non–profit organisations (e.g. Open AI), facilitating coordination on principles and standards (e.g. the Partnership on AI), or stressing the importance of getting prepared to the emergence of strong AI (e.g. the Future of Humanity Institute). Another type of system builders, currently less empowered than the ones mentioned earlier on, are the (platform) workers that support the deployment of AI systems and that are subjects of processes of 'heteromation' (Tubaro et al., 2020).[15] These workers operate at the margins of AI and fill gaps in the working of the technology — they run the so–called 'AI last mile', either fuelling the data necessary for the training of algorithms, verifying their performance or even emulating the results of AI systems.

AI reverse salients. As the system scales and becomes larger, tensions appear. These fault lines are the reverse salients of the system. One recurring source of reverse salients in AI are the system's scarce resources in AI's domains: *being a nascent industry, AI lacks input resources from its domains. The shortage is relative among domains, i.e. the worst performing domain is a source of reverse salient, which can be of quantitative or qualitative kind: delivering an insufficient amount of an input resource, or a qualitatively unfit input.* This holds back or derail the evolution of AI LTS. Quantitatively speaking, AI is data–hungry but other resources whose demand grows faster than supply can become constraining factors as well: among them, AI labour and AI–programming skills, and management trained to lead AI–powered companies. Qualitatively speaking, computation is a reverse salient that falls behind because the dominant design of chips in the semiconductor industry doesn't match the way modern AI operates, even though some alternative

[15] Introduced by Ekbia and Nardi (2017), the concept of heteromation documents the shift, in some sectors, from technologies of automation, in which machines take over tasks from humans, to technologies of heteromation, in which tasks at the margin of value creation that are devoted to the management and update of automated systems are externalised to humans. In this sense, rather the producing unemployment or the end of work, the introduction of automated systems actually increases labour demand; this is however concentrated in labour intensive micro-work activities, with consequent detrimental impacts.

trajectories emerge (Prytkova and Vannuccini, 2020; Hooker, 2020). Atop of the purely technical challenge, there is also a techno–economic one: the competition between cloud and edge modes of organisation of the computing infrastructure. The cloud mode entails allocation of resources with priority on coverage and speed of access networks, and on computing capabilities of cloud–providers. The edge mode emphasises connectedness (which is not the same as coverage), compatibility and computing capabilities of edge devices. Evidently, the choice of the prevalent mode defines the chips' design that will get the lion's share of R&D efforts. The resolution of this qualitative reverse salient from the computation domain will shape the AI LTS with regard to the organisation of computation: inside the chip among its components as well as inside the industry among producers and users.

Another example of qualitative reverse salient is the lack of a well–designed and regulated architecture for data troves, originating from scarcities in the data supply. AI systems can rely (i) on public dataverses and open data or (ii) on the supply of data from data marketplaces. These two alternatives carry with them two philosophically competing views: in the first, the reverse salient is addressed through the creation of common assets; this can give a more democratic style to the AI LTS, but can also open room for nationalistic interpretations of the idea of data 'sovereignty'. The second view might lead to an 'oligopolistic' style of the AI LTS, with a few powerful players shaping the playground at their own advantage. Compared to the 'AI commons' scenario, the oligopolistic one might hasten the growth and impact AITs, but can lead to a more unequal distribution of returns.

Reverse salients emerge also in the domain of AI algorithms. One lies in the proliferation of AI software and programming environments, slowing down the convergence towards a dominant design. Part of the community of AI developers urges technical improvements through recognised contests dedicated to different AI problems[16], open–source platforms to assist the coherence of the community and the development of cross–compatibilities, the establishment of standardised libraries and programming frameworks, and more fundamental theoretical and technological advances (Ben-David

[16]Numerous examples can be found here: https://paperswithcode.com/datasets and, more hardware–oriented, the MLPerf training benchmark: https://mlperf.org/.

et al., 2019; Geirhos et al., 2020; Marcus, 2020).

Another reserve salient is overspecialisation among AI algorithms. Despite AI algorithms become increasingly capable (see Hernandez and Brown (2020) for an assessment of algorithmic performance and efficiency trends), the tendency for ad hoc solutions remains. The reason for that lies in the pursuit of a sole criterion of performance (or its derivatives), namely, out–of–sample accuracy of prediction. The development of algorithms proceeds along this criterion and hence relies heavily on the intensive margin, a trend succinctly expressed as "the bigger the better" — whether bigger refers to the size of a model, of data or of computing power. Figure 4.2 supports this statement plotting accuracy versus model size for two different AI tasks, visual recognition and natural language inference.

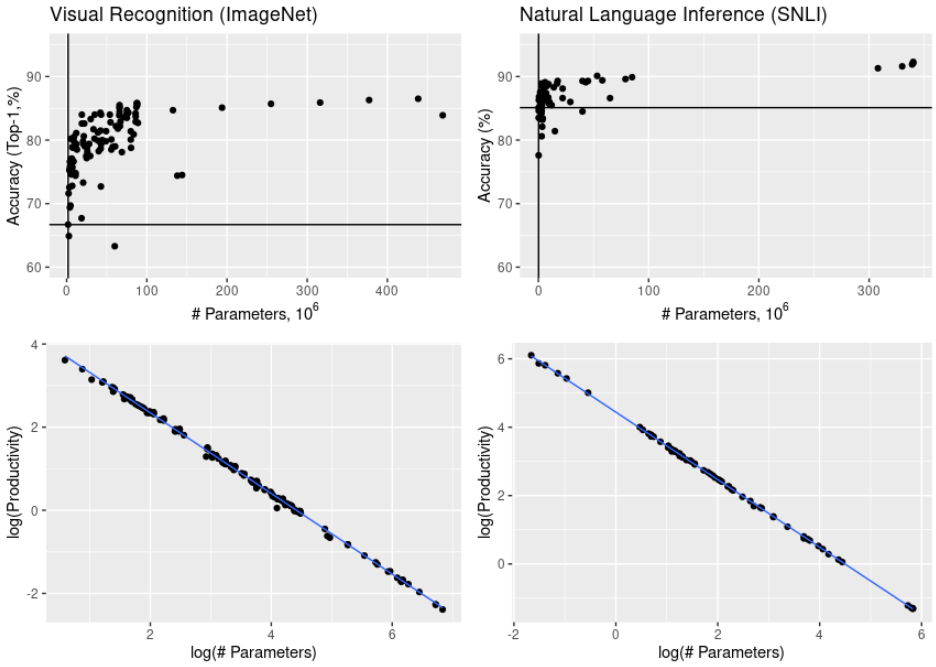


Figure 4.2: Models' retarding performance and decreasing efficiency in visual recognition and natural language inference tasks

Note: Each observation is a model with reported specifications and performance trained on the same data sourced from the respective benchmark test. *Upper panels*: accuracy of out–of–sample prediction plotted versus number of parameters in a model measured in millions. An observation at the intersection of the Pareto–Koopmans criterion borders (black lines) marks the model with the highest return on number of parameters. *Bottom panels*: linearised relation between model size and productivity measured as ratio of accuracy to number of parameters labelled as productivity. *Left panels*: visual recognition task using data of ImageNet competition. *Right panels*: natural language inference task using the SNLI corpus.

The upper panels of Figure 4.2 show decreasing returns to number of parameters in both tasks: as the number of parameters in a model grows, the corresponding gain in accuracy is getting smaller. Black lines represent borders of the Pareto–Koopmans criterion (PKC) (Bogetoft and Otto, 2010); at the intersection of the PKC borders lies a model with the highest accuracy to size ratio, i.e. productivity. The empty second quadrant indicates absence of more efficient observations; after the intersection point the returns on model size are decreasing in terms of accuracy of prediction. The lower panels of Figure 4.2 show that the linearised relation between model size and productivity is strong for both tasks. A deviation upward of the fitted line would indicate a higher return on the number of parameters than expected for a corresponding model size, but there are no such deviations. These results illustrate the claim of algorithms' development along the intensive margin, with accuracy improving at slowing down rate at the expense of accelerating model size. For example, in 2020 GPT–3 model by OpenAI has 175 billion parameters, 100 times larger than models launched two years before. The new frontier of model size, achieved in 2021, is Google's Switch Transformer, featuring more than 1.5 trillion parameters and hence jumping in size by a factor of 8.6 in one year. Techniques like parameter pruning, quantisation, transfer learning, and the usage of lower precision arithmetic might be steps towards more efficient models.

Reverse salients originated in the domain of algorithms have implications for the hardware domain: ad hoc AI algorithms appeal to smaller demand and have short–lived returns, quickly becoming obsolete. At the same time, the design and production of a chip that caters the needs of an ad hoc AI solution has high sunk costs. Therefore, the resolution of reverse salients in the algorithm and hardware domains is entangled, and both remain in a turbulent state until a dominant design emerges in either of the domains. Though connected, the development of the two domains remains driven by rather separate criteria, treating each other as a source of constraint for its own performance rather than a tandem partner with a common goal. The current AI shock shed light on this issue, exposing the shortcomings of such divided approach; there appeared several studies that call for an expansion of performance criteria into various *joint efficiencies*, beyond separately accuracy in the software domain and processing speed in the hardware domain

(Chen et al., 2019; Hooker, 2020; Strubell et al., 2019; Prytkova and Vannuccini, 2020). In line with our discussion on systemic technological dynamism, lack of awareness about the joint nature of improvements brought separate technological trajectories to their exhaustion and further opportunities are recognised in joint efforts of the software and hardware domains.

AI momentum and load factor. A system can be improved by solving tensions created by reverse salients, producing what Sahal (1985) defines *learning by scaling*. The learning occurs by accumulating understanding on which changes (innovations) to implement, and how to redistribute the load in the system while it scales up. In other words, knowing the load factor guides directed interventions to address salients and to allow for improvements *ceteris paribus* the level of technical performance. On top of that, Nightingale et al. (2003) suggest the notion of 'economies of system' to explain the gains a system can enjoy by redistributing activities according to the load factor, dynamically balancing the stress.[17] Economies of system in the AI LTS would occur by rearranging the structural dependencies among its elements when some of them develop unevenly or are overloaded. For example, the shift to federated learning architectures (Li et al., 2020) would represent a system re–arrangement towards a design potentially capable of addressing the computation–related reverse salient: this would be done by distributing workload over the networked components rather than leaving few giant actors to route (and control) finite computing power in the cloud.

Finally, the growing number of actors jumping on the bandwagon of AI successes, the grandiose media coverage of AI advances (in particular for what concerns language models and generative models of speech — the so–called conversational agents like Project Debater of IBM (Slonim et al., 2021)) and the expectations of further ubiquitous diffusion of AI build up a strong momentum for the AI LTS. However, expectations can work in both positive and negative direction. On the one hand, they channel large investments in AI R&D by public and private system builders. On the other hand, the expectations of a large and ubiquitous impact of AI risk remaining unfulfilled: sustained commercialisation and growing competition among

[17] The notion shares similarity with that of architectural innovation (Henderson and Clark, 1990), where improvements in performance are achieved by changing the arrangement of the constituents of a technology while keeping its function invariant.

system builders make them race against each other, undertaking myopic steps in AI development leading to short–term payoff. Stagnating diversity of AI research is among the early signs of such dynamics (Klinger et al., 2020). As the expectations that a new AI winter might be at the horizon start to be considered plausible and the AI hype slows down, the momentum of the system might follow a similar path.

4.3.3 AI LTS: State–of–the–art

Having identified the technological and non–technological features of LTS in AI, we can proceed with a description of state–of–the–art AI LTS: the phase of development the system has achieved, the boundaries that confine the system, the mechanisms of control currently in practice, the distinctive style emerging, and finally the goal or a main function the system embeds.

Current Phase. The 'invention' phase of AI is a contested territory, as the very understanding of what AI is shifts over time; this is why in Section 4.2 we offered a view of current AI. We can claim that following the impressive results in the ImageNet visual recognition competition in 2012 and the subsequent media interest in AI — mostly due to the shadows AI seemed to cast on the future of work — the AI LTS went through the phases of invention and development. The current state of the AI LTS is now in–between the phases of innovation and growth, competition and consolidation, with commercialisation accelerating its pace and increasing technology transfer from academia to business, including a sizeable talent drain of professors and graduate students (Zhang et al., 2021). The very process of growth by expanding to novel application fields generates continuous feedback into the phases of innovation and development. Technology transfer has also accelerated with the increasing efforts of national, supranational and sub–national institutions to govern AI developments as the technology has acquired geo–strategic significance.[18] While in the process of innovation and growth, an interesting question regards whether a process of technological

[18]The reader can refer to the OECD AI Policy Observatory (https://oecd.ai/) or to Nesta AI Governance database (https://www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database/) for a collection of AI–related policy initiatives and institutional strategies.

convergence is taking place within the AI LTS. Technological convergence, a concept introduced by Rosenberg (1963), is a form of 'upstreaming', a process occurring when an activity embedded within diverse sectors or/and tasks exhibits some common features and principles that eventually maturates and unbundle into a fully–fledged sector on its own.[19] We see signs of this process at work in the evolution of the AI LTS: AI–producing companies — key system builders — emerge as specialised suppliers of AI–as–a–service tools, business automation services (e.g. recruitment), scientific discovery tasks ('science–as–a–service') and data analytics tasks. The success of a few system builders to impose their template on the working of the AI sector will influence the evolution of the whole LTS infrastructure in the future, set the stage for novel reverse salients to appear, and orient policy priorities.

Boundaries of the AI LTS. While the AI LTS evolves, it passes from one phase to another and its boundaries change: the system grows larger, usually in an uneven manner. This creates the problem of where the AI LTS 'begins and ends'.

The span of the AI system was at first limited to a pure scientific territory at the intersection of computer science and psychology; then, the entire landscape of business actors that could commercialise AI has been incorporated into the system.[20] The first generation of AI–based products and services has been introduced, and their overlooked flaws — especially with respect to AI safety and ethics — became a fertile soil for the next cohort of actors that address these flaws to enter the system. Progressively, in the described manner, more areas were incorporated in the AI LTS advancing its frontier. The application fields of medical and biological sciences, robotics, automated decision making in business and public administrations and the military sector have recently stepped into the LTS and add new forces to the complex casting of the direction and pace of AI evolution.

[19] The separation of chip production from the rest of the computer industry or the classic formation of the machine tool industry described by Rosenberg (1963) can be considered cases of technological convergence. When technological convergence takes place, a new industry can emerge upstream, producing generic technologies that suit a wide variety of downstream purposes.

[20] Earlier AITs, such as expert systems, extended already in the 1980s the boundaries to Industry and commercialisation activities (Nilsson, 2009).

Control over AI LTS. A relevant issue related to the shifting boundaries of the AI LTS is the possibility that diseconomies of scale exists, leading the system to lose internal coherence and to face 'crises of control' while growing (Beniger, 1986). Under control we understand mechanisms of coordination on progression and management of the deployed applications. The issue of control arises when these mechanisms are non–existent or function inefficiently. Nightingale et al. (2003) study and provide examples of innovations in control technologies as devices to retain control of LTS as they scale up. The issue of control, however, seems to be less relevant for the AI LTS: given the heterogeneity of its components, the system is less coherent as an ensemble compared to more homogeneous LTS. This could be the result of AI being still a 'young' LTS, with a yet fluid distribution of power over the system. In a sense, among LTS, AI is closer to the Internet than to integrated transport systems. As with the Internet, AI displays a mix of centralised and decentralised mechanisms of control and a layering up of commercial and non–commercial areas of development (Greenstein, 2020). Coordination among the system builders is achieved through the convergence on standards and interfaces, which is a non–frictionless process. From the perspective of control, in AI, there are not yet agreed standards for progression of the system, nor an essential need for the centralised maintenance of coherence among the system's parts; also, the actors do not have to be explicitly aware of the infrastructural nature of the AI LTS in order to conduct their operations. Therefore, the minimum requirements to make the AI system working and to avoid the system falling apart are lower than for other infrastructural technologies, while the social impact is potentially larger for AI. This does not mean that AI evolution will continue to follow the same loosely coordinated path: there are coalitions of system builders in formation, supporting the idea of development and implementation of standards (both technical and ethical) to 'govern' AI. Such coalitions have different goals and different shapes, according to the kind of system builders they aggregate: workers and end–users of AI (Crawford et al., 2019), strategic alliances, intergovernmental initiatives (such as the 'Partnership on AI', and fully–fledged consortia. This activism represents the search for directions of progression of the system. In the current phase, the system already gained significant momentum, but 'diseconomies of scale' have not occurred and internal coherence has not been upset due to the compartmental struc-

ture of AI LTS. Instead of the 'death' of the AI system, the failure of control mechanisms could steer AI towards towards detrimental directions at crucial points of its evolution path, where detrimental is intended here in the sense of harmful or exploitative of a share of its actors or users.

Technological style. The AI LTS will display different technological styles in different environments. The specific design of the system constituents and the strength of their interdependence vary according to the priorities, strategies, and decisions taken by system builders, regulators, and users. This becomes evident when considering national (and supra–national) implementations of AI systems. The 'division of labour' and the direction of AI development and deployment depend in part on the structure of AI– and data value chains (Tubaro et al., 2020), but it can be strongly affected by government strategies, resulting in rather distinctive styles. The first distinction can be between the technological style that shapes around the effort of *prolonging the 'age of discovery'* and the one focused on *accelerating the switch to the 'age of implementation'* (Lee, 2018). The first style stresses the role of continuous innovation in AI techniques. The second considers the main AI breakthroughs as already achieved and aims at scaling them up by accelerating diffusion and experimenting with applications to capitalise on them.

AI styles can form under influence of diverse system builders. Let's consider the role of the state in giving the AI LTS its style. Technological style can emerge as a result of the particular policy levers and priorities the regulator decides to pursue. This is reflected in public budget allocations that can channel funds to AI through the university system, the military sector[21], or directly to private actors — for example in form of financial support to AI startups. Beyond the sheer amount of expenditures and the broad direction imposed on the system's evolution, different technological styles for AI LTS can emerge as a result of the specific tools of technology policy used (Steinmueller, 2010). Here, *top–down command and control policy* actions share the stage with more *bottom–up governance initiatives*. Horizontal interventions, such as the design of regulatory frameworks against AI harms,

[21]See the 2021 Report of the US Department of Defense or National Security Commission on Artificial Intelligencehttps://www.nscai.gov/2021-final-report/

misuses and biases, are part of a specific style. The creation of new dedicated institutions (as it has been debated regarding the possibility to create a federal robotics commission in the US — see Calo (2014)) and intermediate bodies to facilitate coordination in the system is another, potentially complementary, option. Examples of policies that can influence technological style are the efforts made by governments to attract, retain and develop AI talent through the visa regime,[22] or the alignment of macro policy levers (e.g. immigration and trade policy) with AI–related strategic priorities. A relevant case of the latter option are export controls policies targeting the semiconductor industry, as this is the producer of key components for AI–tailored hardware and its productive capabilities are a fundamental strategic asset. The use of policy levers in strategic technologies such as AI is not a novelty: the just cited semiconductor industry has been subject of trade policy interventions to shield domestic companies against emerging competitors (Langlois and Steinmueller, 2000). This point highlights the tight link existing between the actions leading to the development of a technological style and the competition between institutional actors at the international level. AI becomes ones of the territories over which geopolitical forces play, so much that some authors discuss whether an AI 'arms race' is ongoing (Asaro, 2019).

Goal orientation of the AI LTS. As a final step in our analysis, we assess whether AI has an identifiable goal orientation. LTS are goal–oriented systems; this means that all components coordinate — easily or at the cost of frictions, negotiations and forced adjustments — to achieve an overarching aim. This characteristic is easy to detect in LTS such as transport systems or water distribution networks (the goals being, respectively, mobility and water supply), while it is less evident for telecommunications, the Internet or, indeed, AI. The reason why the goal is less evident for the last three cited systems is because not only they are large, but also highly heterogeneous and less prone to have centralised control being exerted on the LTS participants. We posit that AI LTS is also goal–oriented, even though the goal might not be clear yet to all the actors involved. We suggest that the goal of the AI LTS is a 'cybernetic' one: *to re–domain the 'fabric of control' of socio–technical*

[22] https://cset.georgetown.edu/research/immigration-policy-and-the-global-competition-for-ai-talent/

systems based on human decisions into an automated one, starting with the transformation of existing activities into instances of adaptive prediction.

Within the closed system of steps required to complete a given task (Agrawal et al., 2018a), currently, AI is focused on perfecting pattern recognition and prediction capabilities. In general, prediction serves the needs of the decision–making stage of task performance, as "each prediction task is a perfect complement to a decision task" (Agrawal et al., 2019c). Essentially, any prediction shrinks the search space that will inform action and lead to a desirable outcome. Up to now, AI is allowed, trusted or able to take decisions only in a few contexts. Simulated, virtual or physically–confined environments such as games, trading and test sites (e.g. for cars, drones and robots) are the contexts at the forefront of AI performing as an autonomous decision maker. In all the other contexts, judgement or decision–making remain overwhelmingly in the hands of the human counterpart. However, the consideration of a task as a closed system suggests looking at the entirety of the process of task performance. To establish full cybernetic control over the performance of a task, AI has to permeate each stage necessary to that task's execution: (i) elaboration of input data (pattern recognition, prediction), (ii) judgement and decision–making, (iii) action and feedback. An example of task controlled by AI along all stages is the industrial control system of cooling facilities in Google's data centres that went completely autonomous in 2018.[23] In general, it is clear that the achievement of the overarching goal of cybernetic control requires the maturity of multiple technologies and institutions, and their coordination. To accelerate or steer this process, reverse salients (technologies, mechanisms, institutions) that are falling behind and holding back AI can be identified adopting the view of AI being an infrastructural technology already now and an LTS in the future. In sum, to use the terminology of Flueckiger (1995), the goal of the AI LTS is to shift further the balance from economies based on operations of transformation to economies based on operations of control — and to automate these.

[23] https://www.technologyreview.com/2018/08/17/140987/google-just-gave-control-over-data-center-cooling-to-an-ai

4.4 Implications for Policy and Strategy

Seeing AI as an LTS rather than a GPT has important implications for policy and strategic decision–making. The core argument here is that the rationale for and the essence of intervention differs between the AI–as–GPT and AI–as-LTS case. To illustrate that, we can compare how the focus of policy might change by changing the categorisation of AI. When a technology is identified as a GPT, the rationale for intervention lies in market failure. The key issue is the under–production of the GPT technologies due to the distributed nature of downstream innovative efforts, which would require coordination. Fixing a coordination failure in the GPT case means kick–starting the dual inducement mechanism, raising the rate of investments in innovation until to foster positive feedback. In this context, public procurement and contract spending can emulate, substitute or subsidise downstream demand. When a technology is an LTS, coordination issues extend beyond simple incentive formation, and become a matter of joint design and production of the whole network of technologies involved in the system. From this perspective, failures take the form of system or orchestration failures, with actors failing to develop the necessary ties and alliances to strike a balanced development of the system (Schot and Steinmueller, 2018; Robinson and Mazzucato, 2019). Rather than facing a stagnating innovation rate, reverse salients appear locally and slow down or disable the whole system, making it work inefficiently or even miss its goal(s) entirely. In system technologies, the source of failure might be located within one component, distributed among several components or even be the very disconnectedness of the system itself. For an LTS, the correct identification of reverse salients and the detection of their composition and reach across the system is a primary step to undertake. Once diagnosed, the task becomes to devise a strategy to tackle the problematic areas of the LTS network, inducing desirable effects and preventing the side effects of the 'treatment'.

From this perspective, the AI LTS requires policy makers to get to know the specificity of the system under consideration: who are the system builders, where are the boundaries of the system, which mode of control is at work at a given moment and locality, how the load factor is measured and distributed. Policy makers must adopt systemic thinking to acquire awareness

of the state of the LTS, its current phase and potential paths of evolution, in order to inhibit detrimental or catalyse dormant useful activities, components and actors, fill gaps and missing links in the system, rebalance control or redistribute load factor, and in general to decide if to opt for command–and–control types of intervention or to prefer indirect forms of governance. Depending on which reverse salient is addressed, policy can opt for a different recipe of science, technology, industrial and competition policy tools (Steinmueller, 2010).

To show how strategy and policy can be discussed from the AI–as–LTS perspective in details, we take the AI reverse salient related to data and summarise dimensions relevant to AI deployment and upon which policy makers can act. Over the last 10 years, we observe a growth of business models that are reliant on the monetisation of data. The diffusion of the Internet and the globalisation of markets at the same time made possible an unprecedented expansion of the consumer base, a boom in the amount of offers from businesses of all kinds, and drastically lowered the related (information) search costs and the cost of tracking the consumption behaviour (content, goods, services, etc.) of online users (Goldfarb et al., 2019). Atop of this abundance of data, new market opportunities for businesses that collect, store, structure and elaborate the data rapidly grew: online databases, search engines, consulting firms, digital platforms, software management systems and many other examples of data–fuelled business models. This is a key transformation: where there is data, there will be AI. AI has the potential to spread into applications where data (i) is generated and can be collected in sufficient amounts, and (ii) its structuring and elaboration creates value–added for the business. These conditions shape the data reverse salient and expose the non–pervasive character of current AI.

Getting the data. First, in order to deploy AI to support any given application, an established and systematic process of data collection is required. In other words, the implementation of AI requires a meaningful representation of business processes (essential or not for a firm) in data — namely, their digitisation. This is why pioneering industries in AI adoption are the likes of Fintech and logistics, which are characterised by highly digitised and measurable processes and had forms of algorithmic automation and optimisation

already in place. The so–called 'Deep Learning revolution' stands precisely in the fact that it provided an effective tool to process raw unstructured data e.g. images, video, audio, making this activity cheaper (and thus economically viable) and less labour– and time consuming. Doing that, Deep Learning expanded the set of tasks that can be solved by AI algorithms. Deep Learning made possible to exploit troves of raw data that were already out there, waiting for an algorithm to harness them. An example is AI–based visual recognition, which emerged as a novel function applied to medical imaging records for diagnostics in many medical disciplines.

The existence of data does not automatically make the case for an AI application. Sometimes data might exist but its accessibility could be either hindered, inefficient or even welfare–damaging. This is partially due to unresolved data ownership and absence of mechanisms such as data markets to coordinate data supply and demand which would ensure the lawful and effective exchange of data ownership rights. An insightful summary of the situation with data markets is expressed in a quote of Edward Snowden: "there is no property less protected and yet no property more private than data" (Snowden, 2019). In some applications, data is a mere representation of an environment's state or processes (e.g. temperature control in data centres). However, when data is an imprint of activities conducted by actors, individuals or organisations that are external to owners of AITs, then data might be considered as a property of the actors that created it (Jones and Tonetti, 2020). Said differently, when data is a public good, ownership issues do not emerge, while the elaboration of data, which has the nature of a private good, requires solutions that address simultaneously consensual data transfer and privacy concerns (personal data that owners might either sell at a very high price or not to sell at all).

- In sum, the collection of data that reflects business processes including demand's feedback loops and establishment of data markets is a necessary though not sufficient prerequisite for AI deployment.

Monetising the data. Second, to persist being used as a useful technology within an economic activity, data elaboration performed by AI has to bring returns. The value of data elaboration can lie in harnessing otherwise

unmanageable amounts and complexity of data or (and) detecting patterns that humans cannot identify. Retrieving information about, for example, highly non–linear relations between a set of covariates and whether or not a person has clicked on an ad is undoubtedly a useful insight, but in order to systematically turn this information into a profit a firm has to build a sustainable business model to monetize on it. Monetisation strategies can vary across applications, which in turn are characterised by different pay-offs from the implementation of AITs. For example, for online retail, the monetisation strategy would involve the structuring of pricing and versioning of the offer given the association revealed by data elaboration. This strategy allows obtaining profit directly and from each offer independently. Differently, an AI algorithm that controls an industrial robot through the processing of sensory data and producing an adequate response in order to perform a routinized task creates value added that is more implicit and grows in a non–linear way with the scale of deployment of the technology.

- In sum, all kinds of data elaboration done by AI has to produce either valuable/unique intermediate result in the firm's production process or contribute to a valuable offer to the consumers, in both B2B and B2C markets, to ensure retention and generate profit.

Investing in assets. Third, sustaining the monetisation strategy requires investments into complementary assets of some kind. The costs of primary collection or acquisition of data from third parties (e.g. the purchase of database licences, cookies or data appends — see Bergemann and Bonatti (2019)), the storage within a firm or purchasing cloud space in order to further elaborate the data with AI, and even contracting micro–work to conduct data annotation (Tubaro et al., 2020), constitute yet another part of the data–related reverse salient. Thus, depending on the revenue from AI–based activity, a firm has to choose between investments into the development of AI systems at least in part in–house (including all domains — data, hardware and software) or into partnership with AI–provider along the AI value chain. The choice between the two alternatives is intertwined with the control aspect of AI and it shapes the distribution of market power among system builders in the nascent AI industry. Obviously, small and

medium–sized enterprises tilt toward outsourcing option to minimise costs. Moreover, even big companies for which AI performs not a core but a side function would be prone to purchase customised but ready–made AI solution in a package, benefiting from sharing the risks and legal responsibilities with the developer. Indeed, among AI–users the emergent strategy of 'join–and–share' AI–as–a–service solutions due to the high costs of every component of AI systems steers AI development towards a form of infrastructure, with the most powerful system builders (AI–producers) meticulously building and gathering pieces of the infrastructure together. The burden of high costs is coupled with cross–domain network effects. For example, depending on the application, the nature of data might vary — pixel matrix for images, text corpus for legal disputes, or panel data for consumer databases. This affects the choices and developments in the hardware domain (bandwidth capacity, memory size and placement, parallel or sequential processing and so on), programming framework (programming language, libraries) and algorithms themselves (loss function, optimization procedure). Together, the initial costs of implementation and cross–domain network effects increase switching costs of an alternative to any component and lead quickly to hard lock–ins for both supply and demand in the software and hardware domains. The result of this dynamics is a trend of over–specialisation in both domains, as we discussed in Section 4.3.2. Investments in more versatile and heterogeneous hardware and algorithms is a long–term strategy, but it has a longer period before returns start and is associated with uncertainty regarding adoption, making such innovation trajectories affordable only to a minority of (rather large) system builders.

- In sum, AI adopters make a choice on how to deploy AI–based solutions and invest in the respective complementary assets. This creates a demand–pull effect steering the innovative efforts of AI–producers further along existing technological trajectories. The opportunity costs in this situation might be substantial, as alternative trajectories are locked out by prohibitively high switching costs.

Eventually, two incentives reinforce each other: competition among AI–producers makes them sensitive to demand's (AI–users) needs, while demand is following the visibility of commercial value to sustain its strategies

to monetise on adopted AI solutions rather than the technical superiority of these solutions in the long term.[24] The task for policy–makers is to make sure that arguments related to high costs and strong network effects are not used as a justification to tilt the development of the infrastructure towards an inefficient realisation. In the case of AI, inefficient in technological sense might mean avoiding following the already–mentioned principle 'the bigger the better' in terms of ever–increasing size of data, amount of compute, complexity of algorithms, number of processors and so on for the sake of marginal improvement in performance (refer again to Figure 4.2, for instance). The fundamentally statistical nature of current AI will always strive for more data as a safe solution not only to achieve better representativeness of a given sample, but also to train deeper ANNs following the trend to increase the size of algorithms. In socio–economic sense, an inefficient instantiation would drain resources and resemble a skewed representation of stakeholders' interests — AI–users, AI–producers, society, public institutions — creating dead–weight losses, violating rights, damaging competition, and producing an asymmetric distribution of gains. R&D investment in scalable AI techniques like federated learning and neural network compression, and hardware technologies such as platform chips and edge computing can soften hard lock–ins and create ways out through compatibility with already existing components of the infrastructure.

Given the discussion above, we can outline a set of insights for policy–making: in order to cultivate technological opportunities to implement AI, policy attention can be directed to address the grey areas of data creation, collection and distribution. A way to do that is to assess how it has been done within the pioneer applications of AI. In particular, focusing on firms, this entails filling gaps such as developing the capabilities to digitise a firm's processes, organising their systemic and structured execution, and creating a digital twin of a firm's activity to be analysed with AITs. From the firms' perspective, the business models that monetise on AITs must be flexible to avoid being locked in solutions offered by dominant actors in monopolistic or oligopolistic markets. From the policy perspective, attention should focus on monitoring, detecting and regulating the whole *network* of AI–related mar-

[24] This dynamics has been observed for Machine tools industry by Rosenberg (1963) and for ICTs by Bresnahan (2019b).

kets, to ensure the conditions for fair competition among system builders, and to lower the cost of exploration and support of alternative technological solutions and partnerships. This would nurture an ecosystem of actors and technologies contributing to the transition to a more distributed mode of control over the AI LTS.

Overall, if AI is an LTS then policy design should be inspired by the priorities set by the LTS framework. Examples of these priorities are: (i) the balanced construction of the system, for example by supporting the development of AI talent, identifying and suggesting new components for the system based on relatedness, providing resources and facilities for experimentation; (ii) curbing the monopolisation of resources in the hand of a few actors across the fundamental domains of AI ensuring equal access for all system builders; (iii) pushing for inclusive or public models of governance by pursuing the identification of technical and non–technical standards.

4.5 Conclusion

Artificial Intelligence is generally considered a breakthrough that is technologically revolutionary and, often, also philosophically existential; it is capable of reshaping societies and economies while at the same time offering a mirror to look inside ourselves and our human intelligence. Adapting an expression of Norbert Wiener, if AI is ready to 'usurp specifically human functions', then our priority should be to understand which is the 'purpose put into the machine' to avoid unintended consequences that can result from delegating tasks to AI.

Indeed, AI has the potential to influence many real world processes. But the very nature of current AI is less romantic than what is usually depicted, even though its impact can still be transformative. Much of the hype around AI is a case of what Braitenberg (1986) called 'the law of uphill analysis and downhill intervention', according to which humans tend to overestimate the complexity of a mechanism, guessing its internal structure from observation. The human–level performance displayed by AI on certain tasks can indeed hide the 'Clever Hans' nature of these technologies. In fact, current

AITs are essentially a new wave of ICT technologies. At the moment, they are usually brittle and function–specific algorithms that are used as 'prediction machines' because they are effective in identifying associations and extracting patterns.

In this Chapter, we focused on the essential mechanisms of AI and offered a novel perspective on its nature — a key exercise, as AITs are increasingly embedded into products and service, commercialised and used in a wide range of applications. In particular, we tested in details the consensus idea that AI is a general purpose technology by evaluating how GPT definitional characteristics fit the features of AI. Our conclusion is that it is premature to consider AI a GPT. This is not because AI is a technology just emerging, and thus *not yet* a GPT, but instead because the GPT 'suit' is structurally inappropriate — and namely too flat — to dress AI. AI is not a stand–alone technology as GPTs are, but a system technology that displays infrastructural properties: it has a dual nature, as a technological artefact and at the same as a socio–technical network.

AI shares some features with GPTs (for example innovational complementarities and technological dynamism), but these have a qualitatively different nature in the AI case. The very differences of AI from the GPT benchmark are what carries useful information. For example, we establish the stylised fact that, at different levels of analysis, AI is not pervasive in a GPT sense: it has many uses, but it is not widely used in the majority of economic activities — it is not as ubiquitous as computers are. Even in the few industries in which it is adopted, diffusion is concentrated in and driven by a few large lead actors. Similarly to the Internet, AI provides an additional layer of functionalities to end–users, and for this reason it spread large in user base and territorial coverage when embedded in final products and services; besides that, AITs leave the rest of the economy almost untouched.

Being a GPT requires penetration in scope and scale, and AI might never reach that status, remaining confined in activities where it is useful the most. Transformation occurs at a deeper level: the level of systems, where AI is implemented as an additional layer in a system to become an autonomous and capable (and, hence, perfect) tool of cybernetic control. In this way AI

becomes an infrastructural technology: superimposed on existing technological layers, and, through that, interwoven into the economic structure.

If GPT is a misspecified model of AI, is there a better model around capable to capture the nature of an infrastructure technology? Our bet is on the concept of Large Technical System and the framework it offers to describe system technologies. In the analysis, we mapped the LTS building blocks on AI. This allowed us to identify AI's system builders, reverse salients, momentum and load factor, as well as to derive useful insights on AI phases of development, boundaries, control, technological style, and goal orientation. Furthermore, we applied the AI–as–LTS scheme to the issue of policy and strategy and showed that the rationales for intervention and the types of policy actions differ substantially between the AI–as–GPT and AI–as–LTS interpretation. As an example, we focused in details on the reverse salient related to data to illustrate the many tensions emerging within this domain of AI.

With our study, we contribute to the nascent Economics of AI and, more generally, to that part of the Economics of technological change and innovation interested in uncovering the structure of technological breakthroughs. Using the LTS framework, we extend the reach of economic analysis of AI to the neighbouring fields of sociology of technology and science and technology studies. As a result, we were able to offer a novel understanding of infrastructural and system technologies through the case study of AI.

As a new fully–fledged industry is rapidly forming around AI, a correct mapping of the technology and its complex nature is necessary to avoid misunderstanding its trajectory, misallocating resources dedicated to its progress, and harmful developments. Understanding AI means understanding its fundamental fabric and design principles: how a system technology is engineered by different actors in a dynamic 'workspace', which forces shape its path of development, and how these same forces can be steered in a direction that contributes to the common good.

Chapter 5

On the Basis of Brain: Neural–Network–Inspired Changes in General Purpose Chips

"This ability of a single box to carry out any process that you can imagine is called universality, a concept first introduced by Alan Turing in 1936. Universality means that we do not need separate machines for arithmetic, machine translation, chess, speech understanding, or animation: one machine does it all."

— Stuart Russell in Human Compatible: Artificial Intelligence and the Problem of Control (2019)

5.1 Introduction

The semiconductor industry has been an upstream supplier of computing devices to a wide range of market segments and during its history has faced various crises. Despite being not new to hurdles, the industry is now facing a

novel, fundamental challenge: chipmakers are exploring new ways of organising computation on a chip to respond to recent breakthroughs in Artificial Intelligence (AI); AI creates a demand for computing devices directly and indirectly induces other markets that adopt AI solutions to demand changes in chips as well. A profound discrepancy resides in the mismatch between the nature of modern AI algorithms and the organisation logic of conventional hardware. Over the decades, and since its establishment as a solid field in the 1950s (McCarthy et al., 1955), AI has been developing mostly as a scientific experiment with its own successes and failures rather than a commercial technology with large potential. For this reason, AI had a small weight as an application segment for the semiconductor industry, and the discrepancy recently exposed appeared at such large scale for the first time. Nevertheless, the status of AI is changing, AI managed to gain traction and now is being experimented with in numerous markets (for example, (Agrawal et al., 2019c)) so that its developers are winning the so–called 'hardware lottery' (Hooker, 2020); there occurred a swarm of new chips that embody alternative architectures capable of executing AI algorithms. Such exogenous 'shock' exerts pressure over the established technological trajectory and is poised to introduce changes in the semiconductor industry. In this Chapter, we analyse this technological discontinuity and how it layers up on the mechanisms and forces at work in the semiconductor industry. On the basis of this analysis we address the question of which product configuration might characterise the next phase in the semiconductor industry life cycle as a result of this shock.

Our study contributes to a number of literature strands. A first one is the nascent economics of AI (Agrawal et al., 2019b), as we analyse the impact of AI on product design and innovation–related decision making in a particular industry. A second contribution is to the literature on the economics of technological change, industrial dynamics, and systems of innovations, as we study the forces that support and contest the technological trajectory (Dosi, 1982) of chip production (Steinmueller, 1992) and the factors driving the evolution of the semiconductor industry (Malerba et al., 2008; Brown and Linden, 2011; Adams et al., 2013). A third domain we build upon is the research on platform products (Baldwin and Clark, 2000), in particular that focused on the computer industry (Bresnahan and Greenstein, 1999) and on

the strategic management of semiconductor firms (Burgelman, 2002; Gawer and Henderson, 2007). A fourth strand we contribute to is the economics of network products and software as a supporting service (Church and Gandal, 1992; Chou and Shy, 1993). We build our line of argument drawing from the AI and computer science literature (Russell, 2019; Hooker, 2020) as well as from that on computation theory and integrated circuits design (Borkar and Chien, 2011).

To expose the discrepancy currently forming between capabilities of chips and their required performance, in Section 5.2 we put together alternative ways of organising computation in a program and the corresponding logic of hardware. This creates a framework that allows understanding the hardware and software domains and their interrelation and helps to highlight the radically different nature of Artificial Neural Networks (ANNs). In Section 5.3, we proceed with an overview of established and novel chip architectures and highlight their strengths and disadvantages in application to different tasks. Comparing different architectures, the important characteristics of a chip's performance become evident: (i) processing speed, (ii) flexibility, and (iii) energy efficiency. Together, these characteristics form a trilateral technological frontier that serves as a benchmark for a chip's performance and guide design decisions. We briefly discuss several directions that chip producers can act upon by introducing improvements in chips design, and the trade–offs that might occur. We conclude that the AI shock at the moment induced two kinds of innovation efforts: (i) the design of a novel processor architecture for the needs of modern AI (especially ANNs), and (ii) the integration of this processor inside a computing system. Section 5.4 rationalises the unfolding situation accounting for the technological and economic factors that affect product development in the semiconductor industry. First, in Section 5.4.1 we introduce a stylised model of demand distribution based on the elasticity of demand with respect to hardware's flexibility (approximated with the variety of supported software) and processing speed and energy efficiency combined. Building on the analysis conducted in the previous parts, Section 5.4.2 outlines two scenarios for the evolution of chips and provides some arguments in support of each of them. In Section 5.5, we place our analysis in context by discussing how the forces and tensions we unpacked in our study align with (or differ from) those identified in related

literature. Finally, Section 5.6 concludes.

5.2 The Computation Framework for Neural Networks

We perform a continuum of tasks with the help of computers. In the words of Baldwin and Clark (2000), "Computers are fascinating, interesting, and delightful to human beings because they are complex. Most of us are not especially intrigued by their raw speed or low cost. It is the many things computers do, and the many different ways they can be configured, that makes them interesting and useful. And it is the ability of computers to fulfill idiosyncratic, even whimsical desires [...] that causes these artifacts to surprise and delight us." In less than a century, computers have gotten firmly entwined with our lives, and computing became an ubiquitous activity. Any program that performs a task has an algorithm that in a structured manner leads to the achievement of a goal. In general, any program is a *virtual machine* that is ran on a physical machine — a computer. Basically, what computers facilitated people to do is the translation of regular tasks and activities into algorithms. Thus, if the performance of a task is a problem, an algorithm is its solution, regardless of the nature of a task — being it writing a document, 3D–modelling or calculating a celestial trajectory. There exist many ways of performing a task, and so do many algorithms. As a solution for a task, an algorithm can be characterized by the level of efficiency with which it achieves the goal. A first intuition would suggest time and probably memory use as inputs that an algorithm needs to deliver the result. However, to get a measure of the efficiency "it is necessary to have at hand a method of measuring the complexity of calculating devices..." (McCarthy et al., 1955, p.2). In other words, the efficiency of a task's solution should be assessed based on joint performance of an algorithm (software) and the device on which the computation occurs (hardware); the design of hardware can take over part of task's complexity so that algorithm remains simple or vice versa. The efficiency issue applies to any algorithm–device tandem and its importance grows together with complexity of a task. This fundamental complementarity between the hardware and software domains

is key to understanding the impact that Artificial Intelligence Technologies (AITs) can have on chips.

Programming Paradigms. Algorithms can approach a task in different manners, called programming paradigms. A paradigm conveys the organisation logic of computations and their execution. There are many programming paradigms — probabilistic, event–driven, automata–based, etc. — but in our analysis we employ two of them as they represent higher–level abstraction approaches to achieving a given task's goal. The first one is the *imperative* or *procedural* programming paradigm, that is concerned with the control over the flow of algorithmic instructions that lead to a desirable outcome. Thereby, an imperative algorithm is an explicit algorithm. The second programming paradigm is *declarative*, that specifies the desirable outcome but not the procedure that leads to it; hence, the algorithm can be implicit. The two approaches exhibit different level of efficiency when applied to different tasks. To illustrate this statement, we consider two examples: the first one is a simple arithmetic task of the kind 'get 8 using only 2s and basic arithmetic operations'; the second is a task of object detection in an image.

In the first task, when the arithmetic rules are well–defined the correct solution can be obtained easily with an explicit imperative algorithm. Now let's imagine that the arithmetic rules are unknown and hence an explicit algorithm as well. Thus, a program can, for example, add before multiplying. In this case there are multiple answers (and the more numbers involved, the more answers are possible). A declarative approach to the task by setting a specific number as an answer would deem other answers incorrect and hence narrow down the set of solutions (i.e. algorithms) to the ones that lead to the correct answer. This approach won't necessarily infer arithmetic rules but can approximate them. Obviously, for this task the imperative approach is much more efficient than the declarative one, as it provides a unique and correct answer in explicit steps.

Now consider a problem of object detection in an image in the context of autonomous driving. To classify an object, for example, as a pedestrian, it is necessary to identify a minimum set of features that characterises it, codify

these features and their variation, and write an algorithm that evaluates the correspondence and 'decides' upon classification. In the simplest case when one feature would unilaterally identify one object, the minimum number of features to pre–program would be equal to number of objects that must be classified. Sometimes, classification can be reduced to the effective minimum of categories to distinguish by making the categories broader, for example, living creatures, mobile non–living obstacles, immobile non–living obstacles. However, the broader the category the larger the variance within a particular feature; if the feature used to classify an object as a living creature is 'presence of a head', the variety of heads' shapes, sizes and textures must be accounted to avoid misclassification into other categories. Depending on the task, the number of objects and their features can vary: more fine–grained classification is required for a high stake loss function (Russell, 2019) such as in autonomous driving. As the number of objects or/and features grows, the task of object detection quickly becomes impractical or even intractable to approach with an explicit, imperative algorithm. In contrast, the declarative approach that allows for implicit algorithms can handle this problem much better as it doesn't need to specify features and their correspondence to objects; instead, it can check whether or not the classification of an object is correct.

The comparison of the two programming paradigms on these example tasks shows two important aspects: (i) efficiency varies between approaches depending on the task to be executed, and (ii) the construction of an explicit algorithm requires some degree of certainty[1] that decreases with the complexity of a task. This is what concerns the algorithms' part of efficiency and overall computability. As pointed out earlier, the way in which computation is organized is fundamentally bound to the design of the computing hardware, and the two have implications for one another.

Models of Computation. The efficiency of a given computing technique should be estimated in connection with the device that performs it. Therefore, it is necessary to consider how the structure of a given physical device has been designed to optimise the joint performance of the device and

[1]Here the notion of certainty refers to the size of search space in terms of the number of (i) laws or rules that a task is subject to and (ii) objects that matter for a task.

the number of virtual machines executed on it. The theoretical concept of model of computation precedes the physical implementation of a computing device. In the theory of computation, a *model of computation* is the conceptual framework that describes how the result of an algorithm is computed given the available components of a computing device and their possible interactions. Not surprisingly, there are several models of computation implemented in hardware.

The first and dominant model implemented in the vast majority of computing devices is the *sequential model of computation*, initially proposed by Alan Turing and named after him as Turing machine. Turing's automatic machine performs computations by scanning one symbol per unit of time from an infinite tape and applying one of its finite configurations (operations) (Turing, 1937, p.231). This organisation of computation mirrors the imperative programming paradigm: control flow programs are sequences of machine instructions with tags indicating which data is needed to perform the respective instruction in a sequence. On the one hand, sequential execution allows for an immense flexibility of manipulations over data, making feasible the performance of complex algorithms, a property which Turing called *universality*. On the other hand, performing a highly complex algorithm in a sequential manner might lead to an impractically long time of execution.

The physical architecture of a computing machine corresponding to the sequential model of computation is so–called *von Neumann architecture*. Due to the property of universality the von Neumann architecture reproduces, this architecture implemented in a processor proved to be fit for the execution of vast amount of virtual machines, allowing to address a large set of tasks where the control flow logic of an algorithm is capable of achieving the goal. Put simply, explicit algorithms with stepwise instructions resemble the way humans reason, which served as inspiration for early computers. During the following decades, due to the positive reinforcement loop in optimising the design of hardware and software, the set of tasks performed on the sequential model of computation kept growing and chips with the von Neumann architecture at the core gained a foothold as the dominant design (Suárez and Utterback, 1995). The development of computers allowed

applying them to increasingly complex tasks, pushing the frontiers of chips performance to keep up with speed, memory, energy efficiency and computability requirements. In the same way with programming paradigms we discussed earlier on, an identical problem can be solved on different models of computation with different efficiency up to the extreme case when one model of computation cannot ensure that an algorithm will convergence to the answer. When an algorithm is implicit and hardly can be expressed in the form of instructions flow sequentially changing the program's state, the efficiency of computing such algorithm on a Turing machine can decrease until it almost disappears. In this case, another model of computation can be more appropriate.

Concurrent models of computation as alternative to sequential models of computation do not focus on the order of instructions; instead, the focus is shifted to other properties of algorithm execution such as timing, parallelism or concurrency (Lee and Neuendorffer, 2005). This class of computation models is a good candidate for tasks where the algorithm is not a linear sequence of instructions but a more distributed one, for example, various instantiations of embedded software[2]. The problem with the concurrent class of models of computation is that it does not have a universal abstraction, a sort of common denominator for this class, unlike the von Neumann architecture for control flow, sequential class of models. Software that implements the concurrent computation model is an ad hoc solution for a specific hardware as opposed to the prevailing general purpose, imperative software that can be installed on any machine. This implies that chip design for the concurrent model of computation supports lower universality (heterogeneity of tasks it can execute), and initial attempts to design such circuits can be tailor–made to a specific family of algorithms and vice versa. To design and manufacture a circuit entails high costs; hence, to return the investments there should be demand from the application markets. Thus, the start of the development process of new circuits that implement the concurrent computation model depends on (i) the technical feasibility of a

[2] "Abstractions that can be used include the event–based model of Java Beans, semaphores based on Dijkstra's P/V systems [29], guarded communication [30], rendezvous, synchronous message passing, active messages [31], asynchronous message passing, streams (as in Kahn process networks[32]), dataflow (commonly used in signal and image processing), synchronous/reactive systems [6], Linda [33], and many others." (Lee, 2002)

common abstraction, (ii) the size or/and number of markets that benefit from such hardware. For a long time concurrent models remained at the fringe of programming and the semiconductor industry, serving specialised niches like avionics and the automotive industry or functions scattered across different industries like signal processing or system modelling.

Neural Networks and AI. Everything changed when Artificial Neural Networks (ANNs) re–entered the toolkit of AI techniques.[3] Representing the dataflow programming paradigm (a subclass of declarative programming) and the eponymous model of computation (a subclass of concurrent models of computation), ANN became a revolution as it is the first program[4] that can operate as embedded software as well as conventional application software while having many distinct uses. In terms of algorithm organisation, ANNs differ significantly from classical programs. An ANN is a multi–layered directed graph. Every layer consists of nodes — instructions represented by some operations over data such as arithmetic functions, e.g. multiply–sum. Connections between nodes in different layers are dependencies between the respective instructions: every possible path in a network is, in a sense, a sequence of instructions. This logic of organising and executing computations describes a dataflow programming paradigm, where the flow of data defines which instructions to perform; when the data required for the execution of an instruction is ready, this instruction can be initiated without waiting for other independent instructions. Differently from the control flow logic realised in von Neumann architectures, where data is stable and a sequence of instructions is applied to it, in dataflow computation models instructions are stable and data floats among the instructions. There are several implications for circuit design that can be derived from this description that we discuss in the next Section.

Being inherently parallel and distributed, ANNs represent implicit algo-

[3]The birth of the connectionist approach to AI centered around ANNs dates back to the 1950s, with ground work of McCulloch and Pitts on neuron–like structures capable of calculations (McCulloch and Pitts, 1943) and Hebb's theory of cell–assembly formation (Hebb, 2005).

[4]We refer to a *program* as a *virtual machine*. Boden (2016, p.4) defines virtual machine as "the *information–processing system* that the programmer has in mind when writing a program". Thus, in this Chapter we use the term *program* in a broad sense to keep the text simple.

rithms where the initial network is a template and Deep Learning (DL) is the tool to establish the ANN's structure; for traditional programs, the algorithm is a sequence of instructions while for ANNs it is a network's structure of connections. The nature of ANNs perfectly fits into the declarative paradigm given the strong goal orientation and the absence of an explicit order of instructions. As already discussed, a potentially large number of tasks is hard to approach with explicit algorithms either because these algorithms are yet unclear or even if known they can be extremely inefficient solutions; the booming number of ANNs' applications confirms this statement, showing the potential of implicit algorithms. It is worth highlighting that the value of ANNs is twofold: for some tasks it increases processing speed resulting in a tremendous reduction in execution time, while for other tasks this is the only computable algorithm that can deliver result. Parallelism is necessarily present in all ANNs solutions primary as a requirement for obtaining the result rather than as an advantage in processing speed.

In general, AI is endemic to the declarative programming paradigm, with strong goal orientation and implicit, exploratory algorithms to achieve it. For example, two important instances of AI algorithms, the Logic Theorist (Newell and Simon, 1956) and ANNs both belong to the declarative paradigm despite representing two distinct approaches to AI — symbolic and connectionist respectively. The difference between the two resides in the strategy used to reduce the search space of options to converge to a goal: Logic Theorist used the rules of propositional logic to cut off irrelevant steps and navigate the convergence towards its goal — the proof of a theorem; ANNs instead use purely data–driven optimisation of a loss function. In both cases, the algorithms are exploratory on the side of *how* to achieve the specified goal. However, the guiding tool of the convergence path for Logic Theorist is logic, a formalisation of explicit reasoning, hence the inference that Logic Theorist emulated is also explicit. Indeed, the program was an attempt to prove theorems whose proof have been previously found through human explicit reasoning. Using logic as a guiding tool has its advantages, but the main problem is that "[l]ogic requires certainly, and the real world simply doesn't provide it" (Russell, 2019, p.40).

Taking stock of the discussion so far, given their many uses, ANNs have the

potential to draw enough attention to the concurrent class of computational models and consequently to trigger and accelerate the development of its physical implementations. The increasing availability of data contributes to the growing applicability of modern AI. As this viable alternative to traditional programs gains traction, so does the exploration of the economic activities and new business models employing or centered around AITs. One particular transformation that is the focus of this Chapter is rejuvenation of the semiconductor industry technological opportunities with the arrival of ANNs. This transformation starts not only with the challenge of addressing the technical properties of ANNs into hardware but also with the need for the industry to engage in technological and strategic foresight estimating the costs and benefits of different product configurations and industry scenarios.

5.3 Computational Models Shaping Hardware Architectures

As discussed in the previous Section, the architecture of a circuit is fundamentally linked to the chosen model of computation. Despite the fact that several models of computation existed on paper, the sequential one implemented in von Neumann architecture prevailed due to its universality (flexibility) and correspondence to the imperative programming paradigm. The dominance was preserved through eight technical and economic crises that forced the semiconductor industry to come up with and implement both incremental and radical innovations (Brown and Linden, 2011). The slowing down of Moore's law as the main roadmap for the industry (Flamm, 2019), as well as rising costs of design and fabrication have already influenced the industry in the past but now seem to come back. Atop of these recurring crises, the novel dataflow architecture is on the rise due to breakthroughs in AITs and threatens to fork the established technological trajectory with von Neumann architecture at the core. Instead of catering the needs of instructions flow mainly concerned with the speed up of computation, the emphasis in dataflow architectures shifts to the energy–proportional and agile data routing. By design, the two architectures have inherent advantages and disadvantages which we shall discuss in the next paragraphs where we com-

pare types of processors that implement these architectures. The first two types — scalar and vector processors — are earlier products that represent the sequential model of computation while the last two — array and neuromorphic processors — are recent implementations of the dataflow model of computation.

5.3.1 An Overview of Architectures' Variety

Scalar Processors. This type of processors performs one instruction over a scalar per one clock cycle. It calls the data one scalar at a time to supply it for an instruction; then results are recorded into memory after every instruction. This architecture is realised in Central Processing Units (CPUs) and represents a physical implementation of the Turing machine with both advantages and limitations. Communication with memory for every instruction allows for the realisation of Turing's universality principle: having heterogeneous instructions (divide, multiply, AND, OR, etc.) in a sequence. However, the same feature creates the so–called von Neumann bottleneck: the transfer of data back and forth from memory for every instruction slows down processing speed (the movement along the instructions' sequence), depends on the bandwidth of the connecting channel, and significantly contributes to the energy consumption of a chip. The true concurrency or parallelism is not implemented in this architecture and can be only simulated through pipelining, a technique that allows performing concurrently a small number of instructions by processing them in a cascade (so–called instruction level parallelism).

Vector Processors. The idea of realising parallelism in computation in order to increase computing power was, however, already around since the 1970s. For example, the products of Cray Research exploited the so–called *vector processors*. A vector processor consists of a large number of cores that are simpler than the few but more complex cores of a scalar processor. A single instruction uses a vector as a unit operand (a batch of data): an instruction fetches a vector from memory and assigns each vector component (scalar) to one of numerous cores to execute this one instruction in parallel

(data level parallelism).[5] Today's Graphics Processing Units (GPUs) embody the same principle. Until mid–2000s vector processors haven't been widespread elsewhere except in supercomputers performing complex computations with large arrays of data, and later in 1990s, with the rise of computer games for the purpose of graphical rendering. GPUs consists of hundreds and even thousands of cores, however less complex and independent than CPU cores. Clearly, the coordinated work of a larger number of cores demands a higher energy consumption and the reorganisation of the computation process according to a specific programming logic; it has to deliver a low idle rate, otherwise the usage of so many cores is not justified (see Amdahl law[6]). Technically, vector processors are still von Neumann machines replicating the same principle for each of its numerous cores: instructions in a sequence can be different but this requires communication with memory for every one of them. This gives GPU the same property of universality (although significantly less than CPU) but the von Neumann bottleneck problem as well. For this reason, GPUs are very well–suited for massively parallel repetitive computations.

The conventional use of GPUs was as a discrete device on a motherboard for graphical rendering in computer games; however, with the rise of ANNs chipmakers started the integration process of GPUs into a chip's system in order to exploit GPUs as a new functional module for a broader set of calculations. The set of functional modules that include CPU, co–processor(s) like GPU, memory system, input and output units placed on a single silicon substrate constitutes a so–called System–on–a–Chip (SoC). The integration of GPUs into SoC opened up a potential to effectively include GPUs in computing processes for more general calculations, rather than just operating with graphical data. Using GPUs for general calculations was dubbed General–Purpose Computing on Graphics Processing Units (GPGPU). Bundled together in a SoC, the CPU performs orchestrating work while the GPU

[5] As an example, the multiplication of two vectors of length 20 in a scalar processor occupies roughly twenty instructions to be executed in a pipelined manner while in vector processor the same multiplication executed in parallel manner takes one instruction.

[6] Amdahl's law describes potential speedup in performance as a function of number of PUs involved (Amdahl, 1967). The exploitation of larger number of PUs represents an "enhanced" or "faster" mode of execution. Thus, said differently,"Amdahl's Law states that the performance improvement to be gained from using some faster mode of execution is limited by the fraction of time the faster mode can be used"(Hennessy and Patterson, 2011, p.39).

executes massively parallel calculations. The main principle of CPU–GPU co–working can be generically described as follows: the CPU dispatches an instruction and points at the related data to the GPU, and the GPU distributes the data across its cores in order to then perform calculations over the data in every core at the same time. The way of computing involving processors with different architectures is called *heterogeneous computing.*

Competing producers developed their own frameworks[7] allowing for heterogeneous computing. Examples are Nvidia's Compute Unified Device Architecture (CUDA) and Fusion System Architecture (FSA), started by AMD, that later transformed in Heterogeneous System Architecture (HSA) by the HSA Foundation, a consortium of companies including AMD, ARM, Qualcomm, Samsung, etc. In 2007, when Nvidia launched CUDA for GPGPU, it was not received with much enthusiasm, certainly not enough to immediately induce a shift from conventional CPU programming to GPU programming. However, Raina et al. (2009) demonstrated the potential of GPU exploitation for ANNs' applications. They showed a method of involving GPU hardware into the training of an ANN which surpasses CPU performance in terms of time by a factor from 12 to 72 depending on the ANN's complexity. The joint success of ANNs and GPUs arrived in 2012 from the field of image recognition through the ImageNet Large Scale Visual Recognition Competition. The ANN AlexNet reached on average a 15.3% error rate in classifying 1.2 million images into 1000 categories (Krizhevsky et al., 2012). Since then, ANNs achieved substantial progress, performing some function above human level capabilities (see Eckersley et al. (2017)). Overall, the use of diverse learning techniques with ANNs opens up a path to the affordable automation of non–routine tasks or improves performance in already automated, routinized ones. Being tightly interconnected, the hardware domain has to respond to accommodate and effectively support this breakthrough in software domain that has multiple and heterogeneous applications.

In sum, comparing CPUs and GPUs, GPUs compute in parallel but are more fit for massive, regular, less sophisticated computations, consume more energy per calculation and are less conventional to program in comparison

[7]The notion 'framework' refers to a complex of hardware, programming language and instruction set library.

with CPUs.

Array Processors. With the advent of ANNs, GPUs acquired a number of new markets; GPUs were available at the time of ANNs' development and contributed to the breakthrough itself. It didn't take long before chip development went further. The first serious challenge for Nvidia's GPU came from Google's Tensor Processing Unit (TPU) in 2016. Google's TPU is a *matrix* (or *array*) processor which removes the von Neumann bottleneck from its cores by creating a systolic array of Data Processing Units (DPUs). A systolic array of DPUs represents a hard–wired network of homogeneous calculating units — meaning that every DPU implements the same set of operations. Once data is uploaded from the memory it travels among DPUs and is processed upon arrival within a DPU without being recorded intermediately into memory. Thus, the TPU emulates the dataflow architecture[8] introduced in Section 5.2. This makes TPUs faster in processing than CPUs and GPUs, performing hundreds of thousands of operations per clock cycle, but allows the execution of only regular instructions such as multiply–accumulate in ANNs (Jouppi et al., 2017).

A TPU is more energy efficient and faster in processing due to its dataflow architecture but it has the shortcoming of being less flexible or universal in computation. Furthermore, data–level parallelism realised in both GPUs and TPUs requires the representation of information in regular form of vector, matrix or array in order to effectively run programs (or parts of them) on this hardware. Google improves its TPUs continuously, issuing a new generation twice as powerful as the previous one roughly every year. However, the company has also refrained from the commercial sale of TPUs, using them only in internal services and providing access to customers through the cloud. In contrast, in early 2019, Intel unveiled Intel Nervana Neural Network Processor (NNP) containing both CPU and tensor cores.[9] More products from competitors followed: Eyeriss 2.0 jointly designed by MIT and Amazon (Chen et al., 2019), Hanguang 800 from Alibaba Group, Infer-

[8] Dataflow architecture is inherently parallel but not all parallel systems belong to the class of dataflow machines (Veen, 1986).

[9] Nervana NNP is a modified 10^{th} generation Intel Core processor (CPU) with, among other changes, replacement of GPU cores by tensor cores.

entia and Trainium by Amazon and many others[10]. Lastly, Huawei's Ascend 910 chip comprises all three types of processors — scalar, vector and matrix — as its functional cores within one SoC.

Neuromorphic Processors. Neuromorphic chips represent a radically different development direction of computing devices. This novel architecture mimics the impulse–based synaptic activity between neurons in the brain. Transistors wired together emulate a network of neurons with electrical synaptic connections. Information is encoded as analog signal and flows across an array of simulated neurons in form of electrical pulses. In effect, the neuromorphic architecture is an analog version of the dataflow architecture of an array processor. Examples of neuromorphic chips are TrueNorth, a joint venture of IBM and DARPA under SyNAPSE program (Merolla et al., 2014), the experimental neuromorphic chip Loihi from Intel (Davies et al., 2018) and the hybrid (CPU and network of neurons) chip Akida from BrainChip. The distinctive feature of such chips is extreme energy efficiency, which makes the neuromorphic architecture a promising competitor.

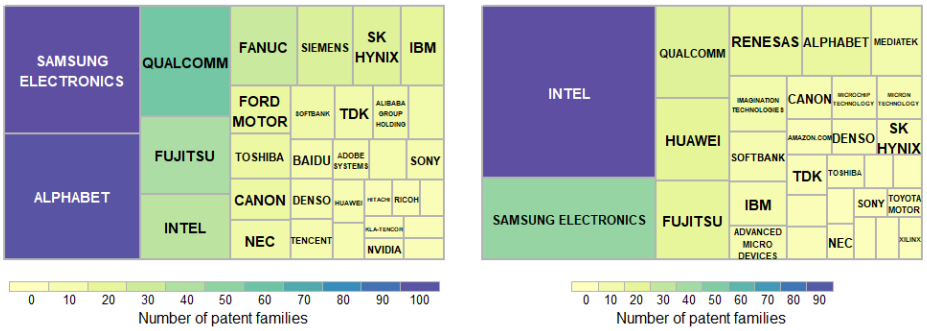
In sum, scalar processors have nearly exhausted their technological opportunities and have lost their exclusive position in computation. Nevertheless, they remain an inalienable component of a computing system. In turn, computing system are experiencing an upgrade through experimenting with novel architectures for co–processor on a SoC. At the moment, specialised processors alone do not deliver end–to–end solutions; they lack generality or flexibility and/or commonly supported and well–developed frameworks to be used for general purpose programming. CUDA for GPUs or TensorFlow for TPUs are examples of such frameworks, but they are immensely smaller than the encompassing and versatile framework created over the decades for CPUs. Overall, chips with different underlying architectures — controlflow sequential and dataflow concurrent — do not seem to be competing technologies (Arthur, 1989) but rather complements for a SoC (Baldwin and Clark, 2000). The development of a new component, its subsequent integration in and reorganisation of the computation process induces architectural innovations[11] (Henderson and Clark, 1990) in a SoC. Examples of

[10] https://github.com/basicmi/AI-Chip

[11] Here we refer to the term *architectural innovation* suggested by Henderson and Clark (1990), that denotes a specific type of innovation when the core function behind a tech-

architectural innovations include techniques of interconnection among processing units (PUs) and memory (e.g. Nvidia's NVlink network–on–a–chip and hierarchical mesh of MIT's Eyeriss) (Borkar and Chien, 2011; Winter et al., 2010; Chen et al., 2019) as well as packaging techniques of chiplets (e.g. Intel's EMIB used in Nervana NNP, TSMC's CoWoS used in Nvidia's GPUs, Swarm communication fabric from Cerebras) (Shao et al., 2019; Lie, 2019). Overall, the direction of innovation efforts seem to "expand beyond the processor core, into the whole platform on a chip, optimising the cores as well as the network and other subsystems" (Borkar and Chien, 2011, p.75)

Currently, the efforts of the semiconductor industry are directed at two targets: (i) to design a novel processor architecture for the needs of AI which emerges from the class of concurrent computational models, and (ii) to integrate this processor onto a SoC. Given that (i) is in development, the first SoC resulting from (ii), though capable of supporting AI applications, are still early and expensive versions whose components yet fall short with respect to the performance characteristics that we discuss in the next Section 5.4.



(a) Dataflow **(b)** Controlflow

Figure 5.1: Top–30 global holders of patent families on chip's architectures 2014–2016
Data: COR&DIP©v.2 IPC classes: (a) G06N 3/02-10, (b) G06F 15/76-82

The efforts in the development of new processors for the needs of AI are already visible in the IPR system. Figure 5.1 shows the patenting activity of top–30 proprietors by type of architecture over the period 2014–2016. The number of patent families filed globally on dataflow architectures (left

nology is preserved while the components and linkages among them undergo changes.

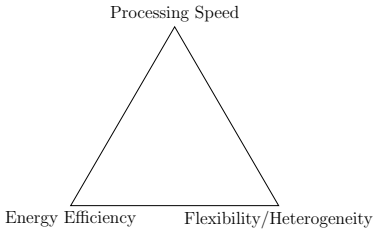
panel of Figure 5.1) overcame controlflow architectures (Turing machines; right panel of Figure 5.1) with a total of 786 patent families by 115 companies versus 405 from 77 firms respectively. Not surprisingly, in both categories the top positions of the ranking are occupied by large international companies such as Samsung, Alphabet, Intel, Qualcomm and Fujitsu, however with a long tail of smaller companies especially in the case of the dataflow architectures. The intensive exploration of the rich technological opportunities for processors embedding dataflow architectures by both incumbents and startup companies creates a swarm of novel and distinct products. For instance, left panel of Figure 5.1 includes *photonic* chips that use light to encode and transmit information instead of electricity (i.e. IPC class G06N 3/067) such as one of the programmable nanophotonic processor from Lightelligence (Shen et al., 2017).

However, there are already early signs of search for a more general or flexible dataflow architecture that might lead to a shake–out in product variety and to the emergence of a dominant design in this class of processors. Sze et al. (2020) discuss in detail the challenges and criteria for the design of Deep ANN (DNN) processors, claiming that "it is increasingly important that DNN processors support a wide range of DNN models and tasks. We can define *support* in two tiers. The first tier requires only that the hardware needs to be able to functionally support different DNN models (i.e., the DNN model can run on the hardware). The second tier requires that the hardware also maintain efficiency (i.e., high throughput and energy efficiency) across different DNN models." In sum, this statement calls for higher "flexibility to cater to a wide and rapidly changing range of workloads" along with speed and energy efficiency, navigating innovation efforts in architectures' design. Moreover, while in general welcoming novel architectures to help advancing AI, Hooker (2020) goes even further in her discussion raising concerns about the premature and costly specialization of novel hardware on ANNs. The dynamic nature of the software domain manifests itself brightly in such an experimenting field as AI. Indeed, "[i]t is an ongoing, open debate within the machine learning community about how much future algorithms will differ from models like deep neural networks"; however, "[h]ardware design has prioritized delivering on commercial use cases, while built–in flexibility to accommodate the next generation of research ideas remains a distant

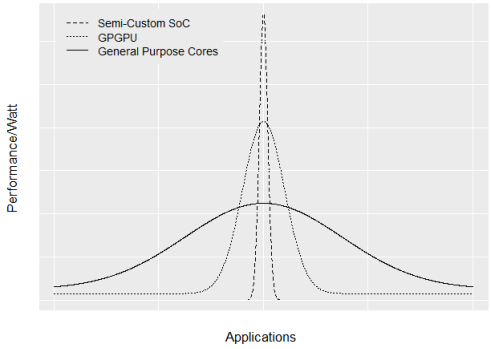
secondary consideration" (Hooker, 2020, p.7).

5.3.2 The Trilateral Technological Frontier

The two architectures we analysed mirror each other in their strong sides and disadvantages: the controlflow architecture is concerned with speed of performance in first place and consumes most of its energy on data movement from and to memory, while the dataflow architecture suffers from lack of flexibility. From the supply side, these three characteristics of a chip's performance — *speed, energy efficiency, flexibility* — form a trilateral technological frontier for the chipmakers, guiding their innovation efforts. From the demand side, these characteristics constitute a chip's value and represent criteria of consumers' choice. Figure 5.2a graphically represents the trilateral frontier.



(a) The trilateral frontier of chip production



(b) AMD's representation of the frontier

Figure 5.2: Different representations of the trilateral frontier

The von Neumann architecture at the core of the majority of chips provided a sufficient level of flexibility for many applications over 40 years, up until approximately year 2010. During this period, the pursuit of miniaturisation strategy through scaling down the size of transistors and, hence, doubling their number provided a 40% increase in speed while keeping the energy consumption constant.[12] In other words, it was possible with one move

[12]Here we refer to Dennard scaling rather than to Moore's law. Moore's law is an empirical regularity relating time and feasible density of elements on a circuit at minimum cost. Dennard scaling is a scaling law based on formal physical principles (Dennard et al.,

(transistor miniaturisation) to achieve improvements in two out of three characteristics, speed and energy efficiency, without harming the third one, heterogeneity. However, approaching atomic scale, transistor size reduction cannot continue at the same pace. Introduced by Intel in 2011, 3D transistor (Auth et al., 2012) instead of planar ones extended the technological trajectory of increasing processing speed due to higher density of elements on a die approximately until 2025. Miniaturisation of elements as a strategy concerns all architectures but eventually will not be further possible, and producers have to decouple speed and energy and to look for other techniques to push the frontier forward.

Relentlessly pressured by demand's needs, the von Neumann architecture evolved in complexity to encompass numerous functions in one chip. Miniaturisation provided more space on a chip to implement not only more but also diverse elements, for example, heterogeneous logic cores, on–die memory (cache), connecting channels. Heterogeneity and multiplicity of elements made the architecture capable of performing a wide range of computations and, hence, algorithms. However, the more complex the architecture the more it is flexible but also the higher the costs of fabrication and the harder the management of its energy consumption. This sets the flexibility aspect at odds with energy efficiency. Indirectly, and returning to the discussion on algorithms, flexibility makes it possible to run sophisticated algorithms, but it is not necessarily associated with high speed of processing them; another emerging tension is therefore between flexibility and processing speed. Implementing the principle of universality, i.e. flexibility, controlflow architectures mostly concentrated on improvement of processing speed and on keeping energy consumption under a constant envelope by miniaturizing elements and adding new modules on a chip (Borkar and Chien, 2011). With the advent of AI, limits of flexibility of the von Neumann architecture started to be seen: it can still run ANNs, but it is poorly–suited for that. Instead, the dataflow architectures are fit for ANNs, but for the moment

1974). It states that (i) reduction of a transistor's dimensions by 30% (to 70% of initial size) allows shrinking its overall area by 50% ($0.7^2 = 0.49$) hence twice as many transistors fit in the same area on a die; (ii) consequently, the transistor's channel length and interconnections reduce as well by 30%, reducing the time of switching and transmission of current across the circuit — "0.7x delay reduction, or 1.4x frequency increase"; (iii) this allows lowering the voltage and hence energy consumption (Borkar and Chien, 2011, p.68)

lack flexibility both within (Sze et al., 2020) and beyond this type of algorithms e.g. embedded software. In the meantime, a growing number of businesses (startups as well as incumbents) experiment and adopt AI–based solutions, expanding the *number of markets* for AITs (see Sections 4.2 and 4.3 in Perrault et al. (2019)). Thus, flexibility is an emerging factor of the frontier and becomes a vital concern of chip producers for either type of architecture as well as for the whole SoC.

In Figure 5.2 we put side by side two representations of the technological frontier. The first one in panel 5.2a, constructed for this Chapter, and the second one in panel 5.2b is reproduced from a 2019 AMD's keynote address at a symposium on high performance chips held at Stanford University (Su, 2019). AMD's representation additionally places examples of chip types into the same framework; number of applications on the horizontal axis represent flexibility or generality, and performance/Watt on the vertical axis combines processing speed and energy efficiency. As expected, General Purpose Cores (CPU) have the widest support of applications but in terms of performance/Watt falls behind GPGPU and Semi–Custom SoC (highly specialised circuits tailor–made for a particular application, also called ASIC, e.g. TPU).

In sum, the trilateral frontier is a coordination mechanism between supply and demand for the development of computing devices, both processors and SoC. On the one hand, each product is characterised by these three metrics. On the other hand, there is a sheer number of application markets that place different weights on each of the frontier's characteristics. Therefore, the size of demand for a particular product can be estimated as a share of markets for whom a product matches the most with consumers' preferences with regard to these characteristics.

5.4 The Future of Chips: Fragmentation vs Platform

In previous Sections we discussed the changing equilibrium between models of computation pulled by AITs and how this reverberates on the choice–set of chip producers for what concerns the architecture of their products. In this section, we construct an economic framework for the development process of a computing device. In this part under computing device we mean both types of products — a processor and a SoC — referring to both as *chip* for convenience. We proceed further in the formalisation of the dynamics unfolding in the semiconductor industry and present a model of demand distribution driven by the value of a chip composed of the frontier's characteristics. The model stresses and illustrates the role of flexibility as a recently aggravated criterion of consumers' choice, and endogenises it through the software environment. Finally, drawing on the analysis of the technological and economic factors and mechanisms, we derive two scenarios for the evolution of the semiconductor industry that differ by the product form at the core of each trajectory and point out issues for further discussion.

5.4.1 Modelling Chips' Flexibility and Demand Distribution

We start with our novel point on the emergence of flexibility as a criterion of chips' performance. To do so, we include the software domain into the model. First, this allows for an indirect modelling of hardware's flexibility, approximated with the variety of programs a chip can support. Second, this modelling choice reproduces the feedback loop between the software domain and the semiconductor industry, in line with the argumentation of the supporting services approach. The supporting services approach is also referred to as indirect network externalities, where consumers are indifferent to the number of users of a product but are interested in the variety of services that this product gives access to. We ground our model on the framework provided in Shy (2011) and modify it for the case of chips by (i) modeling consumers' utility through the value of a chip, (ii) considering partial compatibility (Chou and Shy, 1993) and (iii) deepening the interpretation

of some parameters due to industry–specific features. In what follows, first we outline the model and then we comment it, supporting the results with evidences from the marketplace.

Demand Side. We assume that consumers are uniformly distributed over a unit interval and indexed with x. Consumers can be considered as individuals or application markets. They buy only one chip each, making a choice between a chip i and a chip j. Each of the chips can address the frontier's characteristics to some extent by employing different techniques we discussed in previous Sections. For example, chip i can be either solely based on the controlflow architecture with many homogeneous (scalar) cores and implement quasi–parallelism with the help of software frameworks, or chip i can be mainly based on the controlflow architecture with some additional tensor cores, like Intel's Nervana, or it can represent a highly heterogeneous SoC comprising several architectures such as Huawei's Ascend 910. The parameter δ is usually interpreted as degree of product differentiation. In the case of chips, the parameter δ measures the disutility from purchasing a chip that does not completely match with the type of computation it is bought for by a consumer. For example, if a consumer needs a chip for mainly controlflow–organised computations and only for a small share of dataflow computations, the parameter δ reflects the reduction in utility from buying a non–perfect match to the consumer's needs.

$$U_x = \begin{cases} V^i - \delta x - p^i & \text{if buys chip } i \\ V^j - \delta(1-x) - p^j & \text{if buys chip } j \end{cases} \tag{5.1}$$

Equation (5.1) represents the utility of a consumer from buying one of the alternatives, where the chips' values V^i and V^j are described as:

$$V^i = E^i S^i \tag{5.2.1}$$

$$V^j = kE^i S^j \tag{5.2.2}$$

The value of a chip relates to the frontier's characteristics described in Sec-

tion 5.3.2. The rationale behind the E component is as follows: the processing speed or performance of a chip is measured in operations per second. The energy efficiency of a chip is basically its energy consumption per unit of time, expressed in Watts (W).[13] Thus, we introduce the combined efficiency measure E obtained by dividing performance (in operations/s) over energy efficiency (in W), merging two of the frontier's characteristics into one. Note that the higher the energy efficiency, the smaller its measure in W. This or similar measures are indeed used in the industry. For example, the recent analysis of Open AI on the amount of compute shown by modern AI systems uses FLOPS/W[14] as performance measure which, it is argued, is also correlated with FLOPS/$ (Amodei et al., 2019). Any of these measures fit into the model's logic. The parameter k is a scaling parameter that helps to express one chip's efficiency in terms of the other chip's efficiency, i.e. $E^i = k \times E^j$. For example, if $k = 2$, it means that chip i is twice as good as chip j by either performing twice as many calculations with the same energy consumption or consuming twice less energy to perform the same amount of calculations.

The remaining frontier's characteristic, flexibility of computation, is more subtle to model. From the discussion in Section 5.2 we know that programs can be addressed through either model of computation (sequential or concurrent) and hence performed on any type of chip, however with a sheer difference in time and/or energy endowment. Therefore, there is some degree of interchangeability between chip types that can be expressed in terms of software. Note though, that execution of some programs is so inefficient on a particular chip (e.g. for ANNs parallelism as a requirement to converge to a solution in reasonable time) because of yet absence of developed software environment or framework that would allow efficient execution on another model of computation. In practice, what matters is how many programs can be performed using a specific chip within a reasonable span of time and energy envelope. Therefore, the flexibility of a chip is modeled as

[13] More precisely, under energy consumption we mean the amount of energy required to move an electric charge, expressed in joules (J). Thus, energy efficiency can be measured in joules per second which is equal to Watts: $W = \frac{J}{s}$.

[14] FLOPS stands for floating point operations per second.

follows:

$$S^i = s^i + \rho^i s^j \quad \rho^i \in [0;1] \tag{5.3.1}$$
$$S^j = s^j + \rho^j s^i \quad \rho^j \in [0;1] \tag{5.3.2}$$
$$s^i + s^j = 1 \tag{5.3.3}$$

The total amount of software that can be run on, for example, a chip i is S^i which consists of two components: (i) the amount of chip–specific software s^i, (ii) a share of software written for the other chip that can be interchangeably run on both chips $\rho^i s^j$. The total amount of programs to be performed is normalised to 1.[15] The parameter ρ^i reflects software's *partial* ($\rho^i < 1$) and in most cases *asymmetric* ($\rho^i \neq \rho^j$) interchangeability between chips. In sum, the magnitude of S^i approximates the flexibility of computation that chip i provides.

The difference in values of the two chips can be written down explicitly:

$$\begin{aligned} V^i - V^j &= E^i S^i - kE^i S^j \\ &= E^i(s^i(1 - k\rho^j) + s^j(\rho^i - k)) \end{aligned} \tag{5.4}$$

To analyse the comparative statics of the value difference shown in equation (5.4), we simply take partial derivatives with respect to each variable in the expression.

$$\frac{\partial(V^i - V^j)}{\partial k} = -E^i(s^i\rho^j + s^j) = -E^i S^j < 0 \tag{5.5.1}$$
$$\frac{\partial(V^i - V^j)}{\partial \rho^j} = -E^i s^i k < 0 \tag{5.5.2}$$
$$\frac{\partial(V^i - V^j)}{\partial \rho^i} = E^i s^j > 0 \tag{5.5.3}$$

Equation (5.5.1) shows that the more efficient (higher k) the chip j in terms of combined performance and energy efficiency in comparison with the chip

[15] We purposefully do not model a law of motion for s^i, as we are interested in understanding the allocation of users across systems given a set of available supporting software and their degree of 'multihoming' captured by the ρ parameters.

i, the smaller the gap between the chips' values, other things equal. An increasing capability ρ^j of the chip j to run software s^i written for the chip i also reduces the gap between the chips' values in favour of the chip j. Contrariwise, increasing ρ^i leads to growth of the gap in favor of the chip i, other things equal. The two derivatives with respect to the ρ parameters can be interpreted as the attempts of producers to invest in architectural improvements in order to incorporate the functionality of the competing chip, and therefore to manipulate the indirect network externalities.

The remaining two variables s^i and s^j can be interchangeably expressed according to (5.3.3), hence $s^i = 1 - s^j$ and $s^j = 1 - s^i$. Using this substitution and taking partial derivatives, we obtain the following:

$$\frac{\partial(V^i - V^j)}{\partial s^i} = E^i k(1 - \rho^j) + E^i(1 - \rho^i) > 0 \tag{5.6.1}$$

$$\frac{\partial(V^i - V^j)}{\partial s^j} = -E^i k(1 - \rho^j) - E^i(1 - \rho^i) < 0 \tag{5.6.2}$$

These equations show that with a growing amount of software s^i written for the chip i the gap $(V^i - V^j)$ increases, while the opposite holds for s^j. In general, the technical superiority of a chip i in terms of performance per Watt $E^i = kE^j$, along with more tasks supported on this chip s^i increases its value V^i and hence increases the gap $(V^i - V^j)$. If the software interchangeability parameter of a chip ρ^i would be equal to 1, that would mean that chip i is capable of performing all the tasks that chip j does. In other words, if $\rho^i = 1$, the chip i would support the highest possible flexibility of computation. However, precisely the imperfect interchangeability of chips, captured by $\rho^i < 1$, doesn't allow for completely dismantling either of the chips. Lastly, it is worth stressing that here we want to analyse the dynamics of the values gap when varying each of the component. Therefore, while the expression $(V^i - V^j)$ can grow along, for example, s^i, the absolute value of the gap can be any, positive or negative, namely $(V^i - V^j) \gtreqless 0$.

As highlighted by the construction of the trilateral frontier, the superiority of one chip over the other is based on three factors combined together. For example, it is not sufficient for a chip to exhibit the lowest energy consumption if the processing speed and flexibility available are low. Moreover, even any

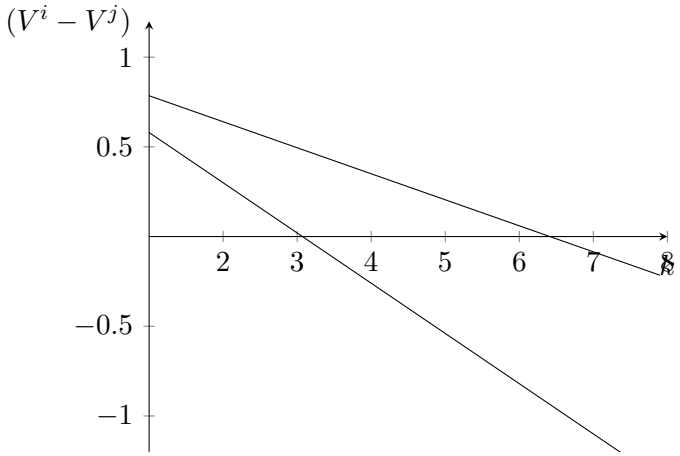


Figure 5.3: Effect of the efficiency multiplier k on the difference between chips' values $(V^i - V^j)$ varying flexibility parameters s^i, s^j, ρ^j

$t = 1$: $s^i = 0.9, s^j = 0.1, \rho^j = 0.05$;
$t = 2$: $s^i = 0.8, s^j = 0.2, \rho^j = 0.1$
$E^i = 1, \rho^i = 0.3$ for both $t = 1, 2$

pair–wise superiority can be outweighed by a deep enough inferiority with respect to the remaining third characteristic. In order to illustrate how the superiority of a chip is reached through the balancing of all the three frontier's characteristics, we constructed a stylised example, visualised in Figure 5.3. In this example, in period $t = 1$ the amount of software performed on chip j is only 10%, $s^j_{t=1} = 0.1$, and only 5% of programs performed on chip i are interchangeably executable on the chip j, $\rho^j_{t=1} = 0.05$. In period $t = 2$, both parameters are doubled, namely $s^j_{t=2} = 0.2$ and $\rho^j_{t=2} = 0.1$. Note that according to (5.3.3) if $s^j_{t=1} = 0.1$ hence $s^i_{t=1} = 0.9$, then if in the second period s^j increased to 0.2 hence $s^i_{t=2} = 0.8$. This doesn't necessarily mean that the absolute amount of tasks performed by the chip i has shrunk; it might simply mean that the amount of tasks performed by the chip j expanded without taking over tasks from the chip i; this can be the case of ANNs. Thus, in period $t = 1$, given the setting, in order to have equal values, $(V^i - V^j) = 0$, for the chip j it is required to be almost 6.5 times more efficient in terms of combined efficiency E (performance per Watt) than the chip i, $k_{t=1} = 6.41$. In the second period $t = 2$, when the amount of software that can be run on the chip j reaches 20%, $s^j = 0.2$, and interchangeability doubles from 5% to 10%, $\rho^j_{t=2} = 0.1$, in order to have equal values chip j

has to be only 3 times more efficient, $k_{t=2} = 3.07$. This numerical example displays the mechanism at work, but might not represent an accurate calibration of the parameters. However, it illustrates the trade–offs and balance through which superiority can be achieved.

Finally, it is extremely important to note that, despite the fact that the numerical expressions of improvements of s^j and ρ^j are incremental from one time period to another, it can be technically rather hard to achieve such improvements, which might also take a substantial amount of time; time periods used in the example are not specified but could be one year or a couple of years, resembling the timing of each next technological node in the industry.

Supply Side. On the supply side we assume a duopoly with price competition. Thus, we are searching for a Nash–Bertrand equilibrium. By equalising utilities from equation (5.1) we obtain the indifferent consumer:

$$\hat{x} = \frac{V^i - V^j + p^j - p^i + \delta}{2\delta} \tag{5.7}$$

Each firm has a profit function:

$$\pi^i = p^i q^i = p^i \hat{x}; \quad \pi^j = p^j q^j = p^j (1 - \hat{x}) \tag{5.8}$$

Maximising profit with respect to price, we derive the Nash–Bertrand pair:

$$p^i_{NB} = \frac{V^i - V^j + 3\delta}{3}; \quad p^j_{NB} = \frac{V^j - V^i + 3\delta}{3} \tag{5.9}$$

For equilibrium prices to be non–negative the condition $-3\delta \leq V^j - V^i \leq 3\delta$ has to be fulfilled. Finally, plugging the results of equation (5.9) into (5.7) we obtain the final formula for the indifferent consumer:

$$\hat{x} = \frac{V^i - V^j + 3\delta}{6\delta} \tag{5.10}$$

The position of the indifferent consumer is defined by two factors. First,

the difference in chips' values ($V^i - V^j$) whose analysis was shown in equations (5.4)–(5.6) can be employed here as well. Second, the disutility δ from the level of mismatch between computations used by a consumer and computations available on the purchased chip. Deepening further the interpretation, this means that the parameter δ reflects a degree of application specialisation or computational convergence. Let's imagine for simplicity that each application market needs a chip to perform one task. One corner case would be when every task is performed by a single algorithm establishing an unequivocal correspondence between the two; in that case δ is the highest because every task is a distinct type of computation. The second corner case would imply that every task consists of all possible types of computation; in that case δ must be low because all tasks are composite. From an economic perspective, in the first case an application market runs only one type of computation, hence it has a strong preference for a chip that runs this calculation better. More generally, if demand consists of consumers each employing highly homogeneous and distinct computation, δ for the chip–making industry would be very high. The realistic case is none of the extreme ones, with demand consisting of consumers each using its own mixture of algorithms and only few consumers representing extreme cases each using either purely homogeneous or purely heterogeneous calculations. That is why δ can be interpreted as a measure of mismatch between the variety of software supported on a chip and the variety of computations used by a consumer.

The model does not contain a cost variable; however, implicitly a higher value is associated with higher costs to achieve it. For example, according to the financial statements of the ASML Holding (the leading company in the market of photolithography systems), the price of an average system sold in the first half of 2019 is in the range of 36–38 millions of euros (ASML Holding, 2019). Such equipment is highly standardised and it would only account for the initial investments to establish the production process. Leaving aside the formal mechanism of cost formation, in our model we deal with its final instantiation — the price. From an economic viewpoint, the purpose of our model is to show how the shares of demand are driven by the frontier's characteristics and prices as a touch–point of supply and demand. Thus, costs are involved implicitly through the cost of production of a chip

with a particular value and its improvement with respect to the frontier's characteristics. It is beyond our analysis to explain *how* a particular value of a chip is achieved, while extensive technological insights regarding chips' characteristics and directions of their improvement currently under exploration in the industry were provided in previous Sections. Here we simply assume that every firm estimates its fixed costs to produce a chip and the quantity demanded in order to understand whether or not the production of a chip will be profitable, exploiting either a high price at low quantity or economies of scale at a low price. If a firm estimates that costs might overweight revenues, it doesn't enter the market.

Parallels in the Marketplace. The frontier's characteristics are operational leverages on which a semiconductor company can act in order to improve its product; producers choose technical approach and the degree of addressing these characteristics based on cost–benefit analysis, aspiring to create a product that appeals to a larger share of demand. In the previous Sections we provided examples of innovations in architecture, elements, materials and techniques that target processing speed, energy efficiency, and flexibility. Hierarchical networks instead of bus interconnect, experimentation with wafer size, new materials and signal types, 3D instead of planar transistors, die stacking, in–memory computing, array and neuromorphic processors, and heterogeneous SoC, all illustrate producers' actions undertaken to act on different segments of the trilateral frontier. Their decision results in the next generation of chips with different values offered to consumers. The real–world example of such behavior is Intel's Process–Architecture–Optimization strategy that was implemented in 2016, replacing the so–called Tick–Tock strategy.

Provided that we model flexibility through the software domain, this implies that producers can allocate their effort to increase the flexibility supported by their products in two ways: (i) introducing changes in hardware to expand functionality and through that encompass more of the existing software from the competitor; in other words, increase ρ^i, (ii) investing in the expansion of the software set written specifically for a producer's own chip, which means increasing s^i. The first way, the introduction of hardware changes, is discussed at length in previous Sections, therefore we now focus on the

second way, software–related changes. As mentioned in Section 5.3 regarding GPGPU, Nvidia developed the CUDA framework to support its products; the consortium Khronos Group works in the same direction of heterogeneous computing with its OpenCL framework designed by Apple. Other open–source platforms like Google's TensorFlow and Microsoft's CNTK are aimed at the collaborative development of dataflow software solutions to run on chips that can support them, such as TPU or CPU–GPU tandem. By adapting the existing software and writing programs that can effectively run on its product, a firm i increases the value of its chip i targeting precisely the s^i component. However, producers of the competing chip j can counteract by developing instruction set architecture (ISA) extensions.[16] Modifying ISA by including additional packages of new commands allows the competing chip j to encompass some functions performed on the chip i. In terms of our model, such effort affects ρ^j. As an example, we can mention Advanced Vector Extensions (AVX) and its further extension Vector Neural Network Instructions (VNNI) from Intel for x86 ISA, Vector Multimedia Extension (VMX also known as AltiVec) by IBM for Power ISA and NEON technology from ARM Holdings for its eponymous ARM ISA.

In sum, our model reveals the mechanism driving the distribution of demand based on chips' technical characteristics, available software and how well overall a chip meets the computational needs of consumers. A better–developed software environment and compatibility indicate higher flexibility of a chip, which can appeal to a larger share of demand. In turn, demand is characterised by degree of differentiation with regard to the frontier's characteristics: the higher the differentiation the more precise features of a chip are required by each application market. In general, a chip can exhibit either (i) Pareto improvements with respect to any of the frontier's characteristics gaining more applications or (ii) a trade–off between each couple of characteristics shifting the set of applications.

[16] In essence, ISA modifications are on the borderline between hardware and software (programmatic) changes. Given ISA's undeniable programmatic element, here we employ example of ISA modifications that, similarly to hardware changes, impact ρ^j.

5.4.2 The Industry at a Crossroad: Alternative Scenarios

From the discussion so far we set out a collection of mechanisms and forces shaping from the outside and within the evolution of the semiconductor industry. Exogenous challenge that arrived from the AI segment tests the robustness of the established technological trajectory. In fact, this time the challenge lies at the fundamental level of the computational model on which chips are built on. Residing in declarative programming paradigm that is inefficiently executed on established sequential model of computation, AITs triggered a wave of innovation efforts that resulted in numerous novel products with the dataflow architecture at their core. It is becoming clear that the simple speedup race between competing chips is not the central issue for the future of the semiconductor industry; rather, the more profound issues of organizing the logic of computation and variety of algorithms that a chip can support in order to appeal to a sufficient share of demand are key. The question now is how chips will evolve this time. Considering all the factors at play, we derived two scenarios on which the industry can converge.

Scenario I Under the relentless pressure of economic factors within the semiconductor industry and the continuous but siloed pull from the downstream markets for market–specific improvements, producers might decide to pursue trajectories tailor–made to subsets of downstream markets, grouped around specialised chips that accurately address needs within given submarkets. A *customisation strategy and hence the fragmentation of the semiconductor industry* might occur.

Scenario II Aspiring to address larger shares of demand associated with greater but probably delayed payoffs, chip producers can make long–term investments at the system level, aimed at the creation of a *platform chip comprising heterogeneous cores*. To achieve that, the overarching architecture must reproduce a composition of components on a chip that ensures scalable, heterogeneous and energy–proportional computing. Developed in response to the call of one segment, the platform chip can diffuse over time among other downstream markets with decreasing cost of production and,

hence, price.

Arguments 'pro' and 'against' exists for each of the scenarios. According to our model, if the demand is significantly differentiated it is harder to acquire a large share of consumers (see 5.10), other things equal. The smaller the size of potential demand aggregated over application markets, the harder it is to return high costs of design and fabrication of an heterogeneous chip. Thus, naturally, if differentiation is high, the viable strategy is that of fragmentation of the semiconductor industry's offer into several distinct chips, each characterised by unique performance with respect to the frontier's characteristics; application markets decide to purchase either one or a set of chips based on their needs.

Thompson and Spanuth (2018) advocate for the first scenario by linking the future dynamics of chips production to the dual–inducement mechanism typical of General Purpose Technologies (GPTs) (Bresnahan and Trajtenberg, 1995). They develop a model of choice between universal and specialised processors based on relative speed up factor and identify a cut–off point from which the specialised processors become more appealing than the universal ones. As more and more downstream markets switch to specialised processors, this leads to the halt of the dual–inducement mechanism for universal processors. Thus, they expect the end of the GPT paradigm of universal processors and envisage a situation of application–based market fragmentation with specialised computing evolving in more compartmentalised domains. This prediction rests on (i) a view of the processor as the singleton GPT technology and (ii) the assumption that processing speed is the sole criterion of the choice of a processor. Concerning (i), from this perspective, processors can be considered as pure competing alternatives. However, we also need to consider the possibility that it is the SoC the candidate for the role of GPT, while processors are complementary blocks. As for (ii), we acknowledge that processing speed is an important factor and included it among the frontier's characteristics. However, we argue that it is not the sole criterion for all applications and might not be the primary one for some share of applications. For the development of AI itself, "[f]ocusing on raw computing power misses the point entirely. Speed alone won't give

us AI. Running a poorly designed algorithm on a faster computer doesn't make the algorithm better; it just means you get the wrong answer more quickly. (And with more data there are more opportunities for wrong answers!) The principal effect of faster machines has been to make the time for experimentation shorter, so that research can progress more quickly. It's not hardware that is holding AI back; it's software." (Russell, 2019). As we pointed out earlier in Section 5.3.1, even at the level of processor for AI there is an ongoing search for a more flexible architecture and "TOPS/W alone considered harmful" (Sze et al., 2020). In general, the software domain is dynamic and evolves faster than hardware due to lower costs inherent to information products (Goldfarb et al., 2019), hence the variety of software solutions that a consumer (market, firm or individual) can employ increases over time. Optimization of hardware for a specific software for the sake of faster processing might result in a very limited set of application markets and in a shorter product life cycle. Instead, flexibility is a more sustainable strategy for a chip producer.

Building on the aspect of flexibility, there are arguments in favor of the platform chip scenario. Given that hardware flexibility can be approximated with the amount of software that is effectively run on a chip, the presence of indirect network externalities does have significant implications for the semiconductor industry. This approach suggests that consumers' decision upon which hardware system to buy is affected by complementary products or supporting services, in this case software, available for each system. In particular, Church and Gandal (1992) model the effect of the decision of software firms upon software provision on the market share and the number of hardware systems that will exists in equilibrium. Their analysis shows that when consumers' preferences on *software variety* are relatively high[17], this leads to the exclusive adoption of one of the hardware systems if a critical minimum amount of software is provided. Furthermore, in the case in which two hardware systems exist, total surplus would be higher under many parameters' values if a standard (a single hardware system) was mandated. Thus, strong preference for software variety is associated with the choice of one hardware system. By translating software variety into hardware's

[17]The benefit from high software variety available on a chosen hardware system has to overweight the disutility from spending on the purchase of various software, the price of hardware and the degree of its differentiation.

flexibility, in our model flexibility is a variable that characterises the chip and producers can act upon it by either writing software supported on their chips or by increasing compatibility with the existing one.

In our analysis we covered the technical performance of a chip and demand's preferences as factors that shape the technological trajectory and steer the development of the semiconductor industry. The last factor that can tip the balance in favor of one or the other scenario is concentration of market power among chip producers. There is a number of big players in the semiconductor industry even at the global level; some of them we already named, such as Alphabet, Amazon, Alibaba, Huawei, Samsung, Nvidia, Intel. Many of these companies are cross–industry actors that comprise a diverse portfolio of assets that they built to pursue internally established goals. The dimension that matters in the context of the semiconductor industry is *edge versus cloud computing*. Some of these big players are cloud–oriented like Google and Amazon and they already direct their innovative effort to develop in–house chips to support AI through their cloud services. The primary focus of such chips would be concurrency and speed to support numerous users working concurrently by providing low latency. The already mentioned TPU of Google and Trainium and Inferentia chips of Amazon serve exactly this purpose, being fast, highly parallel and energy efficient however non–flexible. Thus, cloud computing would rather benefit from a set of dedicated chips combined together to deliver state–of–the–art performance with respect to each frontier's characteristic. Contrariwise, chip producers that place their bid on edge computing lean toward more independent and capable devices and, hence, direct their innovation efforts in the direction of the platform chip. The already named Huawei's Ascend 910 is one example. Another prominent example is Apple's M1 chip that comprises CPUs, GPUs and Neural Engine cores in one SoC.[18] Apple stresses the edge–oriented application of its chip with high performance, low energy consumption and flexibility achieved through integration of heterogeneous cores. In line with our reasoning, Apple acts on hardware's flexibility through software variety as well providing Core ML software framework for programming, and optimising its Big Sur operating system to work with the M1 chip. In sum, there are pieces of evidence suggesting that a dominant design for the dataflow pro-

[18] https://www.apple.com/newsroom/2020/11/apple-unleashes-m1/

cessor is on the way while at the same time there is ongoing experimentation with the configuration of the platform chip.

5.5 Related Literature and Discussion

The analysis offered in this Chapter relates to a number of contributions in the literature. We review the works most related to our study and emphasize similarities and differences. To organise the review, we highlight the dimensions shared by our study and the discussed works with respect to, for example, the level of analysis, the industry considered, the economic and strategic or technological arguments provided, and the role played by demand and supply. We start with a focus on the semiconductor industry to identify the forces and mechanisms shaping its technological trajectory and innovation; then, we progressively move to the computer industry to discuss the similar dynamics produced by the introduction of new products. Finally, we take a more fine–grained perspective centred on the design of platform products, whose rules apply to computers as well as to chips.

Steinmueller (1992) focuses on the semiconductor industry and provides a supply–side analysis of the economic arguments — in particular production economies and dis–economies — that have contributed to maintain chip production for decades on a stable technological trajectory tied to the von Neumann architecture and to miniaturisation as its main innovation direction. The Chapter outlines the trade–off between specialisation and standardisation that characterises the industry. Economies of scope fuelling product variety (and, thus, specialisation) and economies of scale fuelling production expansion (and, thus, standardisation) are at odds with each other, and the semiconductor industry has mostly pursued economies of scale. The reason for this is that chip production is characterised by the so–called *capacity races* — the incentive to engage in mass production in order to amortise large costs of equipment capable of little flexibility. The unprecedentedly big chip produced by Cerebras, which is fabricated on conventional photolithographic equipment is recent evidence of the persistent importance of equipment cost as economic factor. Capacity races produce two economic effects; the first one concerns the incentives for firms to be first movers in

innovation by pre–empting competitors in the introduction of new generations of chips in order to earn additional payoffs out of the investment in capacity. The second effect is the incentive to push for cost reduction in order to hasten the scaling up of production. While Steinmueller advocates for the exploration of flexible chip manufacturing technologies to make product variety economically viable at low output volumes, the forces he highlighted have mostly prevailed and kept the industry on a well–defined trajectory of standardisation and mass production. The same dis–economies of scope that enabled the production at scale of processors embodying the von Neumann architecture can push chipmakers to produce a platform chip. The success of ANNs and the demand for AI applications has drawn chipmakers' attention to concurrent models of computation, but the forces pushing for standardisation can again create a strong incentive to the integration of heterogeneous processors into a single product that can exploit economies of scale.

Adams et al. (2013) study innovative activities in the semiconductor industry in the 80s and 90s. They take a sectoral systems of innovation perspective to highlight the role played by intermediate users' demand in innovation. From the supply–side, a series of technological changes (i.e. the development of Electronic Design Automation tools) allowed for the dis–integration of the chipmaking supply chain. With weaker ties between the design and manufacturing phases, entry barriers for specialised firms (e.g. the so–called 'fabless' firms) at different points of the chain lowered. The new actors could partner with foundries to offer specialised designs to specific market niches and co–exist next to integrated producers, as the latter focus on more systemic innovation that require superior coordination efforts (Kapoor, 2013). From the demand–side, an increasing amount of the market niches started to emerge with the opening of new applications for integrated circuits — in particular wireless communication and mass consumer products; these niches are characterised by the demand for tailor–made chips. The combination of more fragmented production processes and differentiated final demand increased the importance of application knowledge, and thus induced co–innovation by semiconductor and user firms. While Steinmueller's capacity races have confined chip production within a well–defined technological trajectory shaped by economies of scale, the supply chain dis–integration

illustrated by Adams and co–authors has allowed an increase in product variety through the production of specialised chips for market niches. However, this dynamics relaxed but not dismantled the dominance of classic von Neumann chips.

Malerba et al. (2008) analyse to the joint structure of the semiconductor and computer industry and use a 'history–friendly' model to reproduce the discontinuities that technological innovations in semiconductor devices induced on the industries. The Chapter provides a simulation of economic and technological mechanisms on both the supply and demand side to map the co–evolution of two industries' market structures. For example, the authors discuss how the introduction of integrated circuit in the 60s allowed IBM to control both the development of semiconductor devices and their implementation in mainframe systems such as the IBM System/360. The microprocessor, introduced by Intel in the 70s, challenged IBM vertically–integrated production, dis–integrated the supply chain and lead Intel to dominate the semiconductor industry. Each new class of semiconductor devices triggered changes in the industry's structure. The latest technological discontinuity we describe in the Chapter — the embedding of the dataflow model of computation into chips as a result of ANNs 'shock' success — will also reverberate into changes in the industry's organisation. The current turbulence characterised by exploration of product designs and entry by companies dominating in adjacent markets (e.g. Nvidia, Amazon, or Apple) and startups (e.g. Cerebras or Graphcore) might turn into a new structural equilibrium as soon as production economies and dis–economies will play out.

In general, we can consider the problem of producing a new chip capable to integrate sequential and concurrent models of computation as a problem of bundling features into the overall configuration of a product that combines heterogeneous components. Such configuration is an instantiation of a 'platform product' in the context of the semiconductor industry. Platform products have been extensively studied in the context of the computer industry. The industry has been producing computer platforms (an innovation inaugurated by the IBM System/360 — see Baldwin and Clark (2000)) integrating different components — chips being a core and often the highest–

value one. A platform chip and a computer platform are two different types of product; however, the mechanisms at work shaping the configuration of a platform are essentially the same at different levels. Any platform product is subject to dynamic tensions of both economic and technological kind, as the relationship among its components needs to accommodate both innovation and degrees of (backward) compatibility. These tensions emerge at many levels, from the industry to the firm and product level. For this reason, while the focus of the Chapter is a very fine–grained one — the platform chip — we can refer to findings from the literature on computing platforms.

Bresnahan and Greenstein (1999) take an economic perspective on computer platforms, as they study the evolution of technological competition and market structure in the computer industry. Focusing on the industry, their scope of analysis is broader than ours as they consider in detail the industry and segment–wise convergence to equilibrium. However, the key points they make applies to our case as well: first is the need to focus on (platform) products rather than firms as unit of analysis. In fact, computer platforms (such as the IBM System/360, or Apple Macintosh) have been the point of interaction between supply and demand in the computer industry. A second key element is the role of endogenous sunk costs and demand (reflected in the market segments served by the industry) in shaping which platform product gains dominance and persistence. A third important element regards the nature of competition in the industry after what they label the 'competitive crash' of the 90s: platform competition within the same market segment was the result of indirect entry, with new computer platforms first entering a novel market segment with specialised (usually technical) users and then moving to established segments (business and then consumer users — a dynamics illustrated also in Bresnahan and Yin (2010)). Bresnahan and Greenstein's account of the computer industry's evolution around platforms illustrates how industry–wide and within–segment equilibrium are related. "Equilibrium in each segment of the computer industry obeys its own logic of concentration and persistence, determined by buyer/seller interactions" around a platform; indirect entry allows segment dynamics to channel change to the industry level. This dynamics resulted in a 'divided technical leadership' in the computer industry. Our model is a snapshot of this very mechanism at work in the context of semiconductor industry, where

the matching of chips with demand structure (with user needs approximating market segments) determines the industry–wide equilibrium split among alternative technologies. In particular, we can apply their framework to our case by considering AI–users a novel market segment for the semiconductor industry. Through indirect entry, new chips can occupy the segment of AI–users first and then move to compete with established products in other segments. As in the case of the computer industry, the success of direct entry or the insulation of a new platform within a segment depends on several factors — the differentiation of demand's needs in different segments, as well as the technological features of the new chip. A platform chip integrating classic and AI–specialised components could induce a competitive crash with dominant products in many segments served by the semiconductor industry.

Taking stock, the Chapter shares with Malerba et al. (2008) the interest on technological discontinuities. Instead of focusing on the varying structure of the semiconductor and computer industries' supply chains, we are interested in how the push to introduce chips capable of supporting AI applications as a new discontinuity will influence chipmakers product design. Our focus is on how producers will bundle AI–related computations into the functionality supported on a chip, considering that their production choices are influenced by technical feasibility, design and fabrication costs, the matching between product characteristics and end–users demand. Taken together, Steinmueller (1992), Adams et al. (2013) and Bresnahan and Greenstein (1999) provide us with a useful framework to understand the channels through which such new product can emerge and whether it can appeal a major share of market segments (and demand needs), as suggested by our Scenario II. Adams and co–authors and Bresnahan and Greenstein show how the structure of demand (and its participation in innovation) is a potential source of product variety. Steinmueller shows how dis–economies of scope drive the industry back towards standardisation. The current moment is a crossroad. One the one hand, the standard over which the industry settle can emerge through the process of indirect entry and a new competitive crash. In our case, the commercial use of AI and ANNs has indeed induced the exploration of new product space; the successful design of a platform chip integrating AI and non–AI components can enter a specific market niche first and from there diffuse and emerge as a dominant design

for the whole industry. On the other hand, high costs and the appeal to an insufficient share of demand can fragment the product space resulting in non–overlapping demand clusters served by custom chips. All these mechanisms can be rationalised by our model by considering the tension between high differentiation in the structure of demand and at the same time one of the competing systems displaying high flexibility so to serve at scale a large share of the markets.

Cusumano and Gawer (2002), Gawer and Henderson (2007), and Burgelman (2002) shift the analysis from the industry to the firm and product level. As the product platform they study are microprocessors, their work is proximate to ours in terms of the level of analysis. These studies provide a more find–grained analysis of the tensions emerging inside the leading platform sponsor, Intel. The tensions they describe relate to strategic management choices but capture mechanisms at work in our case as well. Cusumano and Gawer (2002) focus on the platform level and discuss the tensions emerging in the process of balancing the relationship between the platform owner (the firm controlling the core architecture of the system) and its complementors. In a platform product, owner and complementors are linked in an ecosystem that displays non–generic and supermodular complementarities (Jacobides et al., 2018). Chipmakers such as Intel experience 'platform dependency' as — to quote the Director of Intel Architecture Lab mentioned by the authors, "(w)e are tied to innovations by others to make our innovation valuable". Intel's strategy with respect to complements is explored in depth in Gawer and Henderson (2007). They study the incentives for the platform owner to enter complementors' markets; the tensions highlighted in this case are prevalently those internal to the organization. In fact, the strategy to expand the demand for microprocessor implied growing the whole computer platform (microprocessors plus complements) and, thus, to allow complementors to grow profits as well. However, a balanced growth of the whole platform contrasted the need of Intel to enter and 'squeeze' profits from complementary markets (for example motherboards, online services, PC peripherals and accessories). To address this tension, Intel entered complementary markets only when the company matched the capabilities of the competitors and prevalently when the complementary markets were 'connector' markets, those producing products that embodied interfaces to the

core technology (e.g. chipsets, or motherboards). Burgelman (2002) zooms further into organizational mechanisms to illustrate the role of leadership in resolving the tensions occurring within Intel, in particular in the period surrounding Andrew Grove's tenure as CEO (1985–1998). Burgelman stresses how innovation taking place at the level of the platform product tied Intel (the producer of the highest value component of computers, microprocessors) to its complementors. This tying resulted in a co–evolutionary lock–in (the platform dependency of Gawer and Cusumano), which has strongly impacted Intel's strategy: to maintain market power, Intel needed to align the pace its technological advances to that of the whole system. The relevance of co–evolutionary lock–in becomes evident in Burgelman's recounting of the resolution of Intel's 'internal battle' between the i860 and the x86 microprocessor architecture (and, respectively tied to them, between the so–called RISC and CISC instruction set). Intel opted for the x86 architecture, especially as it decided to follow the strategy vector leading to focus on the personal computer (PC) market segment; in fact, Intel "increasingly tied its strategic direction and economic fortunes to the evolution of the PC market segment". Co–evolutionary lock–in has been an additional force, this time technology–driven and occurring at the product and firm level rather than emerging from dis–economies of scope and applying to the whole industry, to conserve the semiconductor industry persistent technological trajectory. As dis–economies of scope can push the current state of the semiconductor industry towards our scenario II, the same holds with co–evolutionary lock–in: chipmakers can explore alternative product design, but their economic fortunes are tied to their complementors. As long as uncertainty characterises the future applications and evolution of AI algorithms, the development of a platform chip design would ease coordination among actors and provide a safer bet regarding the future evolvability of the system.

In line with the Chapter, the studies we reviewed highlight tensions and different mechanisms shaping a platform product. However, they all tell a story placed within the established technological trajectory with the sequential model of computation at its core; instead, we look at the impact of a more profound technological discontinuity — the exhaustion of the capability of the sequential model of computation to address an increasing variety of algorithms and the rise of concurrent model of computation. Despite AI

being an active field since the 1950s, this discontinuity did not take place before because AI as an application segment did not have a big weight and only recently entered a commercial phase at scale. Before the current AI commercial boom, the majority of other applications requiring computing devices could get by with the ever–improving von Neumann architecture. The competitive crash among computer platforms described by Bresnahan and Greenstein and the changes in the structure of the semiconductor industry supply chain discussed by Adams and co–authors occurred in response to discontinuities in technology and changes in demand, but were not contesting the organisation logic of chips. Instead, the increasing demand for AI–related computing that we unpacked starts to exert a stronger pressure than the prevailing architecture could accommodate, launching a wave of radical and architectural innovations, respectively developing specialised components and experimenting with the design of a platform chip.

5.6 Conclusion

In this Chapter, we investigated how a technological discontinuity can impact product design and production strategies in a highly technological industry. The industry we focus on is the semiconductor industry, and the technological discontinuity is introduced by the novel application segment of AITs that grows rapidly. In turn, the use of AITs for a growing variety of applications produced an increasing demand for compute. This has triggered a search for the hardware (chips) capable of executing AI algorithms such as ANNs more efficiently. The chips on which the semiconductor industry has built its success and that dominated its technological trajectory for decades are built around the von Neumann architecture, that is ill–suited to execute modern AI algorithms and ANNs in particular. In fact, the commercial boom of AI has shed light on the limitations of this classic architecture despite its continuous performance improvements over time. For this reason, chip producers are shifting attention to alternative architectures to implement in their chips. The prospective candidate is the so–called dataflow architecture. This is the hardware implementation of the concurrent model of computation, a different model compared to the sequential one at the

core of the von Neumann architecture. The properties of the dataflow architecture match better the organisation of computation underlying AI algorithms. This technological discontinuity with respect to the industry's established technological trajectory represents the challenge the chipmakers are currently facing. Thus, we studied the nature of this discontinuity and how forces and mechanisms at work in the semiconductor industry might steer its further development in one of two potential scenarios.

As our study deals with technological innovation in a highly technological industry and stresses the systemic relationship between hardware and software, it is a novel contribution to several strands of literature, from the economics of AI to the study of platform products in the context of the economics and strategic management of the semiconductor and computer industry, as well as to the literature on technological trajectories. In the analysis, we combined insights and perspectives from different fields such as AI, engineering and computer science with modelling approaches from the economics of software and system products. In order to assess the direction in which the AI discontinuity is steering the design of chips, we started our study by overviewing the computational framework for ANNs. We highlighted how ANNs are endemic to the so–called declarative programming paradigm in virtue of their organisational logic as algorithms, and how the concurrent model of computation, as opposed to the sequential one, matches this logic. Given that, we reviewed how models of computation are implemented in hardware architectures and explored the difference between scalar and vector processors embodying the sequential model and novel architectures such as array and neuromorphic ones embodying the concurrent model of computation.

When designing a chip, producers can opt for one or the other architecture or for an integration of them into one SoC. The performance of the resulting chip is measured with respect to the three fundamental characteristics — speed, flexibility, and energy efficiency — that constitute a trilateral technological frontier. The frontier serves as a benchmark for producers, guiding their design decisions. However, the market success of a new chip depends on the demand's preferences with regard to these frontier's characteristics. We captured this mechanism with an analytical model determining the dis-

tribution of demand between two alternative chips based on hardware's flexibility approximated with the software variety available for a chip and an efficiency metric that combines processing speed and energy efficiency. In stylised terms, the model represents the current state of the semiconductor industry, with AI applications expanding demand variety (directly through itself and indirectly through AI–using segments) and the difference among competing chips reflecting the experimentation surrounding the design challenge.

All the forces and tensions we described have derailed the established technological trajectory of the industry and injected uncertainty regarding the novel track on which it will settle. We summarised the outcomes to which the future of chip can converge in two scenarios. In the first, the demand from the AI segment lead to the development of specialised chips but does not induce changes to the industry–level equilibrium — chip production becomes siloed and fragmented. In the second, in response to the increasing variety of algorithms with the advent of modern AI and under the pressure of production economies, chip producers allocate innovative efforts to the flexibility of their products creating a novel platform chip. Such chip would encompass different architectures onto a single substrate and come to serve most of the industry's demand segments. Both scenarios can emerge out of the current technological turbulence in the industry, and we lay out arguments in favour and against them. However, we stress that within the AI field the pace of progress is high, as well as software domain is more dynamic than hardware. The growing variety of (competing) AI techniques and algorithms raises a valid concern with over–specialisation; given the high costs to develop and produce novel semiconductor devices and the high inertia of these processes, the decision to fork the production of chips with specialised products tailored to current AI algorithms while other approaches are yet in their infancy can be a risky bet for chipmakers. Directing innovative efforts towards flexibility and, thus, a platform chip might result in higher pay–offs in the long run. Quoting Marvin Minsky, "[t]he power of intelligence stems from our vast diversity, not from any single, perfect principle"; the same holds for the potential appeal of a platform chip capable to harness the diversity of computations.

Chapter 6

Conclusion

Measuring the diffusion of a particular technology provides a snapshot of its socio-economic importance that, however, already manifested itself. Studying historical patterns of diffusion helps to reveal regularities and mechanisms that can be shared by different technologies and can be used to understand their socio-economic impact ex ante. In this dissertation, I tried to show that apart from neoclassical ones there is a place for evolutionary mechanisms and factors that explain socio-economic the transformations brought by novel technologies such as ICT and AI.

6.1 Main Findings and Novelty

Each chapter of the dissertation contains insights into ICTs or their effects employing a mix of economic, technological and systemic arguments, in line with the research objective. Chapters 2 and 5 focus on the explanation of economic outcomes, namely productivity dynamics and product selection, including into the analysis technological factors. Chapters 3 and 4 represent in-depth studies of technology systems, ICT and AI respectively, providing a more structured view of these technologies with implications for business and policy.

The central idea and the novelty of Chapter 2 is in the assumption that variance in macro productivity dynamics across countries can be explained at least partially with a country's unique industry mix. The discovered dynamics going against the market selection mechanism when less productive industries gain labor shares or highly productive industries lose employment shares might be a sign of structural change. An industry's productivity above the economy's average can be the result of automation processes and skill-biased technical change. This seems to be the case for scale intensive industries that demonstrate a strong and positive contribution to the macro productivity while decreasing their labor shares across the studied countries. Driven away by these processes, labor follows the demand and ends up in labor-intensive economic activities, for example services, where productivity has a "physical" limit. Another finding supporting the technology-based explanation of the macro productivity slowdown is a near-zero improvement of productivity for specialised suppliers coupled with a non-negligible improvement for supplier dominated manufacturing and services. The latter are the receivers of the technological know-how from the former. Thus, such dynamics apparently suggests a transfer of productive potential from specialised suppliers to supplier dominated industries and services but no productive input for the former. This situation can be generated either by exhaustion of established technological opportunities for the specialised suppliers or by lags associated with the implementation of new products and processes. If the exhaustion of productive potential of the technologies currently at use is the case, the source of novel opportunities for specialised suppliers is likely to experience a slowdown as well. As such source lies in the area of scientific and applied research, we estimated trends of research productivity in the studied countries and detected a generalised deceleration, or so-called "innovation slowdown", mirroring the labor productivity slowdown. The literature on productivity slowdown has provided a number of reasons and interpretations for the productivity paradox. However correcting for these explanations still leaves part of the phenomenon unexplained. We believe that our findings related to the compositionality of the productivity trend, structure of the technology flows among industries and presence of slowdown trends at both economy and science and research levels can complement the existing explanations and contribute to the reduction of the unexplained part of the productivity paradox.

A strong synergy between two ICT industries coupled with the exhaustion of established technological opportunities and the struggle to find new ones creates an instantiation of the dynamics described in Chapter 2. In Chapter 5, zooming on the software-hardware tandem illustrates the dynamic interrelatedness between the two industries and how disruption or stagnation in one can hinder or distract progress in the other. The main contribution of Chapter 5 is in highlighting the technological importance of the software industry for the product design in the semiconductor industry, that was downplayed; the strength of the software-hardware connection became evident only recently because of disruption in the software's technological trajectory brought by novel AI techniques as they represent a computation logic different from the currently established one. This proved inefficient the dominant design in the semiconductor industry and exposed an additional technological criterion of hardware performance, namely the heterogeneity of computation supported on a chip. Chapter 5 introduces a model of product selection based on a set of three criteria of chip's performance: classical processing speed, energy efficiency and previously obscured heterogeneity. Depending on consumers' preferences, there is a certain rate of substitution between, for example, processing speed and heterogeneity of computation i.e. reduction in processing speed can be compensated with increase in variety of software supported on a chip and vice versa. We conclude that the future of the semiconductor industry will be shaped by the interplay of the distribution of demand's preferences among the three technological characteristics and the available alternatives of chips resulting from production capabilities as well as strategic decisions of chipmakers. Though previously considered in the literature, the software-hardware bond was largely described through a price and/or quantity relationship, while the technological aspect has been modeled mainly operationalising the compatibility concept. The novelty of Chapter 5 consists in modeling consumers' choice based on explicit technological characteristics combined from both industries. If the importance of heterogeneity is valued by consumers more than processing speed, the outcome might take form of a platform chip; if the situation is the opposite, the semiconductor industry might fragment into smaller niches, each with a specialised chip.

In sum, at different levels of analysis, Chapters 2 and 5 stress the role tech-

nological factors play in producing an observed economic outcome. Nevertheless, both studies contain the systemic aspect as well: whether it is an economy-wide persistent input-output structure of technological knowledge among industries or its particular link between two industries of interest.

Chapters 3 and 4 assign a more prominent role to the system aspect and study, respectively, the structure and properties of the ICT cluster and the systemic nature of AI. In Chapter 3, I distinguish between 13 large classes of ICT and estimate their pervasiveness, technological proximity and commercial co-application. The first novel contribution of Chapter 3 is the proposed indicator of application relatedness based on established connections between industries and ICTs; in conjunction with technological relatedness, the two indicators create a framework that can be used by policy-makers to identify related markets that rely on the same technologies even across industrial boundaries, as well as technologies related through the same industries-applications. This more fine-grained analysis is possible due to the differentiation of the ICT cluster that is usually considered as a monolith; this represents the second novelty of the study. The relatedness analysis showed that the majority of ICT technologies are related through application rather than being proximate in the technological knowledge space; potentially, the technological trajectory of each ICT is influenced by another ICT not directly but through their common application. The generalised trend of decreasing technological proximity among ICTs over time might also indicate deepening of technological knowledge. Such divergence might be the result of strategic specialisation or a misled path taken because of myopic decisions technology development. The latter might be a very probable case, as ICTs with non-technical applications tend to give in to the demand's needs which in turn are driven by obviousness and commercial value and not technical superiority. For example, the case of software and hardware discussed in Chapter 5 might be an illustration of this: for decades, the "one-size-fits-all" remedy of the semiconductor industry was to increase processing speed to aggregate heterogeneous demand as much as possible under the dominant design of the von Neumann architecture instead of paying a special attention to the technical needs of the software industry. Even AI, that existed in different forms since 1950s, on its own didn't gain much attention from the semiconductor industry. Only when the commercial value of scaled up

neural-network-based AI algorithms started surging since around 2010s, the semiconductor industry began experimenting at scale with novel product architectures driven by technical criteria; this led to the current situation of fierce race after a new dominant design.

In general, a lot of academic discussions are devoted to AI as the latest wave of ICT with large transformative potential. Despite its quite recent debut on the commercial stage, AI is already considered as a pervasive technology and has been labelled a GPT. As the very beginning is a very good place to begin, in studying AI we started with understanding the nature of this emerging technology. Undoubtedly, AI deserves a special status, however not every influential technology is a GPT. Chapter 4 is an in-depth comparative analysis of two theoretical frameworks applied to AI: General Purpose Technology and Large Technical System. The relevance of the comparison can be seen as both frameworks are devoted to influential technologies however of qualitatively different nature. In the first part of Chapter 4, we map GPT definitional characteristics and effects on AI analysing mechanisms of technological dynamism, innovation inducement, uniqueness and pervasiveness. The results of analysis conducted in Chapter 3 serve here to provide the picture of scale and scope of AI's diffusion not only in absolute terms but also in comparison with other ICT technologies. Pooled together with the results of other studies and surveys, we demonstrate that AI displays a shallow penetration in the economic structure: only a small fraction within industries, firms, occupation and tasks is affected by AI. The status of newly born technology might raise a valid concern that AI can evolve into a GPT over time. However, the problem of GPT classification of AI might not disappear with time because AI is a system technology and not a stand-alone artefact; therefore, its final stage of evolution is likely to be an infrastructure, akin to the Internet. This is why in the second part of Chapter 4 we apply the LTS framework to AI and evaluate its goodness of fit. The result of the analysis shows that the LTS framework captures features and stylised facts about AI better than does the GPT framework. To demonstrate these differences we discuss policy and strategy implications derived based on the LTS framework that might be flattened or overlooked under the GPT perspective.

In sum, Chapters 3 and 4 investigate the AI technology system nested inside the even larger technology system of ICT. The omnipresent technological connections among this colossal ICT multiplex might remain idle and resilient to economically driven inefficiencies. These inefficiencies might accumulate or can be fixed but if a shock arrives, for example, in a form of radical technological breakthroughs, can resurface and end the life of an established technological trajectory. This is why, while considering the scope and scale of impact a system can generate, it is important to have a complete picture of that system accounting for both the economic and technological dimensions that bind it together.

6.2 Further Research Avenues

The constructed representation of the ICT technology system contains AI as a distinct subsystem. However, as the technology is in its growth phase, the markets and industries for its respective components – data, software, hardware – either emerge as well or, if existed before, experience serious transformations as the one analysed in Chapter 5. In Chapter 3, it has been shown that within the ICT system, AI developed linkages with technologies such as High speed computing and Image and sound technology, respectively representing hardware and data domains, at least partially. Given the quite aggregated level of analysis, some early, yet less intensive connections of AI could remain in the shadow of these primary ones. In favour of more fine-grained (but not less systemic) level of analysis says the fact of a distinctive connection between AI and the group of Other ICTs, hinting at some unclassified technologies that yet relate to AI. Provided that AI's origins as a technology lie beyond computer science and electrical engineering and span to cybernetics, logic, neurobiology, psychology, etc., the compositionality of AI in terms of knowledge combinations is probably diverse. Lastly, accounting for the alleged infrastructural nature of AI investigated in Chapter 4, AI being superimposed on existing digital infrastructure through system-level substitution might exhibit a pattern in acquiring a particular kind of tasks from the production processes and systems it is build upon.

The premises outlined above when brought together draw a prospective

research avenue: the study of the formation of the AI industry through the processes of technological convergence and upstreaming. Technological convergence is "the process by which different industries come to share similar technological base" (Gambardella and Torrisi, 1998). This happens as the industries involved in this process develop, within their production processes, a common set of tasks and functions to perform; examples range from processing raw materials to elaborating data. Initially, these tasks are executed in a compartmentalised manner by narrow, "unskilled" machines; these are technologies that are industry-dedicated, and thus used only within the boundaries of each industry. Over time, as actors learn to use and improve the technologies, it becomes clear that the tasks shared by different industries can be executed effectively by a common set of technologies, by virtue of some generic features or principles these technologies embody. As technological convergence takes place, the set of technologies common to all industries become increasingly autonomous and integrated and able to provide a pool of functions to a range of user industries; thus, "skilled" machines emerge. Through technological convergence, the relevant know-how to produce the technologies is progressively pooled and stored in a newborn industry; this process influences the dynamics of market structures, innovation incentives, and strategic decision making.

Currently, the process of technological convergence appears to be at work in many domains related to business information services and AI in particular. For example: (i) ready-made AI algorithms that are adaptable to different uses (and are based, for instance, on transfer learning techniques) are increasingly developed by specialised companies pushing for a machine learning as a service (MLaaS) business model; (ii) business automation (e.g. in human resources management) is entrusted to third-party providers of AI-solutions, as a next step after companies started to rely on software-as-a-service business models; (iii) warehouse logistics solutions are implemented by robotics firms rather than developed in-house by manufacturers; (iv) corporate strategy increasingly relies on data analysis conducted on external cloud computing platforms (such as Amazon Web Services), serving an heterogeneous array of customers and acting as data warehouses; (v) the design of systems, from products to cities, is explored and refined using algorithm-driven design and 'digital twins' hosted on dedicated plat-

forms and updated in real time through the collection of fine-grained data from sensors and internet-of-things devices , and; (vi) scientific discovery is externalised to automated systems performing brute-force screening and recognition of patterns hidden in the data.

Research in this direction will provide an alternative, less-explored view of the mechanisms of interaction between industry and technology dynamics. Compared with the prevalent literature on automation and employment, in this research the perspective is turned upside down: it is not only work experiencing changes due to technological progress, but technology itself experiencing changes and getting skilled through technological convergence, and in turn impacting industrial organisation, firms' strategies, and employment patterns. The novelty resides in shifting the focus of analysis from skilled (or unskilled) workers to skilled machines and to propose a viewpoint centred on the transformations resulting from technological convergence. This is relevant for industrial organisation, as it will track how the un-bundling and re-bundling of economic functions gives birth to new industries and what that entails, for instance, in terms of market structure and power; it is relevant for firms' strategies, as companies exploit the opportunities of a new industry's emergence to decide which and how many markets to serve, and how to scale and tailor their products; finally, it is relevant to employment patterns, as a new-born industry upstreaming around skilled machines requires specific workers profiles and will therefore impact labour demand and supply.

Bibliography

Abramovitz, M. (1989). *Thinking about growth: And other essays on economic growth and welfare*. Cambridge University Press.

Acemoglu, D., Akcigit, U., and Kerr, W. R. (2016). Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of labor economics*, 4:1043–1171.

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2020). Ai and jobs: Evidence from online vacancies. Technical report, National Bureau of Economic Research.

Acemoglu, D., Dorn, D., Hanson, G. H., Price, B., et al. (2014). Return of the solow paradox? it, productivity, and employment in us manufacturing. *American Economic Review*, 104(5):394–99.

Acemoglu, D. and Restrepo, P. (2017). Secular stagnation? the effect of aging on economic growth in the age of automation. *American Economic Review*, 107(5):174–79.

Adams, P., Fontana, R., and Malerba, F. (2013). The magnitude of innovation by demand in a sectoral system: The role of industrial users in semiconductors. *Research Policy*, 42(1):1–14.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728.

Agrawal, A., Gans, J., and Goldfarb, A. (2017). *What to expect from artificial intelligence*. MIT Sloan Management Review.

Agrawal, A., Gans, J., and Goldfarb, A. (2018a). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., and Goldfarb, A. (2019a). Economic policy for artificial intelligence. *Innovation Policy and the Economy*, 19(1):139–159.

Agrawal, A., Gans, J., and Goldfarb, A. (2019b). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019c). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Agrawal, A., McHale, J., and Oettl, A. (2018b). Finding needles in haystacks: Artificial intelligence and recombinant growth. Technical report, National Bureau of Economic Research.

Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, pages 483–485. ACM.

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., and Sutskever, I. (2019). Ai and compute.

Anderson, S. (2012). Advertising on the internet. In *The Oxford handbook of the digital economy*. Oxford University Press Oxford, UK.

Andrews, D., Criscuolo, C., and Gal, P. (2016). The global productivity slowdown, technology divergence and public policy: a firm level perspective. *Brookings Institution Hutchins Center Working Paper*, (24).

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors*, pages 609–626. Princeton University Press.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The economic journal*, 99(394):116–131.

Asaro, P. (2019). What is an artificial intelligence arms race anyway. *ISJLP*, 15:45.

ASML Holding, N. (2019). Financial statements us gaap q2 2019.

Auth, C., Allen, C., Blattner, A., Bergstrom, D., Brazier, M., Bost, M., Buehler, M., Chikarmane, V., Ghani, T., Glassman, T., et al. (2012). A 22nm high performance and low-power cmos technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density mim capacitors. In *2012 Symposium on VLSI Technology (VLSIT)*, pages 131–132. IEEE.

Autor, D. H., Katz, L. F., and Krueger, A. B. (1998). Computing inequality: have computers changed the labor market? *The Quarterly journal of economics*, 113(4):1169–1213.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729):59–64.

Bakos, Y. and Brynjolfsson, E. (2001). Aggregation and disaggregation of information goods: Implications for bundling, site licensing, and micropayment systems. In *Lectures in E-Commerce*, pages 103–122. Springer.

Baldwin, C. Y. and Clark, K. B. (2000). Design rules, volume 1: The power of modularity. *MIT Press Books*, 1.

Balland, P.-A., Boschma, R., Crespo, J., and Rigby, D. L. (2019). Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9):1252–1268.

Bartelsman, E. J. (2010). Searching for the sources of productivity from macro to micro and back. *Industrial and Corporate Change*, 19(6):1891–1917.

Basu, S. and Fernald, J. (2007). Information and communications technology as a general-purpose technology: Evidence from us industry data. *German Economic Review*, 8(2):146–173.

Basu, S., Fernald, J. G., Oulton, N., and Srinivasan, S. (2003). The case of the missing productivity growth, or does information technology explain why productivity accelerated in the united states but not in the united kingdom? *NBER macroeconomics annual*, 18:9–63.

Baumol, W. J. (2012). *The cost disease: Why computers get cheaper and health care doesn't*. Yale university press.

Beckhusen, J. (2016). Occupations in information technology. Technical report, US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Bekar, C., Carlaw, K., and Lipsey, R. (2018). General purpose technologies in theory, application and controversy: a review. *Journal of Evolutionary Economics*, 28(5):1005–1033.

Belleflamme, P. and Peitz, M. (2018). Platforms and network effects. In *Handbook of Game Theory and Industrial Organization, Volume II*. Edward Elgar Publishing.

Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., and Yehudayoff, A. (2019). Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48.

Beniger, J. R. (1986). The control revolution: technological and economic origins of the information society.

Bergemann, D. and Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11:85–107.

Bernstein, J. I. and Nadiri, M. I. (1989). Research and development and intra-industry spillovers: an empirical application of dynamic duality. *The Review of Economic Studies*, 56(2):249–267.

Bianchini, S., Müller, M., and Pelletier, P. (2020). Deep learning in science. *arXiv preprint arXiv:2009.01575*.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2017). Are ideas getting harder to find? Technical report, National Bureau of Economic Research.

Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press.

Bogetoft, P. and Otto, L. (2010). *Benchmarking with DEA, SFA, and R*, volume 157. Springer Science & Business Media.

Borkar, S. and Chien, A. A. (2011). The future of microprocessors. *Communications of the ACM*, 54(5):67–77.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT press.

Breschi, S., Lissoni, F., and Malerba, F. (2003). Knowledge-relatedness in firm technological diversification. *Research policy*, 32(1):69–87.

Bresnahan, T. (2019a). Artificial intelligence technologies and aggregate growth prospects.

Bresnahan, T. and Yin, P.-L. (2010). Reallocating innovative resources around growth bottlenecks. *Industrial and Corporate Change*, 19(5):1589–1627.

Bresnahan, T. and Yin, P.-L. (2017). Adoption of new information and communications technologies in the workplace today. *Innovation policy and the economy*, 17(1):95–124.

Bresnahan, T. F. (2019b). Technological change in ict in light of ideas first learned about the machine tool industry. *Industrial and Corporate Change*, 28(2):331–349.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The quarterly journal of economics*, 117(1):339–376.

Bresnahan, T. F., Davis, J. P., and Yin, P.-L. (2014). Economic value creation in mobile applications. In *The changing frontier: Rethinking science and innovation policy*, pages 233–286. University of Chicago Press.

Bresnahan, T. F. and Greenstein, S. (1999). Technological competition and the structure of the computer industry. *The Journal of Industrial Economics*, 47(1):1–40.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108.

Brodie, M. L. (1989). Future intelligent information systems: Ai and database technologies working together. In *Readings in artificial intelligence and databases*, pages 623–641. Elsevier.

Brown, C. and Linden, G. (2011). *Chips and change: how crisis reshapes the semiconductor industry*. MIT Press.

Brynjolfsson, E. and Hitt, L. (1995). Information technology as a factor of production: The role of differences among firms. *Economics of Innovation and New technology*, 3(3-4):183–200.

Brynjolfsson, E. and Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic perspectives*, 14(4):23–48.

Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Brynjolfsson, E., Rock, D., and Syverson, C. (2019). Artificial intelligence and the modern productivity paradox. *The Economics of Artificial Intelligence: An Agenda*, page 23.

Brynjolfsson, E., Rock, D., and Syverson, C. (2021). The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–72.

Burgelman, R. A. (2002). Strategy as vector and the inertia of coevolutionary lock-in. *Administrative science quarterly*, 47(2):325–357.

Byrne, D., Corrado, C., and Sichel, D. E. (2018). The rise of cloud computing: minding your p's, q's and k's. Technical report, National Bureau of Economic Research.

Byrne, D., Oliner, S., and Sichel, D. (2017). Prices of high-tech products, mismeasurement, and pace of innovation. Technical report, National Bureau of Economic Research.

Cabral, L. (2018). Standing on the shoulders of dwarfs: Dominant firms and innovation incentives.

Calo, R. (2014). The case for a federal robotics commission. *Brookings Institution. Brookings*.

Cantner, U. (2016). Foundations of economic change—an extended schumpeterian approach. *Journal of Evolutionary Economics*, 26(4):701–736.

Cantner, U. and Krüger, J. J. (2004). Geroski's stylized facts and mobility of large german manufacturing firms. *Review of Industrial Organization*, 24(3):267–283.

Cantner, U. and Krüger, J. J. (2008). Micro-heterogeneity and aggregate productivity development in the german manufacturing sector. *Journal of Evolutionary Economics*, 18(2):119–133.

Cantner, U., Savin, I., and Vannuccini, S. (2016). Replicator dynamics in value chains: explaining some puzzles of market selection. Technical report, Working Paper Series in Economics, Karlsruher Institut für Technologie (KIT).

Cantner, U. and Vannuccini, S. (2012). A new view of general purpose technologies. Technical report, Jena Economic Research Papers.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2020). How to talk when a machine is listening: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research.

Carr, N. G. (2003). It doesn't matter. *Educause Review*, 38:24–38.

Castaldi, C. (2009). The relative weight of manufacturing and services in europe: An innovation perspective. *Technological Forecasting and Social Change*, 76(6):709–722.

Castellacci, F. (2008). Technological paradigms, regimes and trajectories: Manufacturing and service industries in a new taxonomy of sectoral patterns of innovation. *Research Policy*, 37(6-7):978–994.

Castellacci, F. (2010). Structural change and the growth of industrial sectors: Empirical test of a gpt model. *Review of income and wealth*, 56(3):449–482.

Cette, G., Fernald, J., and Mojon, B. (2016). The pre-great recession slowdown in productivity. *European Economic Review*, 88:3–20.

Chen, Y.-H., Yang, T.-J., Emer, J., and Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*.

Choi, J. (2018). The rise of 3d printing and the role of user firms in the us: evidence from patent data. *Technology Analysis & Strategic Management*, 30(10):1195–1209.

Chou, C.-f. and Shy, O. (1993). Partial compatibility and supporting services. *Economics letters*, 41(2):193–197.

Church, J. and Gandal, N. (1992). Network effects, software provision, and standardization. *The journal of industrial economics*, pages 85–103.

Church, J. and Gandal, N. (2005). Platform competition in telecommunications. In *The Handbook of Telecommunications*, volume 2, pages 119–155. Elsevier.

Clark, G. (2016). Winter is coming: Robert gordon and the future of economic growth. *The American Economic Review*, 106(5):68–71.

Cockburn, I. M., Henderson, R., and Stern, S. (2019). The impact of artificial intelligence on innovation. *The Economics of Artificial Intelligence: An Agenda*, page 115.

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*, 1:129–213.

Corrado, C., Hulten, C., and Sichel, D. (2009). Intangible capital and us economic growth. *Review of income and wealth*, 55(3):661–685.

Corrado, C., Lengermann, P., Beaulieu, J. J., and Bartelsman, E. J. (2007). Sectoral productivity in the united states: Recent developments and the role of it. *German Economic Review*, 8(2):188–210.

Couldry, N. and Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4):336–349.

Crafts, N. (2016). The rise and fall of american growth: Exploring the numbers. *The American Economic Review*, 106(5):57–60.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., et al. (2019). Ai now 2019 report. *New York, NY: AI Now Institute*.

Cusumano, M. A. and Gawer, A. (2002). The elements of platform leadership. *MIT Sloan management review*, 43(3):51.

Dabla-Norris, M. E., Guo, M. S., Haksar, M. V., Kim, M., Kochhar, M. K., Wiseman, K., and Zdzienicka, A. (2015). *The new normal: A sector-level perspective on productivity trends in advanced economies*. International Monetary Fund.

Dartnall, T. (1994). *Introduction: On Having a Mind of Your Own*, pages 29–42. Springer Netherlands.

David, P. A. (2007). Path dependence: a foundational concept for historical social science. *Cliometrica*, 1(2):91–114.

David, P. A. and Wright, G. (2003). General purpose technologies and surges in productivity. *The economic future in historical perspective*.

Davies, A. (1996). Innovation in large technical systems: the case of telecommunications. *Industrial and Corporate Change*, 5(4):1143–1180.

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99.

De Loecker, J. and Eeckhout, J. (2017). The rise of market power and the macroeconomic implications. Technical report, National Bureau of Economic Research.

Decker, R., Haltiwanger, J., Jarmin, R., and Miranda, J. (2014). The role of entrepreneurship in us job creation and economic dynamism. *The Journal of Economic Perspectives*, 28(3):3–24.

Dennard, R. H., Gaensslen, F. H., Rideout, V. L., Bassous, E., and LeBlanc, A. R. (1974). Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268.

Dewan, S. and Min, C.-k. (1997). The substitution of information technology for other factors of production: A firm level analysis. *Management science*, 43(12):1660–1675.

Diewert, W. E. and Fox, K. J. (1999). Can measurement error explain the productivity paradox? *The Canadian Journal of Economics/Revue canadienne d'Economique*, 32(2):251–280.

Doganoglu, T. and Wright, J. (2006). Multihoming and compatibility. *International Journal of Industrial Organization*, 24(1):45–67.

Domingos, P. and Lowd, D. (2019). Unifying logical and statistical ai with markov logic. *Communications of the ACM*, 62(7):74–83.

Dorn, D., Hanson, G. H., Pisano, G., Shu, P., et al. (2016). Foreign competition and domestic innovation: Evidence from us patents. Technical report, National Bureau of Economic Research.

Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy*, 11(3):147–162.

Dosi, G. and Nelson, R. R. (2010). Technical change and industrial dynamics as evolutionary processes. *Handbook of the Economics of Innovation*, 1:51–127.

Eckersley, P., Nasser, Y., et al. (2017). Eff ai progress measurement project.

Ekbia, H. R. and Nardi, B. A. (2017). *Heteromation, and other stories of computing and capitalism*. MIT Press.

Essletzbichler, J. (2015). Relatedness, industrial branching and technological cohesion in us metropolitan areas. *Regional Studies*, 49(5):752–766.

Ewertsson, L. and Ingelstam, L. (2004). Large technical systems: A multidisciplinary research tradition. In *Systems Approaches and Their Application*, pages 291–309. Springer.

Federico, G., Morton, F. S., and Shapiro, C. (2020). Antitrust and innovation: Welcoming and protecting disruption. *Innovation Policy and the Economy*, 20(1):125–190.

Feldman, M. P. and Yoon, J. W. (2012). An empirical test for general purpose technology: an examination of the cohen–boyer rdna technology. *Industrial and Corporate Change*, 21(2):249–275.

Fernald, J. G. (2015). Productivity and potential output before, during, and after the great recession. *NBER macroeconomics annual*, 29(1):1–51.

Fitchard, K. (2003). Crossing over: The journey to packets. *Telephony*, 244(16):40–40.

Flamm, K. (2019). Measuring moore's law: evidence from price, cost, and quality indexes. In *Measuring and Accounting for Innovation in the 21st Century*. University of Chicago Press.

Flueckiger, G. E. (1995). *Control, Information, and Technological Change*. Number 6. Springer Science & Business Media.

Foster, L., Haltiwanger, J. C., and Krizan, C. J. (2001). Aggregate productivity growth: lessons from microeconomic evidence. In *New developments in productivity analysis*, pages 303–372. University of Chicago Press.

Freeman, C. (1994). The economics of technical change. *Cambridge journal of economics*, 18(5):463–514.

Frenken, K., Van Oort, F., and Verburg, T. (2007). Related variety, unrelated variety and regional economic growth. *Regional studies*, 41(5):685–697.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, 79(3):733–772.

Galindo-Rueda, F. and Verger, F. (2016). Oecd taxonomy of economic activities based on r&d intensity.

Gambardella, A. and Torrisi, S. (1998). Does technological convergence imply convergence in markets? evidence from the electronics industry. *Research policy*, 27(5):445–463.

Gandal, N. (2002). Compatibility, standardization, and network effects: Some policy implications. *Oxford Review of Economic Policy*, 18(1):80–91.

Gawer, A. and Henderson, R. (2007). Platform owner entry and innovation in complementary markets: Evidence from intel. *Journal of Economics & Management Strategy*, 16(1):1–34.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gökalp, I. (1992). On the analysis of large technical systems. *Science, Technology, and Human Values*, pages 57–78.

Goldfarb, A., Gans, J., and Agrawal, A. (2019). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Goldschlag, N. and Tabarrok, A. T. (2014). Is regulation to blame for the decline in american entrepreneurship?

Gonzalez, J. and Ponemon, L. (2019). 2019 intangible assets financial statement impact comparison report. Technical report, Ponemon Institute LLC.

Gordon, R. J. (2016). Perspectives on the rise and fall of american growth. *Education*, 2(2.5):3.

Graf, H. (2012). Inventor networks in emerging key technologies: information technology vs. semiconductors. *Journal of Evolutionary Economics*, 3(22):459–480.

Graham, S. J. and Iacopetta, M. (2014). Nanotechnology and the emergence of a general purpose technology. *Annals of Economics and Statistics/Annales D'Économie et de Statistique*, (115/116):25–55.

Greenstein, S. (2019). Digital infrastructure. In *Economics of Infrastructure Investment*. University of Chicago Press.

Greenstein, S. (2020). The basic economics of internet infrastructure. *Journal of Economic Perspectives*, 34(2):192–214.

Greenstein, S. M. and Spiller, P. T. (1996). Estimating the welfare effects of digital infrastructure. Technical report, National Bureau of Economic Research.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica, Journal of the Econometric Society*, pages 501–522.

Guerrieri, P. and Padoan, P. C. (2007). *Modelling ICT as a general purpose technology*. The College of Europe.

Hall, B. H. and Trajtenberg, M. (2006). Uncovering general purpose technologies with patent data1. *New Frontiers in the Economics of Innovation and New Technology: Essays in Honour of Paul A. David*, page 389.

Harberger, A. C. (1998). A vision of the growth process. *The American Economic Review*, 88(1):1–32.

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.

Helpman, E. and Trajtenberg, M. (1994). A time to sow and a time to reap: growth based on general purpose technologies. Technical report, National Bureau of Economic Research.

Henderson, R. M. and Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing. *Administrative science quarterly*, 35(1):9–30.

Hennessy, J. L. and Patterson, D. A. (2011). *Computer architecture: a quantitative approach*. Elsevier.

Hernandez, D. and Brown, T. B. (2020). Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv:2005.04305*.

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., et al. (2018). The principle of relatedness. In *International conference on complex systems*, pages 451–457. Springer.

Hidalgo, C. A. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575.

Holm, J. R. (2014). The significance of structural transformation to productivity growth. *Journal of Evolutionary Economics*, 24(5):1009–1036.

Hooker, S. (2020). The hardware lottery. *arXiv preprint arXiv:2009.06489*.

Hughes, T. P. (1983). Networks of power: Electric supply systems in the us, england and germany, 1880-1930. *Baltimore: Johns Hopkins University*.

Hughes, T. P. et al. (1987). The evolution of large technological systems. *The social construction of technological systems: New directions in the sociology and history of technology*, 82.

Iansiti, M. and Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Press.

Inaba, T. and Squicciarini, M. (2017). Ict: A new taxonomy based on the international patent classification. *OECD iLibrary*.

Isdahl, R. and Gundersen, O. E. (2019). Out-of-the-box reproducibility: A survey of machine learning platforms. In *2019 15th international conference on eScience (eScience)*, pages 86–95. IEEE.

Jacobides, M. G., Cennamo, C., and Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, 39(8):2255–2276.

Jian, L., MacKie-Mason, J., Chiao, B., Levchenko, A., Zellner, A., Kmenta, J., Dreze, J., and Oberhofer, W. (2012). Incentive-centered design for user-contributed content. *The Oxford Handbook of the Digital Economy*, page 399.

Joerges, B. (1988). *Large technical systems: the concept and the issues*. Wiss.-zentrum für Sozialforschung.

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58.

Jorgenson, D. W., Ho, M. S., and Stiroh, K. J. (2003). Growth of us industries and investments in information technology and higher education. *Economic Systems Research*, 15(3):279–325.

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., et al. (2017). In-datacenter performance analysis of a tensor processing unit. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pages 1–12. IEEE.

Jovanovic, B. and Rousseau, P. L. (2005). General purpose technologies. In *Handbook of economic growth*, volume 1, pages 1181–1224. Elsevier.

Kapoor, R. (2013). Persistence of integration in the face of specialization: How firms navigated the winds of disintegration and shaped the architecture of the semiconductor industry. *Organization Science*, 24(4):1195–1213.

Klinger, J., Mateos-Garcia, J., and Stathoulopoulos, K. (2020). A narrowing of ai research? *arXiv preprint arXiv:2009.10385*.

Kortum, S. and Putnam, J. (1997). Assigning patents to industries: tests of the yale technology concordance. *Economic Systems Research*, 9(2):161–176.

Koutroumpis, P., Leiponen, A., and Thomas, L. (2020a). Markets for data. *Industrial and Corporate Change*, 29(3).

Koutroumpis, P., Leiponen, A., and Thomas, L. D. (2020b). Digital instruments as invention machines. *Communications of the ACM*.

Kreuchauff, F. and Teichert, N. (2014). Nanotechnology as general purpose technology. Technical report, KIT Working Paper Series in Economics.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kurz, H. D., Schütz, M., Strohmaier, R., and Zilian, S. (2018). Riding a new wave of innovations. *Wirtschaft und Gesellschaft-WuG*, 44(4):545–583.

Laitenberger, J. (2017). Eu competition law in innovation and digital markets: fairness and the consumer welfare perspective. In *MLex / Hogan Lovells event*.

Langlois, R. N. and Steinmueller, W. E. (2000). Strategy and circumstance: The response of american firms to japanese competition in semiconductors, 1980–1995. *Strategic Management Journal*, 21(10-11):1163–1173.

Lee, E. A. (2002). Embedded software. In *Advances in computers*, volume 56, pages 55–95. Elsevier.

Lee, E. A. and Neuendorffer, S. (2005). Concurrent models of computation for embedded software. *IEE Proceedings-Computers and Digital Techniques*, 152(2):239–250.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Lehrer, M., Banerjee, P. M., and Wang, I. K. (2016). The improvement trajectory of pcr dna replication and erp software as general purpose technologies: an exploratory study of 'anchor technologies'. *Technology Analysis & Strategic Management*, 28(3):290–304.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Lie, S. (2019). Wafer scale deep learning. In *IEEE Hot Chips 31. Symposium (HCS 2019)*.

Lybbert, T. J. and Zolas, N. J. (2014). Getting patents and economic data to speak to each other: An "algorithmic links with probabilities" approach for joint analyses of patenting and economic activity. *Research Policy*, 43(3):530–542.

Malerba, F. (2002). Sectoral systems of innovation and production. *Research policy*, 31(2):247–264.

Malerba, F., Nelson, R., Orsenigo, L., and Winter, S. (2008). Vertical integration and disintegration of computer firms: a history-friendly model of the coevolution of the computer and semiconductor industries. *Industrial and Corporate Change*, 17(2):197–231.

Malerba, F. and Orsenigo, L. (1997). Technological regimes and sectoral patterns of innovative activities. *Industrial and corporate change*, 6(1):83–118.

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Maskell, P. and Malmberg, A. (1999). Localised learning and industrial competitiveness. *Cambridge journal of economics*, 23(2):167–185.

Mayntz, R. and Hughes, T. (1988). *The Evolution of Large Technical Systems*. Campus.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). Proposal for the 1956 dartmouth summer research project on artificial intelligence, dartmouth college.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Melitz, M. J. and Polanec, S. (2015). Dynamic olley-pakes productivity decomposition with entry and exit. *The Rand journal of economics*, 46(2):362–375.

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Imam, N., Guo, C., Nakamura, Y., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673.

Metcalfe, J. S. (1994). Competition, fisher's principle and increasing returns in the selection process. *Journal of Evolutionary Economics*, 4(4):327–346.

Metcalfe, J. S. (2003). Industrial growth and the theory of retardation. *Revue économique*, 54(2):407–431.

Metcalfe, J. S. and Ramlogan, R. (2006). Creative destruction and the measurement of productivity change. *Revue de l'OFCE*, (5):373–397.

Michaels, G., Natraj, A., and Van Reenen, J. (2014). Has ict polarized skill demand? evidence from eleven countries over twenty-five years. *Review of Economics and Statistics*, 96(1):60–77.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin UK.

Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684.

Mokyr, J. (1990). Punctuated equilibria and technological progress. *The American Economic Review*, 80(2):350–354.

Mokyr, J. (2014). Secular stagnation? not in your life. *Secular Stagnation: Facts, Causes and Cures*, 83.

Moore, T. and Anderson, R. (2012). Internet security. In *The Oxford Handbook of the Digital Economy*, pages 572–99. Oxford University Press Oxford.

Morton, F. S. and Dinielli, D. (2020). Roadmap for an antitrust case against facebook. Technical report, Omidyar Network.

Nagy, B., Farmer, J. D., Bui, Q. M., and Trancik, J. E. (2013). Statistical basis for predicting technological progress. *PloS one*, 8(2):e52669.

Napoletano, M., Roventini, A., and Sapio, S. (2006). Modelling smooth and uneven cross-sectoral growth patterns: an identification problem. *Econ Bull*, 15(6):1–8.

Neffke, F., Hartog, M., Boschma, R., and Henning, M. (2018). Agents of structural change: The role of firms and entrepreneurs in regional diversification. *Economic Geography*, 94(1):23–48.

Neffke, F., Henning, M., and Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic geography*, 87(3):237–265.

Newell, A. and Simon, H. (1956). The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79.

Nightingale, P., Brady, T., Davies, A., and Hall, J. (2003). Capacity utilization revisited: software, control and the growth of large technical systems. *Industrial and Corporate Change*, 12(3):477–517.

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press.

OECD (2019). Vectors of digital transformation. *OECD Digital Economy Papers*, (273).

Olley, G. S. and Pakes, A. (1992). The dynamics of productivity in the telecommunications equipment industry. Technical report, National Bureau of Economic Research.

Park, K., Seamans, R., and Zhu, F. (2018). Multi-homing and platform strategies: Historical evidence from the us newspaper industry. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (18-032).

Pavitt, K. (1984). Sectoral patterns of technical change: towards a taxonomy and a theory. *Research policy*, 13(6):343–373.

Perez, C. (2010). Technological revolutions and techno-economic paradigms. *Cambridge journal of economics*, 34(1):185–202.

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., and Niebles, J. C. (2019). The ai index 2019 annual report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*.

Prat, A. and Valletti, T. M. (2019). Attention oligopoly. *Available at SSRN 3197930*.

Prytkova, E. (2021). Ict's wide web: a system-level analysis of ict's industrial diffusion with algorithmic links. *Available at SSRN*.

Prytkova, E. and Vannuccini, S. (2020). On the basis of brain: Neural-network-inspired change in general purpose chips.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880. ACM.

Reinganum, J. F. (1989). The timing of innovation: Research, development, and diffusion. *Handbook of industrial organization*, 1:849–908.

Roach, S. S. (1987). America's technology dilemma: A profile of the information economy: Morgan stanley special economic study. *Morgan Stanley*.

Robinson, D. K. and Mazzucato, M. (2019). The evolution of mission-oriented policies: Exploring changing market creating policies in the us and european space sector. *Research Policy*, 48(4):936–948.

Rosenberg, N. (1963). Technological change in the machine tool industry, 1840-1910. *Journal of economic history*, pages 414–443.

Rosenberg, N. (1969). The direction of technological change: inducement mechanisms and focusing devices. *Economic development and cultural change*, 18(1, Part 1):1–24.

Rosenberg, N. (1972). Factors affecting the diffusion of technology. *Explorations in economic history*, 10(1):3.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. VIKING.

Sahal, D. (1985). Technological guideposts and innovation avenues. *Research policy*, 14(2):61–82.

Savona, M. (2019). The value of data: Towards a framework to redistribute it. *SWPS*.

Schmoch, U., Laville, F., Patel, P., and Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*.

Schot, J. and Steinmueller, W. E. (2018). Three frames for innovation policy: R&d, systems of innovation and transformative change. *Research Policy*, 47(9):1554–1567.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Shao, Y. S., Clemons, J., Venkatesan, R., Zimmer, B., Fojtik, M., Jiang, N., Keller, B., Klinefelter, A., Pinckney, N., Raina, P., et al. (2019). Simba: Scaling deep-learning inference with multi-chip-module-based architecture. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 14–27.

Shapiro, C., Carl, S., Varian, H. R., et al. (1998). *Information rules: a strategic guide to the network economy*. Harvard Business Press.

Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., et al. (2017). Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441.

Shy, O. (2011). A short survey of network economics. *Review of Industrial Organization*, 38(2):119–149.

Simcoe, T. and Watson, J. (2019). Forking, fragmentation, and splintering. *Strategy Science*, 4(4):283–297.

Slonim, N., Bilu, Y., and Alzate, C. (2021). An autonomous debating system. *Nature*, 379–384.

Snowden, E. (2019). *Permanent record*. Macmillan.

Solow, R. M. (1987). New york times book review. *July*, 12(1987):36.

Sovacool, B. K., Lovell, K., and Ting, M. B. (2018). Reconfiguration, contestation, and decline: conceptualizing mature large technical systems. *Science, Technology, & Human Values*, 43(6):1066–1097.

Spiekermann, M. (2019). Data marketplaces: Trends and monetisation of data goods. *Intereconomics*, 54(4):208–216.

Srinivasan, D. (2019). Why google dominates advertising markets. *SSRN*.

Steinmueller, W. E. (1992). The economics of flexible integrated circuit manufacturing technology. *Review of Industrial Organization*, 7(3-4):327–349.

Steinmueller, W. E. (1996). Technological infrastructure in information technology industries. In *Technological Infrastructure Policy*, pages 117–139. Springer.

Steinmueller, W. E. (2002). Knowledge-based economies and information and communication technologies. *International Social Science Journal*, 54(171):141–153.

Steinmueller, W. E. (2006). Learning in the knowledge-based economy: the future as viewed from the past. *New Frontiers in the Economics of Innovation and New Technology. Essays in Honour of Paul A. David. Cheltenham, UK: Edward Elgar*, pages 207–238.

Steinmueller, W. E. (2007). The economics of icts: Building blocks and implications. In *The Oxford handbook of information and communication technologies*.

Steinmueller, W. E. (2010). Economics of technology policy. In *Handbook of the Economics of Innovation*, volume 2, pages 1181–1218. Elsevier.

Strohmaier, R. and Rainer, A. (2016). Studying general purpose technologies in a multi-sector framework: The case of ict in denmark. *Structural Change and Economic Dynamics*, 36:34–49.

Strohmaier, R., Schuetz, M., and Vannuccini, S. (2019). A systemic perspective on socioeconomic transformation in the digital age. *Journal of Industrial and Business Economics*, 46(3):361–378.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv:1906.02243*.

Su, L. (2019). Delivering the future of high-performance computing with system, software and silicon co-optimization, keynote address at hot chips: A symposium on high performance chips, edition 31.

Suárez, F. F. and Utterback, J. M. (1995). Dominant designs and the survival of firms. *Strategic management journal*, 16(6):415–430.

Summers, L. H. (2015). Demand side secular stagnation. *The American Economic Review*, 105(5):60–65.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature*, 49(2):326–365.

Syverson, C. (2017). Challenges to mismeasurement explanations for the us productivity slowdown. *Journal of Economic Perspectives*, 31(2):165–86.

Sze, V., Chen, Y. H., Yang, T. J., and Emer, J. S. (2020). How to evaluate deep neural network processors: Tops/w (alone) considered harmful. *IEEE Solid-State Circuits Magazine*, 12(3):28–41.

Taddy, M. (2019). The technological elements of artificial intelligence. *The Economics of Artificial Intelligence: An Agenda*, page 61.

Takahashi, T. and Namiki, F. (2003). Three attempts at "de-wintelization": Japan's tron project, the us government's suits against wintel, and the entry of java and linux. *Research Policy*, 32(9):1589–1606.

Teulings, C. and Baldwin, R. (2014). *Secular stagnation: Facts, causes, and cures–a new Vox eBook*, volume 15. Voxeu.

Thoma, G. (2009). Striving for a large market: evidence from a general purpose technology in action. *Industrial and Corporate Change*, 18(1):107–138.

Thompson, N. and Spanuth, S. (2018). The decline of computers as a general purpose technology: Why deep learning and the end of moore's law are fragmenting computing. *SSRN 3287769*.

Timmer, M. P. and Van Ark, B. (2005). Does information and communication technology drive eu-us productivity growth differentials? *Oxford Economic Papers*, 57(4):693–716.

Trajtenberg, M. (2018). Ai as the next gpt: a political-economy perspective. Technical report, National Bureau of Economic Research.

Trajtenberg, M. (2019). Artificial intelligence as the next gpt. *The Economics of Artificial Intelligence: An Agenda*, page 175.

Tubaro, P., Casilli, A. A., and Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society*, 7(1):2053951720919776.

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society*, 2(1):230–265.

UN (2008). *International Standard Industrial Classification of All Economic Activities Revision 4 (ISIC rev.4)*. Number 4. United Nations Publications.

Van Ark, B., Inklaar, R., and McGuckin, R. H. (2003). Ict and productivity in europe and the united states where do the differences come from? *CESifo Economic Studies*, 49(3):295–318.

van der Vleuten, E. (2009). Large technical systems. *A Companion to the Philosophy of Technology*, pages 218–222.

Van Roy, V., Vertesy, D., and Damioli, G. (2020). Ai and robotics innovation. *Handbook of Labor, Human Resources and Population Economics*, pages 1–35.

Vannuccini, S. and Prytkova, E. (2020). Artificial intelligence's new clothes? from general purpose technology to large technical system. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.3704011*.

Veen, A. H. (1986). Dataflow machine architecture. *ACM Computing Surveys (CSUR)*, 18(4):365–396.

Verspagen, B. and De Loo, I. (1999). Technology spillovers between sectors. *Technological Forecasting and Social Change*, 60(3):215–235.

Vázquez, M., Henarejos, P., Pérez-Neira, A. I., Grechi, E., Voight, A., Gil, J. C., Pappalardo, I., Credico, F. D., and Lancellotti, R. M. (2020). On the use of ai for satellite communications.

Winter, M., Prusseit, S., and Gerhard, P. F. (2010). Hierarchical routing architectures in clustered 2d-mesh networks-on-chip. In *2010 International SoC Design Conference*, pages 388–391. IEEE.

WIPO (2019a). Patentscope artificial intelligence index. Technical report, WIPO.

WIPO, W. (2019b). Technology trends 2019: Artificial intelligence. *Geneva: World Intellectual Property Organization*.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., et al. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

Erklärung nach §4 Abs. 1 PromO

Hiermit erkläre ich,

1. dass mir die geltende Promotionsordung bekannt ist;
2. dass ich die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe;
3. dass ich bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes keine unzulässige Hilfe in Anspruch genommen habe;
4. dass ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
5. dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;
6. dass ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultät als Dissertation eingereicht habe.

Ort, Datum

Unterschrift