Combinatorial Inventions in Artificial Intelligence:

Empirical Evidence and Implications for Science, Technology, and Organizations

by

Jieshu Wang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2023 by the
Graduate Supervisory Committee:

Andrew Maynard, Chair
José Lobo
Katina Michael
Sébastien Motsch

ARIZONA STATE UNIVERSITY

August 2023

ABSTRACT

Artificial Intelligence (AI) is a rapidly advancing field with the potential to impact every aspect of society, including the inventive practices of science and technology. The creation of new ideas, devices, or methods, commonly known as inventions, is typically viewed as a process of combining existing knowledge. To understand how AI can transform scientific and technological inventions, it is essential to comprehend how such combinatorial inventions have emerged in the development of AI.

This dissertation aims to investigate three aspects of combinatorial inventions in AI using data-driven and network analysis methods. Firstly, how knowledge is combined to generate new scientific publications in AI; secondly, how technical components are combined to create new AI patents; and thirdly, how organizations create new AI inventions by integrating knowledge within organizational and industrial boundaries. Using an AI publication dataset of nearly 300,000 AI publications and an AI patent dataset of almost 260,000 AI patents granted by the United States Patent and Trademark Office (USPTO), this study found that scientific research related to AI is predominantly driven by combining existing knowledge in highly conventional ways, which also results in the most impactful publications. Similarly, incremental improvements and refinements that rely on existing knowledge rather than radically new ideas are the primary driver of AI patenting. Nonetheless, AI patents combining new components tend to disrupt citation networks and hence future inventive practices more than those that involve only existing components.

To examine AI organizations' inventive activities, an analytical framework called the Combinatorial Exploitation and Exploration (CEE) framework was developed to measure how much an organization accesses and discovers knowledge while working within organizational and industrial boundaries. With a dataset of nearly 500 AI organizations that have continuously contributed to AI technologies, the research

shows that AI organizations favor exploitative over exploratory inventions. However, local exploitation tends to peak within the first five years and remain stable, while exploratory inventions grow gradually over time.

Overall, this dissertation offers empirical evidence regarding how inventions in AI have emerged and provides insights into how combinatorial characteristics relate to AI inventions' quality. Additionally, the study offers tools to assess inventive outcomes and competence.

In memory of my father, who I wish could have seen this adventure.

ACKNOWLEDGEMENTS

My sincerest appreciation goes to those who have supported me throughout my doctoral journey. I am profoundly grateful to my advisor, Prof. Andrew Maynard, whose patience, guidance, and invaluable advice have been a driving force during these past five years. His encouragement to explore the technical aspects of AI has been instrumental in the success of my dissertation. Moreover, both he and his wife, Clare, have provided comfort and support during challenging times.

I wish to express my gratitude to Prof. José Lobo, whose unwavering faith in me allowed me to pursue projects that sparked my curiosity. His support has been essential in sustaining me throughout my Ph.D. journey, and I have acquired a wealth of knowledge and research skills under his mentorship. Working with him has truly been an honor.

I also extend my gratitude to my other committee members, Prof. Katina Michael and Prof. Sébastien Motsch, for sharing their invaluable insights and expertise. I am particularly grateful for Katina's support in attending the CSPO conference and her guidance on accessing research data. I am thankful to Prof. Erik Johnston, for his resourcefulness and genuine concern for my well-being. His trust allowed me to engage in various research projects and refine my research skills. His emphasis on prioritizing my physical and mental health has been invaluable.

My heartfelt appreciation goes to my loving and supportive husband, Bilal, who has accompanied me throughout this journey. His steadfast belief in me, technical expertise, and encouragement have significantly contributed to the success of my dissertation. I am truly blessed to have him as my life partner.

I am indebted to Wendi Taylor and Andra Williams for their assistance with paperwork, milestone procedures, and logistics. I would also like to thank the ASU library's librarians and data science experts, including Mr. Dan Stanton, Ms. Anali

Perry, Ms. JoAnn Mulvihill, and Dr. Michael Simeone, for their help with research resources. Additionally, I acknowledge Prof. Heather Ross, the current director of HSD, and Prof. David Guston for their support.

I offer special thanks to the professors from whom I have had the pleasure of learning, including Prof. Subbarao Kambhampati, Prof. Clark Miller, Prof. Jameson Wetmore, and Prof. Faheen Hussain. Their teachings on AI, socio-technical systems, and research methodology have been invaluable.

I am grateful for the opportunity to collaborate with professors such as Prof. Lauren Keeler, Dr. Michael Bernstein, and Prof. Shade Shutters on pressing issues like the future of aging in smart environments and technologies to mitigate climate change. My conversations with Prof. Debby Strumsky have been particularly inspiring and have shaped my research.

I extend my gratitude to my colleagues, including Dr. John Harlow, Julia Hsu, Prof. Myeong Lee, Prof. Susan Winter, and my classmates and friends, Pooja, Sangha, Vanya, Farah, Martin, Yiama, Max, Nicole, Elma, and Danielle. Their camaraderie and support have made this journey all the more enjoyable.

Lastly, as a student at ASU, I wish to recognize and honor the ancestral homelands upon which the Tempe campus resides. For centuries, this land has been home to various American Indian tribes, including the Akimel O'odham (Pima) and Pee Posh (Maricopa) peoples, who have maintained a deep and enduring connection to this place.

The completion of this dissertation would not have been possible without the collective support, guidance, and encouragement of all those mentioned. I will be forever grateful for their contributions to my academic and personal growth.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

APPENDIX Page

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Artificial Intelligence (AI) is a rapidly developing field that carries the promise of influencing almost every aspect of society, including the inventive practice of science and technology. Inventions, or the creation of new ideas, devices, or methods, have been conceptualized as a process of combining existing knowledge. In order to assess how AI can transform the process of scientific and technological inventions, it is crucial to understand how such combinatorial inventions have occurred in the development of AI. How are inventions within the domain of AI occurring? For instance, is it via the development of new technological functionalities and new scientific understandings, or is it mainly through combining existing concepts and methods or introducing radically new ideas? Answering these questions can inform an assessment of how AI will amplify or exceed human effectiveness and displace human labor. Invention in AI is also a venue to study the production of new scientific knowledge and technological change. Is knowledge creation in AI similar to previous episodes and inventive activities in general? Or is it the case that by dealing with the very nature of intelligence, reasoning, and learning, the creation of knowledge in AI marks a departure from historical patterns in the advancement of science and technology? Addressing these questions requires understanding the nature and underlying mechanisms of combinatorial inventions in AI and how they affect the quantity and quality of AI inventions. The nature and mechanisms of combinatorial inventions in AI are the overarching research question of this dissertation.

This dissertation investigates combinatorial inventions in the development of AI from three perspectives — the perspective of scientific research, the perspective of

technological invention, and the perspective of organizational inventions, which constitute the three main chapters of this dissertation (Chapters 2, 3, and 4). Specifically, this dissertation aims to address the following three aspects of combinatorial inventions in AI: (1) how knowledge is combined to create new scientific knowledge related to AI; (2) how technical components are combined to create new technologies related to AI; and (3) how organizations create new knowledge related to AI by searching and combining knowledge constrained by the organizational and industrial boundaries.

To address these questions, a series of datasets were constructed, including three primary novel datasets: an AI publication dataset, an AI patent dataset, and an AI organization dataset. The methodologies involved in this dissertation include bibliographic methods, scientometric methods, network analysis, and statistical analysis.

This research's investigations reveal that scientific research related to AI is driven by combining existing knowledge in highly conventional ways, and the resulting academic publications are also the most impactful. Similarly, in patents related to AI, incremental improvements that heavily rely on existing knowledge rather than radically new ideas are the primary driver of knowledge creation in AI patents. Like AI publications, AI patents that combine existing technologies tend to have higher citations. Nevertheless, AI patents combining new components tend to disrupt the knowledge network to a greater extent. By comparing scientific publications and patents related to AI, this dissertation find that novelty in AI patents is better rewarded by future citations than AI publications, implying that technological innovation has a larger appetite for novel ideas and combinations than scientific research. To examine AI organizations' inventive activities, an analytical framework was developed in this dissertation— the Combinatorial Exploitation and Exploration (CEE) framework — to quantify the extent an organization accesses existing knowledge and discovers new knowledge within the constraints of organizational and industrial boundaries. Using

the CEE framework, this research finds that AI organizations prefer exploitative over exploratory inventions. Nevertheless, locally exploitative search tends to peak within the first five years, while exploratory inventions tend to grow gradually if the organization remains active in AI. Moreover, the CEE framework is also demonstrated to be able to compare inventive competence and capacities intra-organizationally and inter-temporally. Several additional empirical examples are given to showcase other use cases, such as analyzing the geographical aspects of organizations' inventive behaviors.

1.1 Literature Review

1.1.1 Artificial Intelligence

AI methods, techniques, and procedures are already transforming medical diagnostics, transportation planning, manufacturing processes, data analysis, automated-decision-making, speech and optical pattern recognition, software design, and even legal analysis and investment decisions (National Research Council, 1997; Matheny et al., 2019a; World Economic Forum, 2018; Mozer et al., 2019; Correia and Reyes, 2020; Chang, 2020; Chalmers et al., 2021; Surden, 2019; National Academies of Sciences, Engineering, and Medicine, 2021). AI is expected to transform the workplace in much the same way the Industrial Revolution and electrification did (Frank et al., 2019a). Furthermore, deployment and mastery of AI are seen by governments as a cornerstone of national security and economic prosperity (State Council of the People's Republic of China, 2017; Eric and Work, 2021). "Machine-learning" and "deep learning," algorithmic and computational procedures able to learn from and recognize patterns in copious amounts of data, are now able to surpass human capacities in important problem areas (LeCun et al., 2015; Sejnowski, 2020). A leading science publication selected AI algorithms' ability to predict the three-dimensional structures

of proteins as the scientific breakthrough of 2021 (Service, 2021). AI-powered programs, such as *Midjourney AI* and *DALL-E*, can generate artistic images with text prompts within a few seconds (Oppenlaender, 2022). With numerous parameters and trained on large quantities of data, Large language models (LLMs) like OpenAI's GPT families and Google's LaMDA, with properly-designed interfaces like ChatGPT, can produce extremely lifelike answers when engaging in conversations with users, arguably summarize knowledge from the entire internet, perform a wide range of tasks, and are believed to be able to revolutionize how people access knowledge and even the entire higher education system (Lund and Wang, 2023). AI is thought to have the potential to not only aid the process of invention, but to become a method of invention because of its ability to rapidly and effectively search high-dimensional solution spaces that are characteristic of many types of problems in science and technology (Agrawal et al., 2018; Cockburn et al., 2019; Crafts, 2021).

AI is both a field of scientific and engineering research and a domain of technological development, which seek to invent, that is, to generate intellectual novelty. AI studies the mechanisms and procedures of intelligence, including learning, reasoning, perception, and pattern recognition. It aims to develop devices, computational processes, and algorithms that display intelligence, often with greater speed and accuracy, and the ability to process data volumes far beyond human capability (Mitchell, 2019). Even if initially animated by an interest in understanding and replicating human intelligence, AI's focus now is on intelligence *sui generis* (Crevier, 1993). AI draws upon mathematics, computer engineering, linguistics, neuroscience, cognitive psychology, decision-making theory, search theory, machine learning, data science, speech, pattern recognition, and the theory of computation (Russell and Norvig, 2020).

A Brief History of AI

The creation of AI Some of AI's basic ideas and exploratory attempts already existed before the term AI was conceived. Examples include the Turing Test (Turing, 1950), the primitive model of artificial neurons that can be switched "on" or "off" (McCulloch and Pitts, 1943), and the first neural network computer SNARC (the Stochastic Neural Analog Reinforcement Calculator), built in 1950 (Crevier, 1993). Nevertheless, a workshop at Dartmouth College in the summer of 1956 has been widely acknowledged as AI's birthplace. The term "AI" was coined in the proposal for the workshop submitted to the Rockefeller Foundation in 1955 by John McCarthy, Marvin Minsky, Nathan Rochester, and Claude Shannon (McCarthy et al., 1955). They were later recognized as the founding fathers of AI. In the proposal, AI was defined as a machine that can simulate "every aspect of learning or any other feature of intelligence." The attendees at the Dartmouth Workshop attempted to find "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al., 1955).

1950s-1970s: Symbolic AI The Dartmouth Workshop created a brand-new academic space and collaborative relationships rather than actual research output. The conception of AI as a new, separate field with well-articulated goals but little establishment opened up exploration opportunities. Many theoretical foundations were laid in the 1950s and the 1960s, and they remain relevant even today. As C. S. Peirce (1960) put it, "we think only in signs;" and, Terrence Deacon (1997) saw humans as a "symbolic species" whose defining characteristic is manipulating symbols. It is not surprising that the first wave of AI was characterized by its deep roots in mathematics and logic, focusing on symbol manipulation. For example, Newell and Simon (1976) designed the General Problem Solver (GPS) that mimicked humans in

solving logic problems, paving the way to their award-winning achievement – physical symbol system hypothesis. Newell and Simon (1976) believed that "symbols lie in the root of intelligent action" regardless of the actor, be it a human or a machine. By manipulating and structuring symbols correctly, intelligence can arise from any system (Newell and Simon, 1976). Those symbolic AIs are referred to as "GOFAI" (Good-old Fashioned AI) (Haugeland, 1985).

The birth of connectionism During this early stage of AI, in addition to "GOFAI," a completely different approach to AI, referred to as "connectionism" exemplified famously later on by deep learning, had its foundation laid as well. Even though humans think in abstract signs, on the physical level, our brain is made up of large quantities of almost homogenous neurons, whose activation and interaction function to process information, store and retrieve memories, and give commands to our bodies. Maybe, through mimicking such structures, machines can learn how to learn intelligently without pre-programmed prior knowledge. This is the basic idea behind connectionism. In 1958, Frank Rosenblatt at Cornell proposed the idea of perceptron, an algorithm to achieve binary classification through a network of weighted neurons (Sammut and Webb, 2011). He built its first implementation with 400 artificial "neurons" made of "cadmium sulfide photocells" (Bishop, 2006). However, with the absence of adequate computing power and programming tools, both symbolic and connectionist AI of that time had focused on simple "toy problems" such as checkers instead of real-world problems. Neither symbolic or connectivist AI had fulfilled their big promises. Rosenblatt's claim that his perceptrons would gain the ability to read, walk, and talk like humans within a year and with just $100,000 in research funds never came to fruition (NYT, 1958).

1980s: Expert systems Besides the limited computing power, another big problem of AI that limited its practical use in the first decades was its lack of domain-

specific knowledge. In the 1970s, knowledge-based systems, known as expert systems, started showing promise in solving real-world problems. An expert system generally comprises two components: the knowledge base, where domain-specific knowledge (rules and facts) is represented symbolically, and a general-purpose inference engine that describes how to manipulate these symbols (Kaplan, 2016). To build an expert system for a specific domain, extensive interviews with experts had to be conducted to acquire the knowledge. Expert systems achieved great commercial success in the 1980s partly because of the investment stimulated by the competition between Japan and the U.S. Expert systems were helping tasks such as design, monitoring, and diagnosis in a wide range of industries, including but not limited to computers, accounting, oil drilling, banking, and power plants, saving each company tens of millions of dollars every year (Leonard-Barton and Sviokla, 1988). By 1988, almost every major U.S. company was either using or looking into expert systems (Russell and Norvig, 2020).

AI winter However, expert systems failed to achieve satisfactory performance because their frameworks lacked "sufficient expressive power to capture the breadth of expert knowledge and behavior." Besides, extensive and expensive handcrafting and hard coding were needed (Kaplan, 2016). Many companies failed to fulfill their promises. From 1987 to 1993, due to the daunting prospect of expert systems' practical potential, investment in AI fell significantly. AI had entered a period referred to as the "AI winter" (Church, 2011).

The rise of machine learning The collapse of expert systems did not obstruct the progress of AI. If it is challenging to teach knowledge to machines, how about creating machines that can learn the knowledge by themselves? Humans learn from our interactions with the environment. Learning is undoubtedly a crucially important aspect of human intelligence. This idea formed the basis of machine learning, which started to be treated seriously by researchers in the early 1990s. The pendulum began

to swing from symbolic AI to connectionist AI. Many new developments occurred. For instance, the re-introduction of back-propagation, a technique that allows connectionist models to learn from data and adjust the weight of each neuron during the learning process (Rumelhart and McClelland, 1986), opened up the possibility of learning from big data.

In recent years, deep learning, a kind of artificial neural network that consists of many layers, has shown great potential in machine learning due to the increased availability of training data, improved computing power, and unprecedentedly sophisticated algorithms. It has achieved or even surpassed human capacities in some domains, such as playing Go (Silver et al., 2016) and recognizing faces (Lu and Tang, 2014). Today, AI is thriving in many areas of research and development, including but not limited to robotics, speech recognition, computer vision, natural language processing (NLP), self-driving cars, and game playing.

1.1.2 Combinatorial Inventions

Invention and *innovation*, although often used interchangeably, have different meanings. According to the Cambridge Dictionary, an invention refers to "something that has never been made before, or the process of creating something that has never been made before." An invention involves the discovery of new knowledge and may not have practical applications or be commercially viable. On the other hand, an innovation refers to "the use of new ideas and methods," or as Maclaurin (1953) put it, "a new or improved product or process" as the result of an invention "introduced *commercially*." Innovation involves the development of products that are useful and valuable to society. An invention precedes the innovation that uses the invention. This research focuses on inventions rather than innovation. Therefore, the documentation of new ideas and processes (scientific publications and patents)

instead of commercialized products is examined in this research.

Inventive-like processes emerge in both the natural world and human societies. In the natural world, invention occurs through Darwinian evolution, continuously introducing new organisms, physiological functions, and new ecosystems, filling the Earth with diverse ecosystems. In human societies, an invention consists of ideas or technologies, either new or already in use, brought together in a way not previously observed (Schumpeter, 1934). Inventions take various forms, such as patents, music, movies, publications, and paintings. Continuous inventive activities foster scientific and technological development, bring about cultural change, maintain economic growth, and facilitate social progress.

Regardless of whether they are natural or anthropogenic, inventive processes involve similar mechanisms such as trial and error, descent with modification, horizontal information transfer, and combining existing components (namely *combinatorial inventions*) (Wagner and Rosen, 2014). Combinatorial inventions are created by reassembling existing factors in new ways (Becker, 1982; Schumpeter, 1934; Arthur, 2009). Combinatorial processes are considered an essential source of invention. Due to this combinatorial nature, inventing can be conceptualized mathematically as a search problem over a space of combinatorial possibilities (Kauffman et al., 2000; Kauffman, 1993; Wagner and Rosen, 2014).

Scientists, economists, management researchers, artists, policymakers, and others have observed and studied combinatorial inventions. Different domains differ in the forms of inventions, the components to be combined, and how they are combined.

Combinatorial Inventions in Different Domains

Combinatorial inventions in art In art, music, film, media, and culture, a remix of existing elements can produce creative works and novel experiences (Borschke, 2017;

Navas, 2012). For example, Seth Austen and Grahame-Smith (2009) recombined the story of Jane Austen's *Pride and Prejudice* with B-grade zombie pulp fiction to create a new novel. Mark Vidler recombined Madonna's "Ray of Light" (1998) and the Sex Pistol's punk music to create "Ray of Gob" (Gunkel, 2016), not to mention Andy Warhol's famous *Marilyn Diptych* (1962) that combines fifty paintings based on the same photo from the film *Niagara* (1953).

Combinatorial inventions in technology Technologies are no exception. As Brian Arthur (2009) put it in his book *The Nature of Technology*, "technologies are combinations," and novel technologies are constructed from existing ones. Disassembling any new device at your home, be it a phone, laptop, TV, toaster, or fridge, you will recognize that almost all parts consist of existing technologies, such as cameras, CPUs, glass, and screws. Each disassembled part is a subsystem combining another set of technologies. Because inventions can be conceptualized as a search problem in a landscape of possible inventions, entities that invent, such as organizations and inventors, are solving a combinatorial optimization problem by considering their initial conditions, boundary conditions, and available resources (Kauffman et al., 2000; Macready et al., 1996).

Combinatorial inventions in science Every little step in science is built upon previous knowledge. As Isaac Newton (1675) remarked, "if I have seen further, it is by standing on the shoulders of giants." Even what Thomas Kuhn (1962) would refer to as "revolutionary science" that fundamentally shifts paradigms needs to incorporate existing knowledge acquired through "normal science." For instance, Einstein's general relativity theory was not widely accepted until its prediction of the starlight's gravitational deflection was confirmed by the Eddington experiment designed and carried out in the "normal" way (Coles, 2019). Scientific inventions usually take the form of academic publications. In scientific publications, one way existing knowledge

is recombined is through references. Previous literature is reviewed and referenced to elaborate on the relevance, gaps, methodologies, contribution, and other aspects of this new research.

Carriers of Combinatorial Inventions

How existing components are combined to create new knowledge has been viewed as the "holy grail" of innovation studies (Gruber et al., 2013; Schumpeter, 1934). In literature, the evidence of combinatorial inventions has been examined in various carriers and on different levels. The carrier of the inventions is the form with which the invention is recorded and recognized. The forms could be technical documentation, patents, academic publications, commercial products, software, music, films, or others. At the same time, the scope of inventions could be examined on different levels organized either by individual pieces of work (e.g., patents or scientific papers), people (e.g., engineers or companies), geographic scope (e.g., industrial parks, cities, regions, or nations), domains (specific fields or sub-fields), or across time. The carrier and the level together determine the components that serve as existing knowledge to be combined in the study.

Patents as invention carriers Patents are often considered inventions' carriers by researchers and policy-makers. Because patent offices use "technology codes" to classify patents, these codes are suited to represent the components for combinatorial innovation (Clancy, 2015; Gruber et al., 2013; Youn et al., 2015).[1] A patent with multiple codes can be seen as a combination of previous technologies of each code. In this way, Youn et al. (2015) explored the code-to-code combinatorial dynamic of inventions recorded in US patents dating from 1790 to 2010 and found a constant

[1]Detailed explanation and discussion regarding "technology codes" can be found in Chapter 3 and Chapter 4.

rate of exploitation and exploration even though the creation of new technological capacities has significantly slowed down.

Scientific publications as invention carriers In science, inventions are recorded mainly in theses, dissertations, peer-reviewed papers, books, and monographs. In the literature that uses scientific publications as carriers of scientific inventions, references are usually used to represent the existing knowledge combined (Mukherjee et al., 2016; Uzzi et al., 2013).[2] In some studies, referenced journals rather than referenced papers have been chosen to represent existing knowledge. In a method developed by Mukherjee et al. (2016) to analyze combinatorial inventions in science, each of the journals cited in a given paper represents a piece of previous knowledge. A journal pair co-cited in the paper represents the recombination of two pieces of previous knowledge. A journal pair can be classified as typical or atypical depending on whether the two journals are co-cited in the dataset more or less frequently than random. Using this method, Mukherjee et al. (2016) created a taxonomy of novelty to categorize scientific papers into four types. They considered the entire Web of Science dataset and found that papers that mix typical and atypical combinations tend to become top papers in citation counts. In addition to referenced literature, keywords extracted from publications can be seen as combinatorial components. For example, in a study on diversity and innovation conducted by Hofstra et al. (2020), a dissertation that combines two keywords (e.g., HIV and monkey) for the first time in the *ProQuest*[3] database is considered an invention.

[2]Detailed discussion about this method developed by Uzzi et al. (2013) and Mukherjee et al. (2016) can be found in Chapter 2.

[3]`https://www.proquest.com`

Levels of Combinatorial Inventions

Combinatorial invention on the individual level An individual creator of inventions, such as a patent inventor or a software developer, embodies the existing knowledge mastered before the invention, which strongly relates to whether they can invent and what and how they invent. Gruber et al. (2013) studied how inventors' individual-level characteristics affect the breadth of the technological recombinations in their patent applications. They found inventors with scientific degrees rather than engineering degrees, and inventors with doctoral degrees tend to combine existing knowledge across disciplinary boundaries. Collaborations between individuals expand the existing knowledge available to be combined. Methodologically, co-authorship network analysis is often used to study how collaboration creates new knowledge. For example, Fleming and Marx (2006) examined the patent co-authorship networks (or co-patenting networks) in Boston and Silicon Valley and found that small-worldness has become the "new inventive basis of innovation."

Combinatorial inventions on the organizational level A group of inventors organized in a certain way, such as an R&D department in an organization or a conference focusing on a specific topic, can be seen as an embodiment of existing knowledge collectively mastered by the members. Collaboration between departments, organizations, or institutions is also believed to stimulate innovation by providing an opportunity to fuse knowledge from different domains. Mervin Kelly, the president of Bell Labs in the 1950s, created a physical environment at Bell Labs that facilitated inter-departmental groups of physicists, chemists, metallurgists, engineers, theoreticians, and experimentalists. This "interdisciplinary nature" at Bell Labs was believed to be an essential factor in its success (Gertner, 2012).

There is also a body of literature in innovation studies that explores innovation

systems and the complex interactions among the diverse actors within these systems. Those innovation systems' boundaries are usually defined on a sectoral, regional, or national level (Edquist, 2005). These studies consider a variety of actors, including universities, corporations, and governments. They care about the impact of institutional, social, and economic factors on inventions and innovation and how stakeholders such as business managers and policymakers can make better decisions (Christensen, 1997).

1.1.3 Combinatorial Inventions regarding AI

AI is an inherently multidisciplinary field that integrates diverse fields of knowledge, a defining feature that provides the empirical means to investigate the novelty of the scientific and technological output propelling the field. Studies of science and technology development conceptualize invention as combining existing factors in new ways (Schumpeter, 1934; Arthur, 2009; Uzzi et al., 2013; Youn et al., 2015). The history of science is one of accumulating insights as every new development builds upon previous knowledge (Arthur, 2009; Uzzi et al., 2013). Even what Thomas Kuhn called "revolutionary science" that fundamentally shifts conceptual paradigms incorporates existing knowledge acquired through "normal science" (Kuhn, 1962; Mitchell, 2019; Crevier, 1993). Combining existing knowledge or introducing a completely new idea occasionally results in a startling intellectual transformation (Ziman, 1978; Weinberg, 2015; Wootton, 2015; Gal, 2021). A shared perspective in economics, management science, sociology, anthropology, and history posits that the recombination of new and existing technological capabilities is the principal source of technological novelty (Arthur, 2009; Tarde, 1903; Kaempffert, 1930; Usher, 1954; Kirby et al., 1956; Derry and Williams, 1993; Hippel, 1988; Basalla, 1988; Pacey, 1990; Richerson et al., 2013).

AI and inventions are interconnected and mutually reinforce each other. The rela-

tionship between AI and combinatorial inventions is three-fold: (a) the combinatorial inventions within AI, (b) AI as a component combined with other domains, and (c) AI as a means for combinatorial inventions, in other words, using AI as a tool in searching over the space of possibilities, either for creating search space, searching over such space, or optimizing search strategy.

Combinatorial Inventions within AI

As Section 1.1.1 demonstrated, the history of AI *per se* is a process of combinatorial inventions. Many AI ideas are initially borrowed from other disciplines, such as biology, mathematics, logic, psychology, linguistics, or philosophy. For example, the first model of artificial neurons by McCulloch and Pitts (1943) was a combination of ideas rooted in three disciplines: the physiology and function of human neurons, the formal analysis of propositional logic, and Turing's theory of computation (Russell and Norvig, 2020). Similarly, the framework of machine learning to deal with uncertainty was provided by probabilistic theory in mathematics (Ghahramani, 2015). In addition, AI applications are nearly always built with multi-disciplinary knowledge. For instance, the capabilities allowing AI to interact with humans and the environments physically, verbally, and emotionally requires knowledge from domains such as robotics, natural language processing (NLP), computer vision, and psychology. Furthermore, many AI researchers have multi-disciplinary backgrounds. For example, Newell and Simon are both mathematicians.

Much research has implicitly investigated the semantics of combinatorial inventions within AI. However, very few are devoted to quantifying, measuring, or characterizing the general process of combinatorial inventions within AI. Some researchers attempt to identify AI inventions' significant trends. For example, Cockburn et al. (2019) found that the AI community is shifting towards more application-oriented

learning research. Niu et al. (2016) surveyed where the highest productivity resides in AI. Raghupathi and Nerur (1999) examined the author-to-author co-citation networks and identified some research themes and trends in AI. Others embed AI in broader knowledge landscapes, aiming to build a holistic understanding of the innovation system around AI. For example, Frank et al. (2019b) studied the interdisciplinary knowledge flow embodied in the citation relationships between AI and other domains. They found that social science and other non-AI areas are not keeping up with AI development's fast pace in recent years. They suggested that it might explain why AI has had negative social impacts recently.

AI as a Combinatorial Component

Because of its ability to automate physical and complex cognitive tasks, AI has been increasingly recognized as a general-purpose technology that can spawn further inventions (Brundage, 2019). In many cases, AI is incorporated into and becoming a part of the final products of inventions. For instance, Apple's Siri and Amazon's Alexa integrate AI components such as speech recognition. In literature, any survey on AI applications is essentially a demonstration of AI as a component for combinatorial inventions, for instance, AI used in smart grid (Li et al., 2018), healthcare (Bali et al., 2019), law (Liu, 2020), and engineering in general (Shukla et al., 2019). Furthermore, the rise of domains such as AI ethics and philosophy can be also considered combinatorial inventions that integrate AI with other knowledge.

AI as a Means for Combinatorial Inventions

Another perspective of AI as a general-purpose technology is that AI has become a method for inventions that facilitates further inventions, praised as a "new method for invention" (Cockburn et al., 2019). AI techniques have been routinely incorporated in

the research and development in domains not traditionally around AI's disciplinary "neighborhood," for instance, biology and archaeology (Davis et al., 2019). The difference between AI as a component and AI as a means is in which step AI is engaged in the inventive process. The former involves AI in the final result, and the latter involves AI during the process but does not necessarily involve AI in the final product.

AI as a means for combinatorial invention has two meanings. First, AI can be used for searching over the space of possible inventions, either for searching or optimizing search strategy. It is made possible because of AI's promising performance in search problems, fast iteration, heuristics, and learning from experience, for example, machine learning in polymer design (Kim et al., 2021), or genetic algorithm optimization in drug design (Fernandez et al., 2011). Another examples is the Material Genome Initiative (MGI), which advances new paradigms such as AI for material discovery and design (de Pablo et al., 2019). For example, Gómez-Bombarelli et al. (2016) utilized machine learning to screen through 1.6 million molecules of organic light-emitting diode (OLED), and identified a set of promising molecules with state-of-the-art external quantum efficiencies. In these examples, AI is involved in the R&D process but not the resulting products, i.e., drugs and polymers. Also, because search problems constitute a significant part of the problems that AI researchers traditionally attend to and AI has been remarkably fruitful in this regard, in light of the conceptualization of inventing as a combinatorial optimization problem, AI is particularly suited for such search problems, either for algorithm-oriented metaheuristics problems or problem-oriented, context-specific problems (Blum et al., 2011). For example, Kulkarni et al. (2016) has introduced an AI-based technique called cohort intelligence to solve combinatorial optimization problems in supply-chain domains. Bezerra et al. (2021) used AI as a combinatorial optimization strategy to help with

cellulase production innovation using peach-palm waste.

Secondly, AI could be used as a research tool to study combinatorial inventions due to AI's applicability in knowledge representation, network analysis, and graph analysis in general. For instance, Choi et al. (2015) constructed a model that processes patent data for predicting technology transfer using social network analysis, regression analysis, and decision tree.

Those studies have deepened our understanding of how combinatorial inventions occur within AI and how AI is embedded as combinatorial components in broader socio-technical innovation systems. However, little effort has been devoted to characterizing combinatorial inventions in AI in terms of their novelty and impacts. How novel are AI inventions? What characteristics of AI inventions are related to their impact on subsequent knowledge creation? Those are the questions explored in this research.

1.2 Research Question

The overarching research question of this dissertation is *how new knowledge is created through combinatorial processes in the field of AI*. Further, how does such a combinatorial process relate to the quantity and quality of the knowledge created?

I set out to address such a research question by investigating three aspects of knowledge creation in AI — the creation of scientific knowledge as recorded in academic publications, the creation of technical knowledge as recorded in patents, and the creation of technical knowledge within organizational and industrial boundaries.

Several sub-questions are posed in addressing the creation of AI-related scientific knowledge. First, how is existing scientific knowledge combined to create new knowledge in AI? Secondly, how are knowledge recombinations associated with scientific impact? In other words, what kind of knowledge recombination predict higher

citations? Thirdly, how do novel and conventional combinations in AI emerge and evolve?

In investigating the combinatorial inventions in patents related to AI, several sub-questions are addressed, including how novel, diverse, and disruptive AI patenting is. Furthermore, this research asks how public support has fueled AI patenting and what inventor team sizes affect AI patents' impact and disruption. In addition, what is different and what is similar between AI publications and AI patents?

Three sub-questions are tackled when addressing the question regarding organizational inventions in AI. First, how can organizations' exploitative and exploratory searches be measured to reflect how much organizations have exhausted the search space? Second, how can those measurements be mathematically formalized? Last, how can such measurements be visualized to inform organizations' inventive competence and policy-making?

1.3 Structure of the Dissertation

This dissertation consists of three main essays (Chapters 2, 3, and 4) in addition to the introductory Chapter 1 and the conclusion Chapter 5. Each of the three main chapters addresses a separate set of research questions regarding combinatorial invention in AI.

Chapter 2 engages the research questions of how novel AI research is and how the combinatorial characteristics of AI research affect its scientific impact. I collected a dataset of AI research publications that contains nearly 300,000 records from the Web of Science (WOS) using a snowball data collection technique. Then, Chapter 2 utilized the modified version of an analytical framework developed by Uzzi et al. (2013) and Mukherjee et al. (2016) to study the combinatorial characteristics of this dataset. I found that AI publications are growing faster than scientific publications

in general. AI publications that feature highly conventional combinations of existing knowledge account for the highest percentage and receive the most citations. Such AI publications also have a higher chance of becoming top-cited. This implies that scientific endeavor related to AI has not deviated significantly from the traditional "normal" scientific research that relies heavily on incremental advance. It implies that it is debatable how revolutionary AI transforms scientific research.

Chapter 3 investigates extensively how AI patenting is advancing and how combinatorial features relate to AI patents' other characteristics, such as novelty, diversity, impact, and team sizes. I constructed an AI patent dataset comprising over 250,000 U.S. patents related to eight AI component technologies, including knowledge processing, speech, AI hardware, evolutionary computation, natural language processing, machine learning, vision, and planning and control. Like AI publications that grow faster than scientific publications in general, AI patents grow faster than utility patents in general. Nevertheless, contrary to anticipation, AI patents have been less novel and less diverse than utility patents. AI patents also have a lower percentage of highly-disruptive patents and a more substantial reliance on public support than utility patents. AI patents that only refine existing technologies or combine existing technical components in new ways account for the largest share. Nevertheless, this research found that AI patents that combine existing components in new ways tend to have higher citations. In contrast, highly original patents tend to disrupt subsequent knowledge creation to a greater extent. A previous report concluding that "large teams develop and small teams disrupt" is found to hold in AI patents. In addition, I found that highly original patents tend to have smaller teams, while refinement patents often have larger teams. Compared to AI publications, the novelty in AI patents is better rewarded by citations. Regardless of publications or patents, conventionality is increasing while novelty is declining.

Chapter 4 starts with the development of an analytical framework — the **C**ombinatorial **E**xploitation and **E**xploration (CEE) framework. The CEE framework quantifies organizations' combinatorial exploitation and exploration with eight parameters in three dimensions — knowledge access, inventive outcome, and knowledge discovery. These eight CEE parameters assess the extent to which organizations have exhausted their possible combinations rather than merely counting the number of patents. I explored the application of the CEE framework through a case study using AI patents granted by the United States Patent and Trademark Office (USPTO). The case study showcased how the CEE framework can be used and visualized to improve our understanding of organizational inventions and industrial development and inform decision-making. Through this case study, this research found that AI organizations prefer exploitative over exploratory invention. Nevertheless, exploitative efforts often peaks within the first five years after an organization initiates AI patenting and then slightly declines. Meanwhile, the exploratory search would gradually increase or remain stable if the organization remains active in AI patenting.

The conclusion Chapter 5 first reiterates the findings of the three essays (Chapters 2, 3, and 4). Then, the implications of this research are discussed. To the research communities concerning the history of science and technology, this research implies that revolutionary technology can be and will likely be born from routine R&D practices. I then offer insights for individual researchers and inventors regarding how to produce high-quality publications and patents. For funding agencies, universities, and organizations, this research emphasizes the significance of public funding and provides additional tools to assess inventive performance. Last, plans to extend and perfect this research were proposed.

1.4 Research Tools

This research uses Python v.3.9.4[4] (pandas v.1.5.2,[5] NumPy v.1.24.1,[6] Matplotlib v.3.6.3,[7] seaborn v.0.12.2,[8] statsmodels v.0.13.2[9]) and PyCharm Integrated Development Environment (IDE) v.2022.3.1[10] to wrangle, analyze, and visualize data.

To analyze and visualize network structures, *Gephi*,[11] the open graph viz platform (v.0.10.0[12]) is used (Bastian et al., 2009).

[4]See Python release note: https://www.python.org/downloads/release/python-394/

[5]See pandas release note about this release: https://pandas.pydata.org/docs/whatsnew/v1.5.2.html

[6]See NumPy release note about this release: https://numpy.org/devdocs/release/1.24.1-notes.html

[7]See Matplotlib release note: https://matplotlib.org/stable/users/prev_whats_new/github_stats_3.6.3.html

[8]See seaborn release note about this release: https://seaborn.pydata.org/whatsnew/v0.12.2.html

[9]Release note: https://www.statsmodels.org/dev/release/version0.13.0.html

[10]See PyCharm release note: https://www.jetbrains.com/pycharm/whatsnew/

[11]https://gephi.org

[12]See realease note: https://gephi.wordpress.com/2023/01/09/gephi-0-10-released/

Chapter 2

KNOWLEDGE RECOMBINATION IN ARTIFICIAL INTELLIGENCE RESEARCH AND ITS SCIENTIFIC IMPACT

2.1 Introduction

Artificial Intelligence (AI) research, development, and utilization are both intrinsically innovative and serve as a tool and source of innovation for research and development in a variety of domains.[1] In recent years, expectations have risen that AI will revolutionize innovation and profoundly impact nearly every aspect of society. As such, there is an assumption that AI can be considered "revolutionary science" rather than "normal science." Nevertheless, is this assumption supported by empirical evidence? Using a newly-compiled dataset of nearly 300,000 AI research publications, this chapter investigates the degree to which AI research is novel. The work of this chapter is accomplished through knowledge combination analysis. This chapter surveys how knowledge recombination relates to the impact of AI publications. This research found that while the number of publications from AI research is growing exponentially and faster than science in general, when combining existing knowledge to create new knowledge, AI is still advancing incrementally. This chapter's analysis indicates that research combining existing knowledge in highly conventional ways without introducing radically new ideas is a substantial driving force in AI. Despite

[1]Part of the results in Chapter 2 will be adapted and submitted for publication. The manuscript is written with the contribution of Dr. Andrew Maynard, Dr. José Lobo, Dr. Katina Michael, Dr. Sébastien Motsch, and Dr. Deborah Strumsky. Figures in Chapter 2 were generated with the help of Dr. Sébastien Motsch.

assumptions to the contrary, AI has been advancing in ways resembling "normal science," i.e., growing cumulatively and incrementally rather than like "revolutionary science," which dramatically revises existing scientific practice.

Artificial Intelligence (AI) has developed quickly in recent years. AI shows tremendous potential to affect the search for solutions in numerous domains ranging from medicine (Bali et al., 2019) to transportation (Donnellan, 2018), from education (Zawacki-Richter et al., 2019) to military (Szabadföldi, 2021), from manufacturing (Maynard, 2015) to social media (Fernandez-Luque and Imran, 2018), from playing Go (Silver et al., 2016) to astrophysics (Biswas et al., 2013), and within many other areas. Many issues and concerns have been raised around specific features and consequences of AI applications, including algorithmic bias (Akter et al., 2021), technological unemployment (Brynjolfsson and Mitchell, 2017), autonomous weapons (Welsh et al., 2018), surveillance (Introna and Wood, 2002), and the potential disruptions to laws, norms, and social institutions (Vesnic-Alujevic et al., 2020). However, AI is widely expected to become a general-purpose technology and a "new method of invention" that will transform the processes of invention and innovation. This is partly due to AI-based efficiencies in searching over highly complex solution spaces and new approaches to solving "needle-in-a-haystack" types of problems prevalent in science and technology (Agrawal et al., 2018; Brundage, 2019; Cockburn et al., 2019). AI-enhanced search is crucial for knowledge creation which has been long understood as a combinatorial process through which existing knowledge is re-combined to create new knowledge (Gruber et al., 2013; Schumpeter, 1950; Arthur, 2009; Youn et al., 2015; Strumsky and Lobo, 2015a). As accumulated knowledge has grown, the scale of possible combinations of knowledge units has increased faster (a combinatorial explosion). The promise of AI to transform invention largely stems from the capacity to search over vast combinatorial spaces and identify viable, thus possibly valuable,

new knowledge combinations.

Thomas Kuhn famously distinguished "revolutionary science" from "normal science," with revolutionary science introducing radical novelty and paradigm shifts, while normal science produces knowledge in an incremental fashion that builds mainly upon well-accepted previous knowledge (Kuhn, 1962). AI's capabilities are seen as possibly facilitating the search for novel knowledge combinations constituting revolutionary science. However, how is AI itself advancing? Is the field advancing in a revolutionary manner or in a way reminiscent of previous episodes of change in other areas of science and technology?

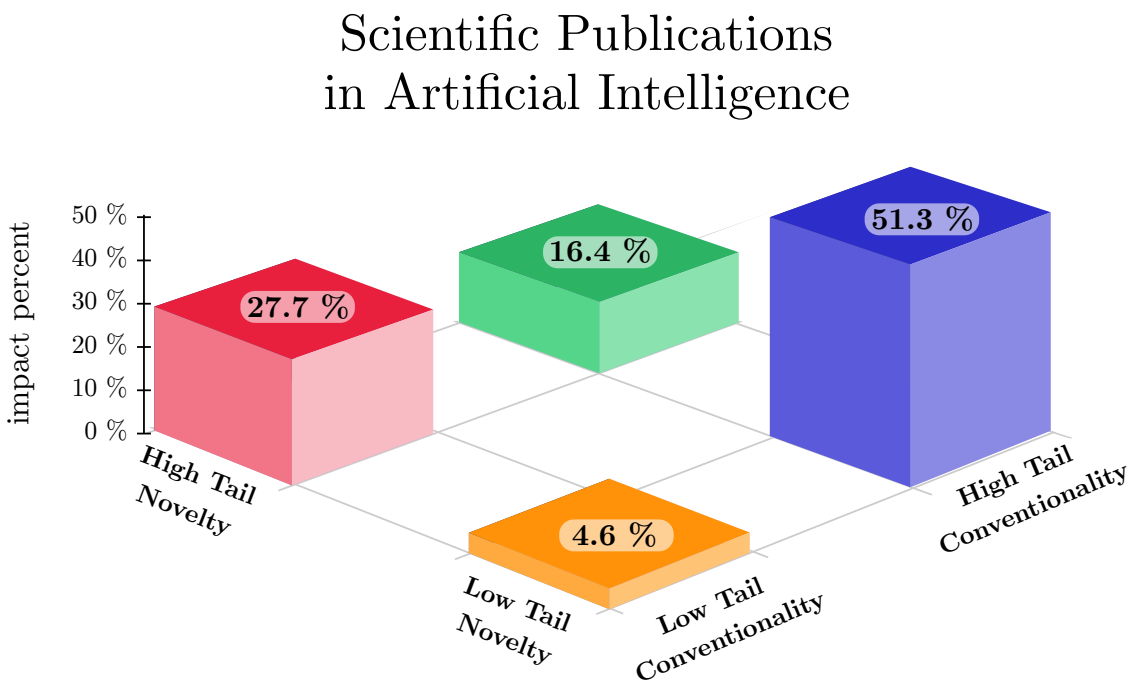


Figure 2.1: The scientific impact of AI publications, conditional on novelty and conventionality. The figure presents citations received by different AI publications conditional on two dimensions: whether an AI publication exhibits (i) high or low tail conventionality and (ii) high or low tail novelty, as defined in this paper. AI publications that are highly conventional (blue bar and green bar combined) have generated a majority of the citations. Specifically, AI publications that combine high conventionality with low novelty (blue bar) have received more than half (51.3%) citations of all AI publications. The sample includes AI-related scientific publications recorded in the Web of Science (WOS) from 1946 to 2020.

This chapter examined nearly 300,000 scientific publications related to AI through the perspective of knowledge recombination. The number of AI publications has grown exponentially and at a faster rate than science-based publications in general. However, AI has not significantly deviated from the historical progression associated with normal science. Specifically, AI publications that are highly conventional and with little novelty represent the largest share of new AI knowledge. The percentage of highly conventional knowledge combinations is even higher in the highest-impact publications. Furthermore, AI publications that feature highly conventional knowledge combinations create higher scientific impact (Figure 2.1). In addition, they have a higher probability of becoming top-cited papers and a lower probability of becoming never-cited papers. The results presented in this chapter imply that exploiting conventional combinations of existing knowledge has become the dominant driver of new knowledge creation in AI research. Despite expectations to the contrary, AI research is accurately described as normal science rather than revolutionary science.

2.2 Research Questions

This exploratory study addresses the research question of *how existing knowledge is re-combined to drive innovation in and development of AI research.* It is underpinned by three sub-questions.

1. How is existing knowledge combined within the field of AI to create new knowledge?

2. How is knowledge recombination associated with scientific impact? And

3. How do novel ideas and conventional ideas in AI emerge and evolve?

2.3 Method

This chapter used an analytical framework, a variation of the method developed by Uzzi et al. (2013), to identify and measure the novelty of knowledge recombinations in the scientific literature. Specifically, the framework identifies pair-wise combinations of journals cited by a scientific publication to represent the recombination of existing knowledge. Two statistical features of the journal pairs associated with a publication are identified as the measurements for *conventionality* (frequent pairings of knowledge units) and *novelty* (atypical or novel pairings). Based on those two measurements, the framework also supports a taxonomy to categorize the scientific literature into one of four categories (see Figure 2.2).

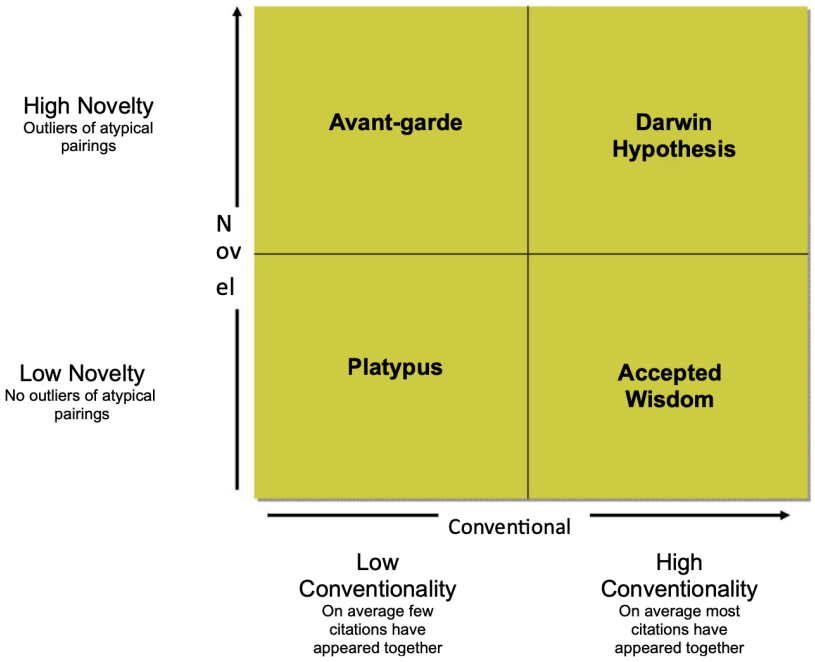


Figure 2.2: Categorization of Types of Scientific Papers Based on their Combinations of Conventional and Novel Pairings of Prior Published Work (Image from Mukherjee et al. (2016))

2.3.1 Data Collection

The goal of data collection in this study was to construct a dataset of AI publications that effectively represents AI-related research. Data collection was designed to include publications in computer science and other disciplines that either contribute to AI or utilize AI as a research tool.

Previous research on AI publications has often relied on data collected through either downloading pre-constructed, domain-specific datasets such as AI-related fields in the Microsoft Academic Graph (MAG) (Frank et al., 2019b), or searching AI-related keywords in multi-disciplinary literature databases such as Web of Science (WOS) and compiling the returned records (Niu et al., 2016; Cockburn et al., 2019). The first way is to use pre-constructed, domain-specific datasets to approximate AI research without searching. For instance, in their study, Frank et al. (2019b) used all the papers in AI-related subfields in computer science (CS) in the Microsoft Academic Graph (MAG) database. This dataset is appropriate for answering their research questions. However, it may be overly inclusive of the CS papers that do not address AI and, at the same time, exclude papers outside selected CS sub-fields related to AI, such as those in computational linguistics or cognitive science. The second way to construct an AI publication dataset is to search keywords in specific multi-disciplinary literature databases and compile the returned records. Different research differs in databases and search terms. For instance, Niu et al. (2016) searched in the titles, abstracts, and keywords in three databases (SCI-Expanded, CPCI-S, and WOS) with one search term – "*artificial intelligence*," in which the wildcard asterisk symbol "*" allows for the inclusion of any word that contains "artificial intelligence." A total of 22,072 publications were retrieved in their study. This approach omitted the publications that do not explicitly mention the term "artificial intelligence" in their titles and

abstracts. Those omitted publications make up a non-negligible proportion of AI publications. A search on WOS using the term "*artificial intelligence*" returns only 37,045 records, but searching "machine learning" returns over 107,000 results, almost three times as many. Authors of another study searched WOS using 42 keywords that encompass three main sub-disciplines identified by the researchers (symbols, learning, and robotics) (Cockburn et al., 2019). Eventually, 98,124 publications were retrieved.

This chapter sets out to incorporate AI-related publications as accurately, comprehensively, and inclusively as possible across disciplinary boundaries. The dataset for this study was compiled through keyword searching on WOS with snowball sampling to achieve this. The data collection method is described as follows.

Searched database This study's analysis draws upon a dataset that captures a set of AI publications from the WOS Core Collection[2]. WOS is one of the largest repositories for scientific publications, currently maintained by Clarivate Analytics (previously the Intellectual Property and Science business of Thomson Reuters). The WOS Core Collection contains 74.8 million records in 21,100 peer-reviewed scholarly journals worldwide, conference proceedings, and books in 254 science, social sciences, and arts & humanities disciplines (Web of Science, 2020).

Search field The dataset was collected through snowball sampling (Biernacki and Waldorf, 1981) that utilized the "*Topic*" search in WOS in July 2020. Search terms input into the *Topic* field in WOS were searched in the title, abstract, author keywords, and *Keywords Plus* of a record in a case-insensitive fashion. It is worth noting that *Keywords Plus* are words or phrases generated by an algorithm developed by Clarivate Analytics to identify keywords that frequently appear in the titles of an article's reference but do not appear in the title of the article itself (Clarivate

[2]The WOS databases searched include Science Citation Index Expanded, Social Science Citation Index, Arts and Humanities Citation Index, and Emerging Sources Citation Index (only 2015-2020).

Analytics, 2018). Thus, a *Topic* search is sufficient to capture the main topics of a publication.

Sampling Snowball sampling is a non-probability sampling technique commonly used in sociology to recruit future subjects through existing subjects. It has also been used in bibliometric research to acquire further literature through the references in existing literature (Roetzel, 2018). This study uses the snowball sampling method to incorporate new search terms through the records acquired using existing search terms.

The dataset of this study was collected and compiled using the following steps (illustrated in Figure 2.3):

1. ***Initial search.*** Using four initial search terms, "Artificial Intelligence," "AI," "machine learning," and "artificial neural network*," 149,428 records were retrieved from the WOS using *Topic* search.

2. ***"Snowball" the search terms.*** The keywords of the 149,428 records retrieved from the initial search were counted and ranked. Then, the top 50 keywords' synonyms were manually and semantically aggregated. For instance, "CNNs" is aggregated into "convolutional neural networks." The replaced synonyms can be found in Appendix A. A new ranking of the keywords was generated based on the aggregated keywords. The terms ranked in the top 20 keywords directly and unambiguously referring to AI were selected and compiled into the search terms for the next iteration. The terms that contain ambiguous meanings or potentially point to non-AI publications were discarded to avoid false positivity, for instance, "prediction," "data mining," and "big data." The search terms for the next round are listed in Appendix B.

3. ***Second iteration.*** The top search terms generated in the previous step (see

Appendix B) were fed into the second round of *Topic* search. It is worth noting that while conducting the second search round, the terms used in the previous round(s) were negated using the Boolean syntax "NOT" to avoid retrieving duplicate records. After the second round, in total, 324,666 records were retrieved.

4. ***Repeat the second and third steps*** until no new terms appear in the top 20 keywords after synonyms aggregation.

5. ***Confirm no top keywords are missing.*** In this study, only one round of the snowball process, in addition to the initial search, was conducted before no new terms emerged in the top 20 keywords. In other words, two iterations of the search were conducted. It is confirmed that after the two iterations of the search, each of the top 20 keywords (see Appendix C) had already either been searched in the previous round(s) or discarded because of ambiguity. Therefore, the results from the two iterations of the search can be considered sufficient to capture an effective representative set of AI publications and thus could serve as a dataset for research with interest in AI publications.

The metadata of the resulting 324,666 records for papers published between 1946 and 2020 was retrieved. Each record's metadata is coded in 73 two-character field tags, including title, authors' names, keywords, abstract, publisher, year published, cited references, citation count, and funding agency. The raw data was processed using *Metaknowledge*, a Python package for bibliographic analysis (McLevey and McIlroy-Young, 2017). It is worth noting that because this chapter considered knowledge combinations as recorded in a publication's referenced journals, 24,674 publications that did not reference journal articles or did not have the necessary information, such as publishing years, were removed in subsequent analysis. From the remaining dataset, 3,614 publications referencing only one journal article were discarded because

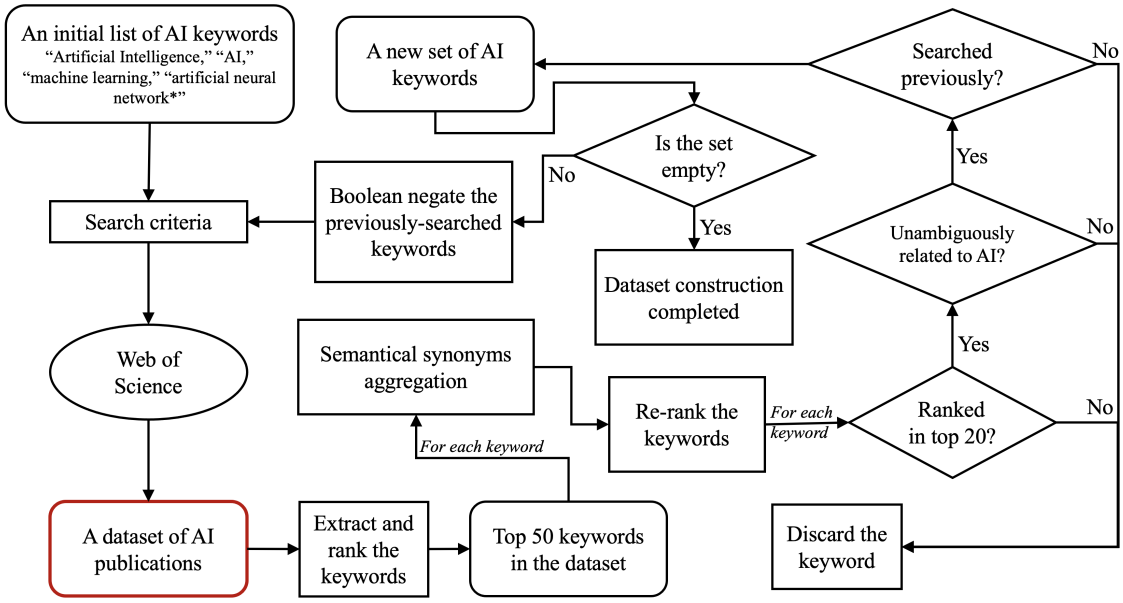


Figure 2.3: The process of constructing AI research publication dataset. The AI publication dataset was constructed through a snowball sampling process that consisted of several search rounds. This research started with an initial list of four AI-related keywords and incrementally incorporated the top keywords into search terms from the results of the previous search rounds. The ultimate, unprocessed raw dataset contains 324,666 records of papers published between 1946 and 2020.

no journal combinations can be formed with merely one single article. The final dataset contained 296,378 AI publications.

2.3.2 Measuring Typicalities of Knowledge Combinations

In this study, the typicality of a knowledge combination is defined as the observed frequency of two knowledge units, or components, co-appearing together, standardized against all the combinations during a select period.

For a scientific publication, this chapter considered cited papers as representing the previous knowledge combined in the publication. Specifically, each referenced journal is considered as a unit of knowledge, as a scientific journal is often domain-specific, representing the established body of existing knowledge in that domain. The pairwise combinations of the referenced journal articles, or co-cited journal pairs, are identified

as the recombination of existing knowledge. As an example, if an article referenced a paper from *Advances in Cognitive Psychology* and another paper from *Journal of Economic Geography*, it can be reasonably inferred that this article at least combines existing knowledge from the areas of cognitive psychology and economic geography in the creation of its knowledge.

For a co-cited pair of journal i and journal j in a given year t, its cumulative typicality, or z-score, can be computed as:

$$z_{(i,j),t} = \frac{\sum_{n=1946}^{t} x_{(i,j),n} - \mu_t}{\sigma_t} \tag{2.1}$$

where $x_{(i,j),n}$ is the observed frequency of journal-pair (i, j) in year n; μ_t is the mean value of cumulatively observed frequencies of all journal pairs up to year t, in other words, from 1946 to the year t; and σ_t is the standard deviation of the cumulatively observed frequencies up to the year t.

For instance, let us say a paper is published in 2013. It cites five journal papers published in three journals (A, A, A, B, C). Therefore, a list of four z-scores for journal pairs AA, AB, BC, and AC can be generated. In this example, the pairs of AA, AB, and AC occur three times; and the BC pair occurs once. Suppose only this paper is considered in the dataset and only one year (2013) is considered. In that case, the dataset's cumulative co-citation frequencies can be represented as an array: $\{AA : 3, AB : 3, AC : 3, BC : 1\}$. In this array, the mean value is 2.5; and the standard deviation is computed as 0.866. So, the z-score of each journal pair is computed using Equation 2.1 into an array: $\{z_{(A,A),2013} = 0.577, z_{(A,B),2013} = 0.577, z_{(A,C),2013} = 0.577, z_{(B,C),2013} = -1.732\}$.

In this way, the value of $z_{(i,j),t}$ indicates how frequently i and j are co-cited from 1946 to the year t compared to all the other journal pairs that were co-cited from 1946 to the year t. The higher $z_{(i,j),t}$ is, the more frequently they are co-cited cumulatively.

Thus, a high z-score indicates a frequent or *typical* combination. On the other hand, if $z_{(i,j),t}$ is negative, it indicates that i and j are co-cited less frequently than average, and it can be considered an *atypical* pairing. In the example described above, AA, AB, and AC can be seen as typical pairs because their z-scores are larger than 0, while BC can be seen as an atypical pair because its z-score is less than 0. Similarly, in the computed data, *IEEE Transactions on Power Systems* and *International Journal of Electrical Power & Energy Systems* have been referenced together 103,156 times from 1946 to 2019, and this pair has a cumulative z-score as high as 202 in 2019. In contrast, the *British Journal of Management* has been only co-cited once with *Fisheries* from 1946 to 2019. Their negative z-score (-0.045) indicates a rare and atypical pairing.

2.3.3 Knowledge Recombination Taxonomy

After a z-score is computed for every journal pair for every year, each AI publication is associated with a set of z-scores, describing the standardized cumulative frequency of a journal pair co-cited by this publication. Two statistical attributes are extracted for the z-scores of each publication to characterize its conventional and novel combinations, respectively, described as follows.

1. ***Tail Conventionality*** ***(TC_p)*** is the 80th percentile of the set of z-scores associated with a given paper p, featuring the typicality of the paper's right tail, where z-scores are relatively high, and pairs appear more conventional. The TC_p characterizes the paper p's tendency to combine conventional pairs. A paper can be considered to have high or low conventionality if its *Tail Conventionality* is in the upper or lower half of the *Tail Conventionalities* of all the papers published up to the given year. The higher the TC_p is, the higher degree of conventionality the paper has.

2. ***Tail Novelty*** ***(TN_p)*** is the 20th percentile of z-scores of a given paper p, characterizing the paper's more unusual journal combinations where novelty might dwell (Mukherjee et al., 2016). A paper can be seen as of high or low novelty if its *Tail Novelty* is below or above zero, respectively. The lower the TN_p is, the more novel the paper is.

A paper falls into one of four categories in a taxonomy developed by Uzzi et al. (2013) and described in detail by Mukherjee et al. (2016):

1. ***Darwin's Tower*** is a category where papers have both high conventionality and novelty. In other words, a paper in this category has a high TC and a negative TN. Papers in this category cite highly typical journal pairs and highly atypical journal pairs. The category is named after Charles Darwin, who used the same approach to present his idea in *On The Origin of Species* (Mukherjee et al., 2016).

2. ***Avant Garde*** is a category where papers have low conventionality but a high novelty. In other words, a paper in this category has a low TC and a negative TN. The category is named after the term for works of art that are unorthodox and radical.

3. ***Accepted Wisdom*** is a category where papers have high conventionality but a low novelty. In other words, papers in this category have high TC and positive TN. Papers in this category tend to cite journals that are always cited together, most likely from the same fields. At the same time, they tend not to cite journals from different disciplines.

4. ***Platypus*** is a category where papers have low conventionality and low novelty. In other words, the tail conventionality (TC) is in the lower half, while the 20th

percentile z-score (TN) is larger than zero. The category was named after the "neither-nor-like" quality of the *platypus* (Mukherjee et al., 2016).

It is worth noting that Uzzi et al. (2013) had selected the 10th percentile z-score and the median z-score to characterize a scientific publication's novelty and conventionality, which were defined as *Tail Novelty* and *Median Conventionality*, rather than the 20th and 80th percentiles as defined in this study. However, in their supplementary material, Uzzi et al. (2013) also described and provided justifications for alternative definitions for *Tail Novelty* including the 1st, 5th, and 20th percentiles, the last of which was selected in this study to describe *Tail Novelty*. In the data used here, the distribution of 20th percentile z-scores of all AI publications is less skewed than that of the other three (1st, 5th, or 10th percentile) while maintaining a sufficient describing power for the left tail of individual publications' z-score distributions. On the other end of the distribution, the 80th percentile z-scores (*Tail Conventionality*) is selected rather than the median z-scores (*Median Conventionality* as defined by Uzzi et al. (2013)) to characterize conventionality, because the median z-score of a publication often does not sufficiently differ from the 20th percentile. Thus, the 20th and median z-scores cannot capture a publication's distinct features of novelty and conventionality. Examining the distributions of z-scores of individual publications sampled from multiple groups stratified by the number of references reveals that the 80th percentile is among the values that can characterize the tendency of the right tail effectively. Therefore, this chapter selects the 80th percentile z-score of a publication to capture the right tail instead of choosing the median to capture the middle mass of the z-score distribution. As a result, this conventionality measurement is defined as *Tail Conventionality* instead of *Median Conventionality*.

A Case Example

To demonstrate how this method applies to a single record, let us take the paper

Deep Learning by LeCun et al. (2015) as an example. This paper was published in *Nature* in 2015 and reviewed the current development of deep learning. Being cited 14,357 times[3] when the record was retrieved in July 2020, it was ranked as the fifth-highest cited paper in the dataset. This paper references 103 other publications, with 69 journal articles and 45 distinct journals observed.

First, the z-scores of all the journal pairs co-cited up to 2015 were computed. Up to 2015, a total of 16,825 journals were cited by AI publications, making up 3,571,994 distinct journal pairs with an average co-citation count of 19.6.

Then, the journal pairs co-cited by the paper *Deep Learning* were selected. Table 2.1 presents a sample of the paper's journal pairs and their z-scores. The higher the z-score is, the more frequently the two journals were co-cited by AI publications up to 2015. As shown in Table 2.1, *Science* and *Nature* are co-cited very frequently – 55,913 times. Domain-specific journal pairs with high z-scores are often in the same areas, such as *Machine Learning* and *Neural Computing*, co-cited 12,296 times. The journal pairs with low z-scores are more likely made up of journals from different domains, for instance, *IEEE Journal of Solid-State Circuits* and *Journal of Chemical Information and Modeling*, which belong to the areas of physics and chemistry respectively. These two journals were co-cited only once by AI publications up to 2015.

The distribution of the z-scores associated with this paper is shown in Fig. 2.4. The 20th percentile of z-scores or TN (*Tail Novelty*) is computed as 0.19, larger than 0, while the 80th percentile z-scores or TC (*Tail Conventionality*) is computed as 16.86. This is higher than the median value of all TCs up to 2015, which was

[3]The citation count (14,357) is from the "Z9" tag of the WOS database, which represents "Total Times Cited Count" that includes WOS Core Collection, Arabic Citation Index, BIOSIS Citation Index, Chinese Science Citation Database, Data Citation Index, Russian Science Citation Index, SciELO Citation Index. If only WOS Core Collection is considered, this paper is cited 13,619 times.

Table 2.1: Selected samples of the journal pair frequencies up to 2015 and z-scores for an illustrative paper, *Deep Learning* by LeCun et al. (2015).

Journal pairs	**Co-citation Frequency up to 2015**	**z-score**	
IEEE T PATTERN ANAL[a], PROC CVPR IEEE[b]	23,105	64.8	More typical combinations
NATURE, SCIENCE	55,913	156.84	
MACH LEARN[c], NEURAL COMPUT[d]	12,296	34.45	
ARTIF INTELL[e], COMMUN ACM[f]	6,133	17.15	
Z-score = 0 means observed co-citation frequency up to 2015 is as likely as average.			
BIOINFORMATICS, J FIELD ROBOT[g]	5	-0.04	More atypical combinations
IEEE T AUDIO SPEECH[h], J CHEM INF MODE[i]	8	-0.03	
IEEE J SOLID-ST CIRC[j], J CHEM INF MODEL	1	-0.05	

[a] *IEEE Transactions on Pattern Analysis and Machine Learning.*

[b] *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[c] *Machine Learning.*

[d] *Neural Computing.*

[e] *Artificial Intelligence.*

[f] *Communications of the ACM.*

[g] *Journal of Field Robotics.*

[h] *IEEE Transactions on Audio Speech and Language Processing.*

[i] *Journal of Chemical Information and Modeling.*

[j] *IEEE Journal of Solid-state Circuits.*

computed as 3.14. Therefore, this paper has a low novelty and high conventionality and is categorized as an *Accepted Wisdom*.

2.3.4 Assessing Scientific Impact

With AI papers categorized using the taxonomy described above, this chapter then investigated how a paper's categorization relates to its scientific impact.

In scientometrics research, citation count has arguably been considered one of the most reliable indicators for scientific quality and impact, especially at the highly-cited end of the distribution (Phelan, 1999). However, both citation counts and citation

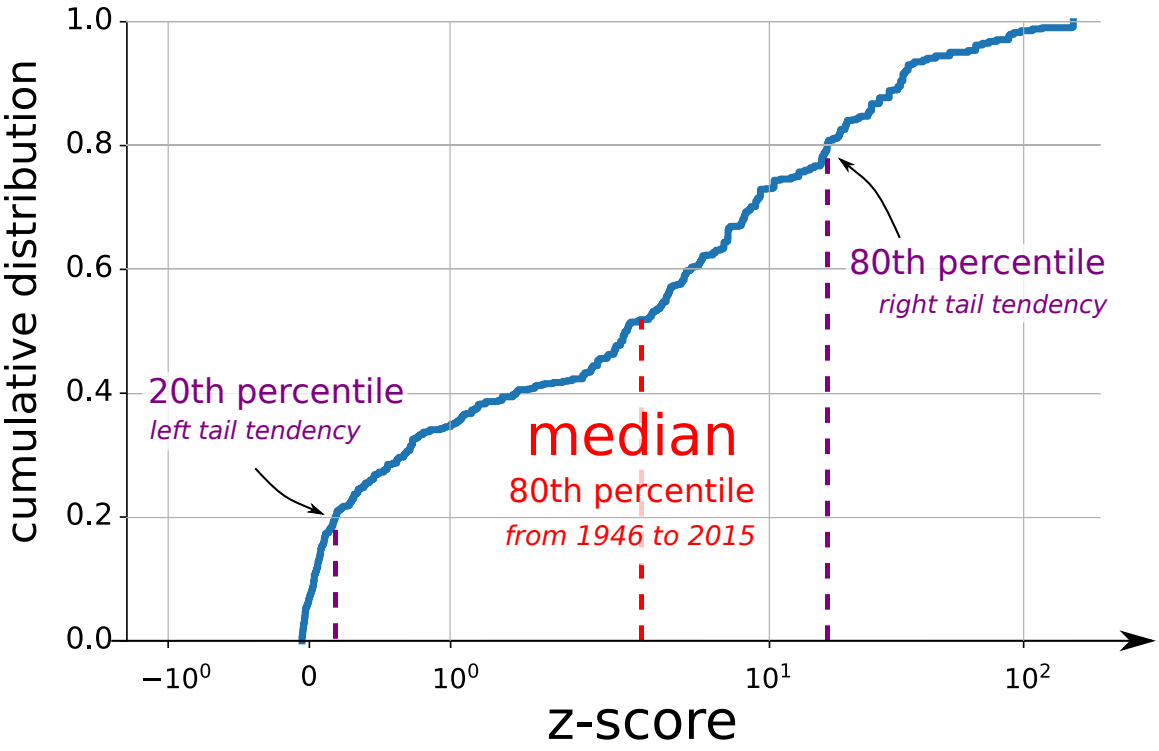


Figure 2.4: Cumulative distributions of z-scores associated with the paper *Deep learning* by LeCun et al. (2015). The z-score of each journal pair is computed using Equation 2.1. The 20th percentile of z-scores that features the left tail tendency (unusual or atypical pairing) is calculated as 0.19, larger than 0, indicating a low novelty. In contrast, the 80th percentile that characterizes the right tail tendency (more usual or typical pairing) is computed as 16.86, greater than the median value of the 80th percentiles of all AI publications published between 1946 to 2015, which is computed as 3.14. Therefore, this paper has a low tail novelty and high conventionality. It is categorized as an *Accepted Wisdom*.

rates (those received in a given year) do not stay constant over time. Typically, a paper's citation rate grows slowly during the first 5 to 20 months after publication and accelerates until a saturation plateau is reached. After that, citation rates tend to decrease (Ponomarev et al., 2012). This pattern makes it problematic to directly compare the citation counts of two papers published in different years. Instead, researchers often select the mean annual citation rate to compensate for this time effect of citation counts (Das et al., 2019; Unger et al., 2018). The mean annual citation rate (from now on referred to as "annual citation" or AC for short) of a given publication p is computed as the ratio of total citation counts up to year t

divided by the number of years since its publication, as in Equation 2.2.

$$AC_{p,t} = \frac{\sum_{n=0}^{t} CR_{p,n}}{t} \tag{2.2}$$

where t is the number of years since publication; $CR_{p,n}$ is the citation rate of paper p in year n.

This study uses the *annual citation* to measure an AI paper's average scientific impact. A paper ranked among the top papers in annual citations is considered to have a high scientific impact, hence referred to as a "hit papers," or simply a "hit" (Mukherjee et al., 2016).

2.4 Results

2.4.1 Data Description

The dataset used in this research contains metadata from 296,378 AI publications spanning from 1946 to 2020. It is worth noting that some early years are absent from the dataset, indicating no relevant publications in those years. These gaps include 1947 to 1951, 1952 to 1958, and 1965. The publications in the dataset reference 19,474 scientific journals, making up 7,779,502 distinct journal pairs. They collectively represent 5,912,959 citation counts.

The dataset contains two types of publications:

1. Publications that contribute to AI directly, for instance, a paper introducing a new architecture of Artificial Neural Networks (ANNs) for more efficient learning.

2. Publications that engage AI as a research tool, including:

   (a) Studies that use AI as a tool to produce research output, for instance, an

astrophysical study on gravitational wave detection that uses the random forest to filter out noise signals (Biswas et al., 2013),

(b) Studies that develop AI methods or tools for research, for instance, a study that develops an approach using support vector machines for identifying cancer biomarkers (Chen et al., 2011).

The distributions of the number of previous works referenced in AI publications are shown in Fig. 2.5. The blue histogram represents the distribution of the number of references of each AI publication regardless of their publishing type. It includes but is not limited to journal articles, books, thesis, preprints repositories, blogs, and web pages. The orange histogram represents the number of journal articles referenced by each AI publication; the green histogram represents the number of distinct journals referenced by each AI publication. Distinct journals refer to the non-repetitive list of journals referenced by a publication, which should be smaller than or equal to the number of journal articles referenced because some articles may be from the same journal(s).

It is revealing to distinguish the number of journal articles and the number of distinct journals referenced by a publication because, while the former demonstrates the number of previous works, the latter illustrates previous works that are sufficiently distant. It is worth noting that, as mentioned in Section 2.3.1, the publications that reference only one journal article are removed. Therefore, in Fig. 2.5, the blue and orange histograms both start from 2.

I found, on average, that an AI publication has 41 references, among which about 28 are journal articles from roughly 16 distinct journals. The 40% difference between the number of journal articles and distinct journals implies that AI publications often cite articles from the same journals. For example, one AI publication with the longest

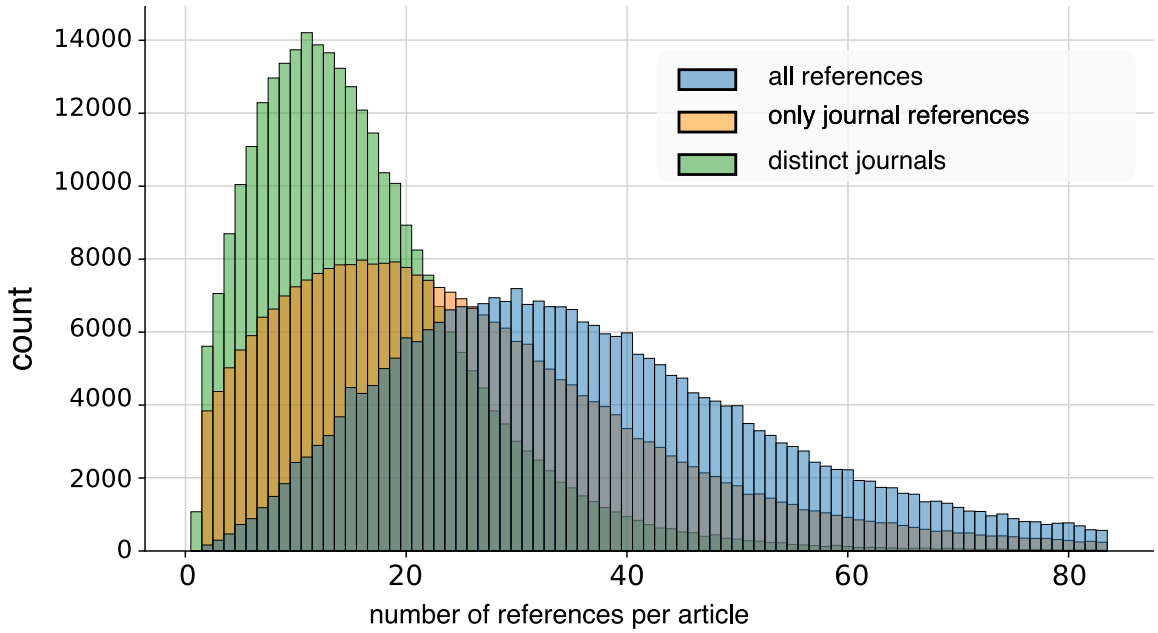


Figure 2.5: Histogram plot showing the distribution of AI publications' reference count (blue bars), referenced journal-article count (orange bars), and distinct journal count (green bars). Long-tails are cutoff at 84, corresponding to the 95th percentile of reference counts.

reference list is a review regarding analytical chemistry by Crouch et al. (1998). This publication referenced 1,427 previous works, among which 1,304 are journal articles. Nevertheless, these articles are published in only 305 distinct journals, and more than three-quarters of the articles are from the same journals as the rest.

2.4.2 Exponential Growth of AI Publications

I found, formally, that the number of new AI papers published each year is growing exponentially, as shown in Fig. 2.6. This result was reported in a previous paper (Wang et al., 2022a). Before 1984, AI's annual publication count never exceeded 100. It took another ten years for the number to grow beyond 1000. In 2019, over 50,000 publications were recorded. Linear regression between years and natural logarithms of annual publication counts resulted in a coefficient of 0.1736 with an R-squared of

0.97 and a p-value of 1.52×10^{-48}.

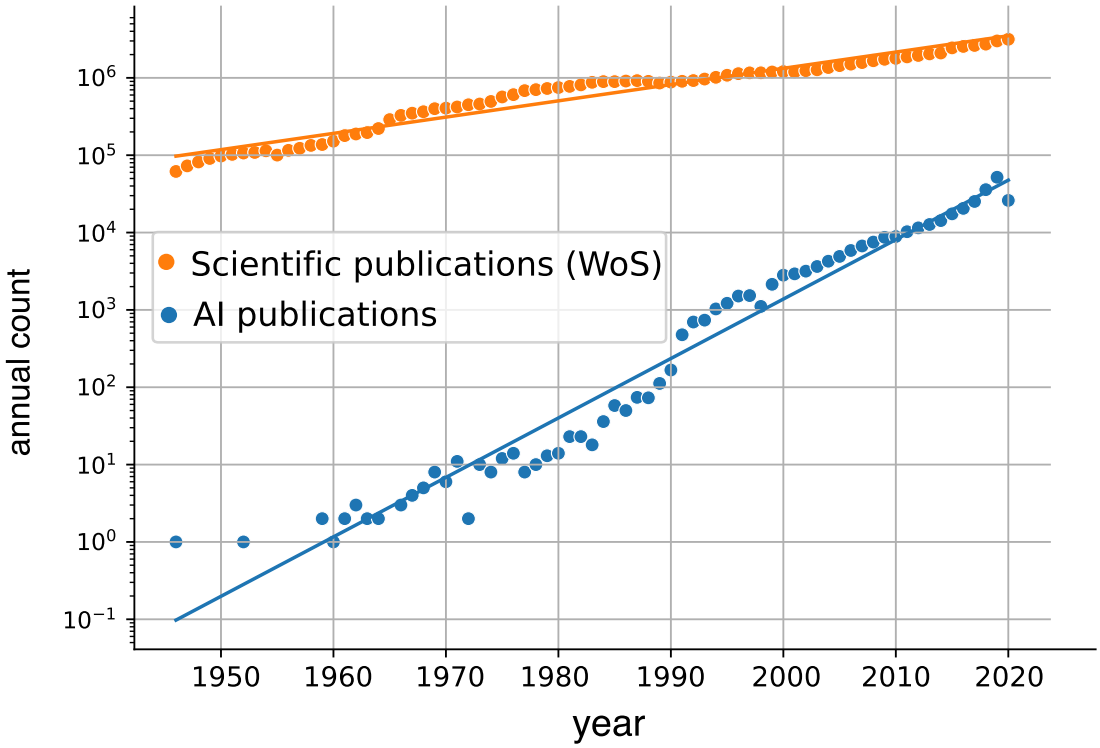


Figure 2.6: Time Series of the number of AI publications and the number of scientific publications in general (log scale). AI publication is growing faster than scientific publication in general.

It is noticeable that there is a broader trend in exponential growth amongst scientific publications, as shown in Fig. 2.6. However, the slope of the fitted line for the natural logarithm of the annual count of all scientific publications against time is 0.0484 ($R^2 = 0.949$, $P = 4.50 \times 10^{-49}$), less than 30% of the magnitude for the coefficient of AI publications. Consequently, the ratio of AI publications compared to scientific publications is increasing, as shown in Fig. 2.7. By 2019, AI publications made up 1.73% of all scientific publications according to the data used here for analysis.

Fig. 2.6 shows that there were a few years where the number of AI publications fluctuated and declined slightly, particularly in the early 1970s and late 1980s, cor-

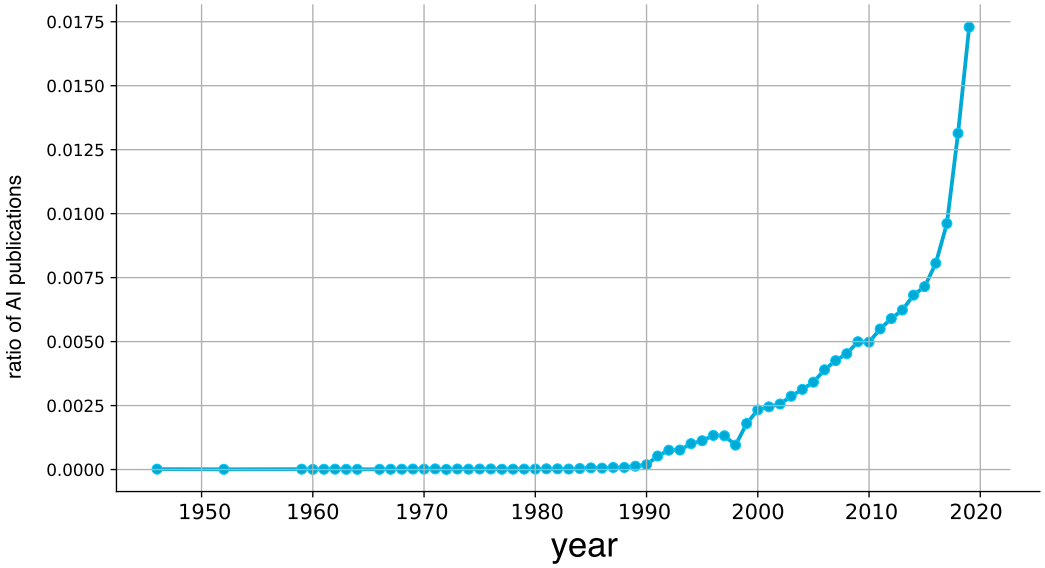


Figure 2.7: Time Series of the ratio of AI publications in scientific publications in general. AI publications account for an increasingly larger share of scientific publications. By 2019, the ratio reaches 1.73%. In other words, in every 100 scientific publications, about two are related to AI.

responding to the historical periods referred to as "AI winters," when funding and interests in AI research were drastically reduced (Howe, 2007; Hendler, 2008). As a result, AI publication output in these periods is diminished. Nevertheless, AI has seen steady and robust growth over the past two decades. Indeed, the world appears to be in the midst of an "AI spring" (Maclure, 2020).

2.4.3 The Skewness of AI Research

As mentioned in Section 2.4.1, AI publications in the analyzed dataset have generated almost 6 million citation counts in total. On average, each publication is cited almost 20 times. However, the distribution of citations received by AI publications is extremely skewed. This phenomenon has been observed and reported frequently in scientific publications in general (Seglen, 1992; Albarrán et al., 2011). I found that in AI publications, 19.02% have not been cited at all. 9.94% have been cited once,

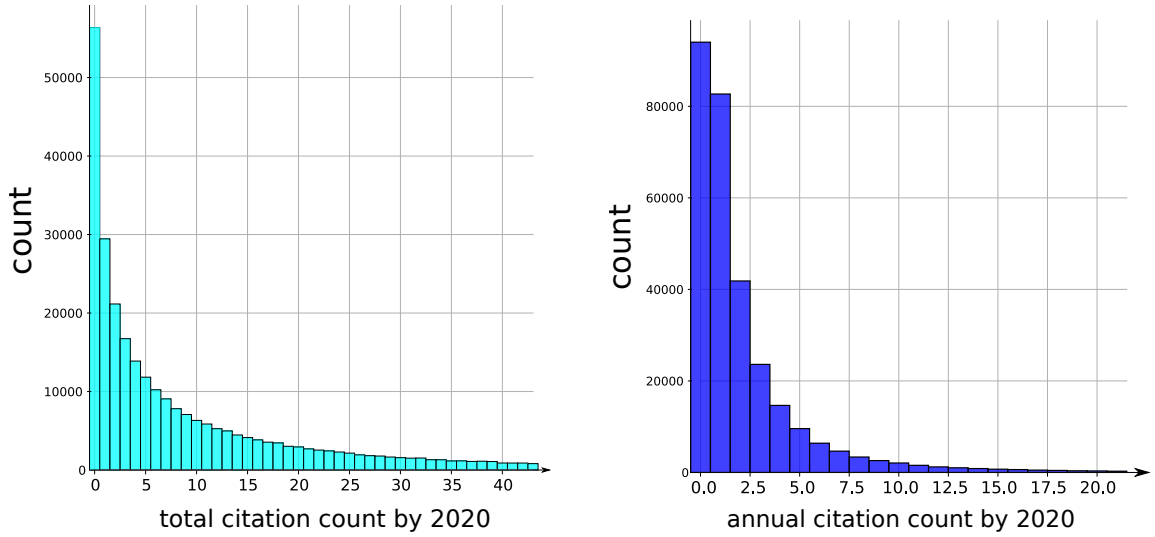
and 50.41% have no more than five citations. Moreover, AI publications can serve as another empirical demonstration of the Pareto principle of the 20/80 rule, which states that roughly 80% of consequences come from 20% of causes (Pareto et al., 1964). The AI publications ranked among the top 20% in citation counts and generated 4,632,360 citations, making up 78.3% (almost 80%) of all citations received by AI publications. The top 10%, 5%, and 1% produced 62.34%, 48.74%, and 26.49% of all citations respectively. This skewness is shown in Fig. 2.8a.

In addition to citation counts, the distribution of annual citations, as Fig 2.8b shows, remains considerably skewed. Slightly over half (53.08%) of all AI publications have no more than 1 citation each year. The top 20% AI publications in annual citations ($AC \geq 3.17$) collectively generated 71.40% of the sum of annual citations of all AI publications, again, very close to the Pareto principle of 20/80 rules. The skewed distribution of AI publications' citations illustrates that, like other areas of science and technology, many AI papers have no citations, at all and very few have produced substantial impacts.

2.4.4 Conventional Knowledge Driving AI Growth

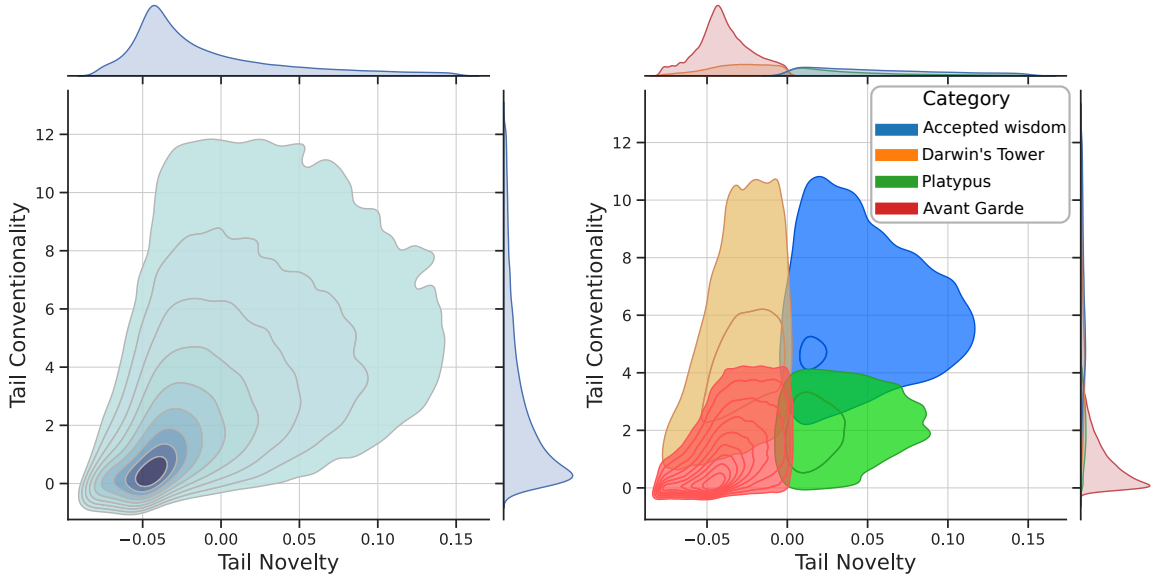
I implemented the knowledge recombination taxonomy described in Section 2.3.3 by computing z-scores for all referenced journal pairs and categorizing each AI publication into one of the four categories based on two statistical features of its z-scores (TN and TC). The joint distributions of TN and TC are shown in Fig. 2.9. It is worth noting that there are areas of overlap between categories in Fig. 2.9b, especially along the y-axis. This is because median TC, the value used to classify high and low conventionality, varies from year to year.

I found the category of *Accepted Wisdom*, which combines high conventionality with low novelty, accounts for the largest share (46.4%) in AI research. Another



(a) (b)

Figure 2.8: (a) Distribution of AI publications' citation counts and (b) average annual citation counts by 2020. Long tails are cut off at 90% and 99%, respectively. The two panels both show a significant level of skewness.



(a) (b)

Figure 2.9: Joint distribution of *Tail Novelties* (*TN*s) and *Tail Conventionalities* (*TC*s) of AI publications. Panel (a) shows the distribution of *TN* and *TC* of all AI publications. Panel (b) shows the distribution of the two in each of the four categories. Long tails are cut off. It is worth reiterating here that the *higher* the value of *TN* of a publication is, the *less novel* its left tail is.

category with high conventionality is Darwin's Tower, which makes up 11.5%. These two highly conventional categories collectively constitute more than half (57.9%) of AI publications. The time series for the percentage of publications each category accounts for, Fig. 2.10, reveals a significant trait: *Accepted Wisdom* in AI research has been advancing steadily over the last three decades. By 2020, new publications categorized as *Accepted Wisdom* reached 52.4%. *Avant Garde*, the second largest category, has been stable with a slight decrease in recent years. *Darwin's Tower*, the category that mixes high novelty with high conventionality, has been diminishing, from above 20% in the 1990s to only 6% in 2020. This can lead to the conclusion that new knowledge that relies heavily on conventional combinations has become AI's most significant driving force.

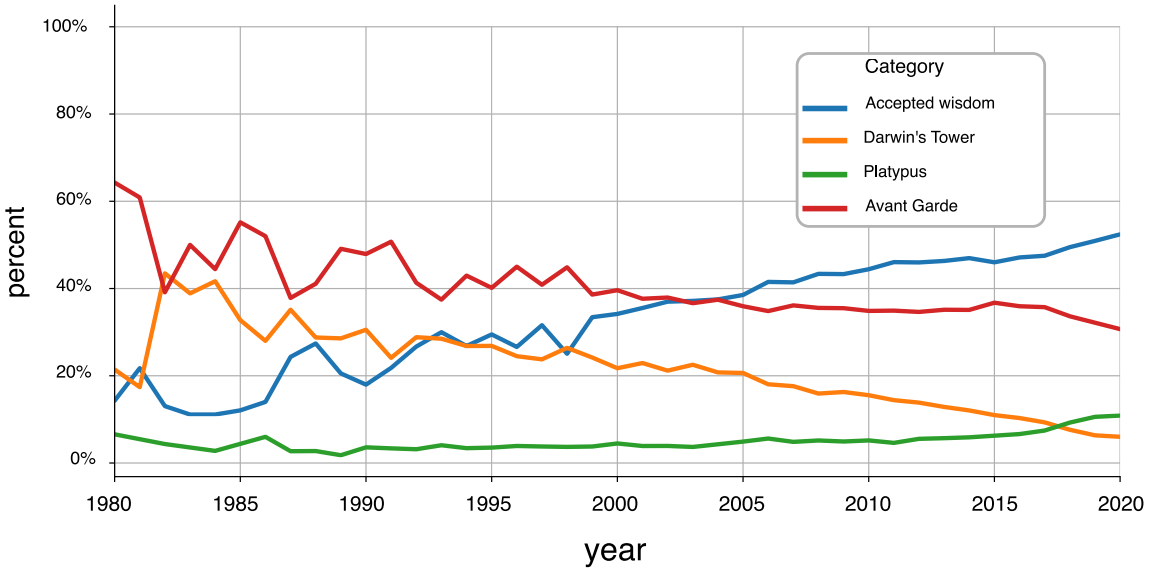


Figure 2.10: Time Series of AI publications' composition regarding knowledge recombination (1980-2020). *Avant-Garde* and *Darwin's Tower* have been declining while *Accepted Wisdom* is increasing in its share and has become the largest category.

2.4.5 Conventional Knowledge Exerting Greater Impact

As discussed in Section 2.4.3, the impact of AI publications, measured using citations, exhibits a very skewed distribution, regardless of whether total citation count or annual citation count is considered. The question is whether an AI publication's taxonomic category is related to its scientific impact. The answer highlights the role of *Accepted Wisdom* in predicting impact.

Among those publications that are highly ranked for annual citations, *Accepted Wisdom* occupies an even greater percentage (for instance, as high as 65.57% in the top 1%). Furthermore, the higher the group's rank, the larger its share becomes. As addressed in Section 2.4.4, *Accepted Wisdom* makes up 46.39% of all AI publications. In the AI publications ranked in the top 10% in annual citations, this category accounts for 59.73% of publications. In the top 5% and top 1%, this percentage rises to 61.96% and 65.57%, respectively. In addition, in the top 10 papers, seven (70%) are categorized as *Accepted Wisdom*. In contrast, *Avant Garde* papers that feature a high novelty and low conventionality decreased relative to other categories, with 34.68% in all papers yet declining to only 18.72% in the top 1% AI publications. Moreover, there is no *Avant Garde* paper in the top 10 papers. On the other hand, the share of *Darwin's Tower* category tends to remain stable across percentiles, its percentage not fluctuating much no matter if considering the top 1%, 5%, 10%, or all papers. The percentage of each category in different groups is shown in Table 2.2.

Table 2.2: Percentages of Each Category in Top AI publications

Category	**% in all**	**% in top 10%**	**% in top 5%**	**% in top 1%**	**% in top 10 papers**
Accepted Wisdom	46.39%	59.73%	61.96%	65.57%	70%
Darwin's Tower	11.54%	10.99%	11.19%	10.85%	20%
Platypus	7.40%	5.88%	5.71%	4.86%	10%
Avant Garde	34.68%	23.39%	21.13%	18.72%	0%

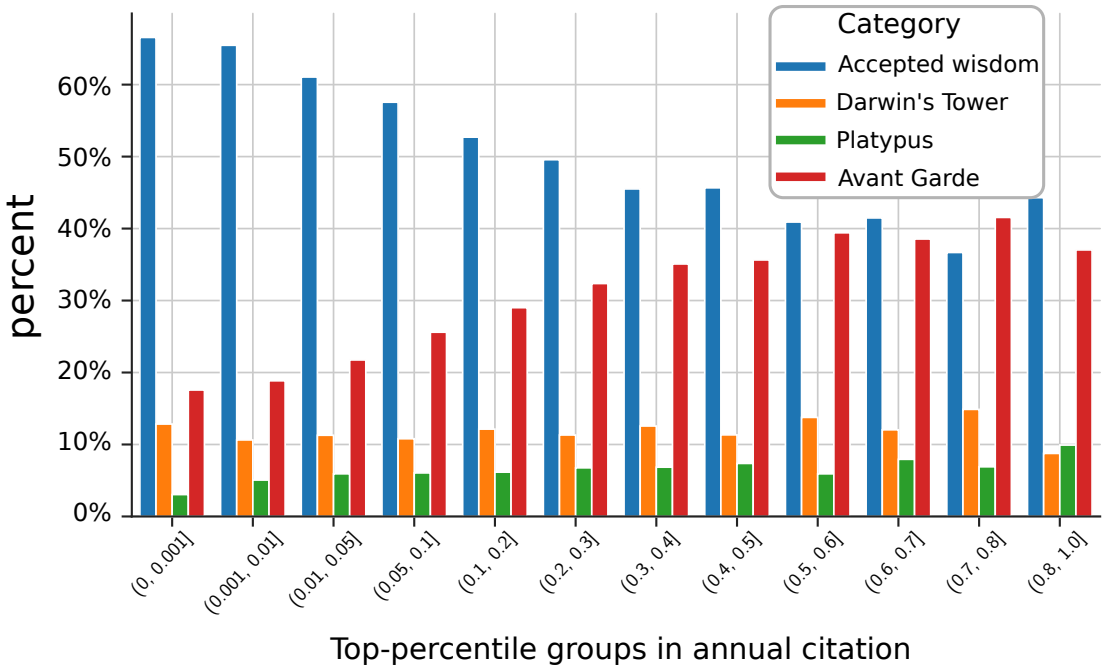


Figure 2.11: Percentages of each category in AI publications grouped by annual-citation-percentile (measured between 0 and 1). The leftmost group represents the group of AI publications with the highest average annual citations, ranked in the top 0.1%, while the rightmost group represents the publications ranked among the bottom 20% (the 80th-100th percentile) or the least cited annually.

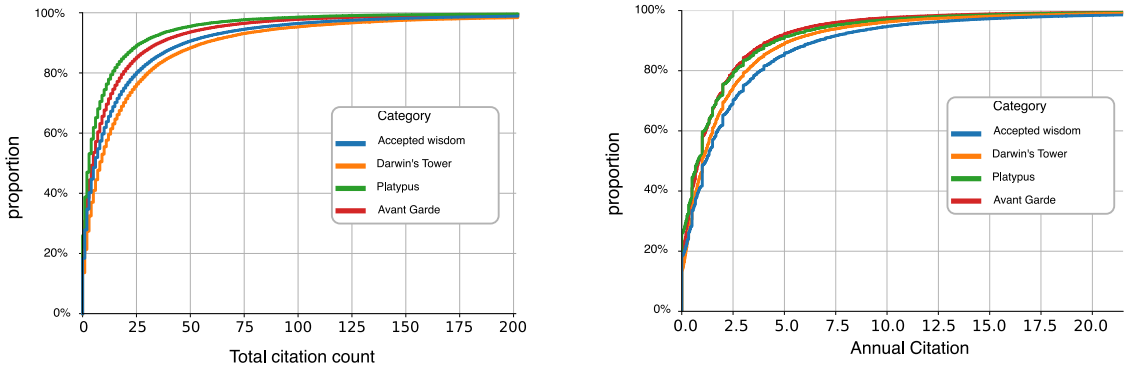
Figure 2.11 provides additional evidence for *Accepted Wisdom*'s dominance in the top tiers of AI papers. In Fig. 2.11, AI publications are arranged into 12 groups based on the ranking percentile in annual citations. Then, the percentages of the four taxonomic categories are plotted for each group. For example, the leftmost group consists of AI publications of which the annual citations are ranked in the top 0.1% (0 to 0.001), and the second group comprises publications that are ranked from top 1% to top 0.1% (0.001 to 0.01), and so on. Fig. 2.11 clearly shows that *Accepted Wisdom* (blue) is increasing towards the left, while the *Avant Garde* category is decreasing and *Darwin's Tower* and *Platypus* are relatively stable.

Moreover, as shown in Table 2.3 and Figure 2.1, *Accepted Wisdom* accounted for more than half (51.31%) of all citations, larger than its share in publication count (46.39%). It is also highly ranked regarding mean and median annual citations.

However, the *Darwin's Tower* category has the highest mean citation count (28.41) and median citation count (8). This suggests that, although *Darwin's Tower* makes up a much smaller proportion than *Accepted Wisdom* in AI publications (as shown in Table 2.2), it is less skewed regarding citations. In other words, *Darwin's Tower* has a smaller proportion of publications ranked at the bottom. Fig. 2.12a confirms this by showing that among the four categories' cumulative distributions of citation counts, the orange curve that represents *Darwin's Tower* has the lowest curvature, indicating a smaller degree of skewness in citations.

Table 2.3: Selected Citation Features of the Four Categories

Category	**Total count**	**Total citation**	**Citation percentage**	**mean citation**	**median citation**	**mean annual citation**	**median annual citation**
Accepted Wisdom	**137,481**	**3,034,185**	**51.31%**	22.07	6	**3.04**	**1.17**
Darwin's Tower	34,203	971,780	16.43%	**28.41**	**8**	2.54	1.00
Platypus	21,922	270,211	4.57%	12.33	3	1.97	0.78
Avant Garde	102,772	1,636,783	27.68%	15.93	5	1.86	0.83



(a) (b)

Figure 2.12: Cumulative distribution of (a) citation counts and (b) annual citations of the four categories.

To further verify the relationship between the taxonomic categorization of an AI publication and its scientific impact, it is necessary to conduct a regression analysis. Because citation count is essentially discrete count data with a skewed distribution, a

Poisson regression is generally applicable. However, the variances of citation counts in each category are substantially larger than the mean values (see Appendix D). Due to this over-dispersion, the assumption of Poisson regression is not satisfied. A negative binomial regression is considered a more suitable model to compensate for such an over-dispersion. With the control of years (t), the following regression equation can be assumed:

$$\ln Y = \beta_0 + \beta_1 C + \beta_2 t + \epsilon \tag{2.3}$$

where Y denotes the number of citation counts an AI publication receives as recorded in the dataset by 2020. C denotes the category of the AI publication (a categorical variable), and t denotes the year of publication. The summary of negative binomial regression results of Equation 2.3 can be found in Appendix E. The coefficient β_1 for the four possible values of variable C are computed as 1, -0.29, -0.46, -0.52 for *Accepted Wisdom*, *Darwin's Tower*, *Platypus*, and *Avant Garde* respectively, with Accepted Wisdom as the controlled value and the other three as treatments. All p-values are smaller than 0.001. The coefficients indicate that, among the four categories, an *Accepted Wisdom* tends to have the highest citation counts, followed by *Darwin's Tower*, *Platypus*, and *Avant Garde*.

In summary, *Accepted Wisdom* papers collectively exert a more significant impact. Moreover, such papers take the highest share in the top percentiles of AI publications regarding annual citations. Regression analysis confirms that an AI publication in *Accepted Wisdom* tends to have the highest citation counts. These results imply that research activity that exploits conventional combinations of existing knowledge without introducing radically novel ideas (*Accepted Wisdom*) not only dominates AI research in general but also creates a higher scientific impact.

2.4.6 Accepted Wisdom Has the Highest Hit-rate

I have reported in Section 2.4.5 that *Accepted Wisdom* occupies the largest share in highly-impactful AI publications, followed by the *Avant Garde*, *Darwin's Tower*, and *Platypus* categories. Is the categorization of an AI publication strongly associated with its chance of becoming highly impactful? Does being a paper categorized as *Accepted Wisdom* increase the chance of it becoming highly cited? To answer this question, it is helpful to examine the cumulative distribution of the percentiles of each category in terms of annual-citation-ranking, as shown in Fig. 2.13, in which panel (b) is a proportion of the panel (a) cut off at 0.1 on the x-axis.

I first ranked all AI publications based on their annual citations. The higher a paper's annual citation is, the smaller its ranking number will be. With this approach, the paper with the highest annual citation is ranked as "1." Publications with the same annual citations are ranked the lowest in the group. Then, the rankings are normalized to range between 0 and 1 so that the normalized ranking of a publication illustrates what proportion of the dataset has higher annual citations than this one. For instance, a paper with a normalized rank of 0.4 indicates that 40% of AI publications have higher annual citations than it has. In turn, this paper can be considered as being located at 40% from the top in terms of annual citations. The cumulative distributions of this normalized ranking of each category can illustrate the "hit probabilities" of each category (Fig. 2.13a). Points in the area above $y = x$ represent the hit probabilities higher than the background probability ($y = x$) and vice versa. Here, background probability refers to the cumulative distribution of papers' rankings if only one category exists. Based on the methodology description above, it is not difficult to derive that a paper's background hit probability equals its normalized ranking. Therefore, the background probability distribution can be described as the

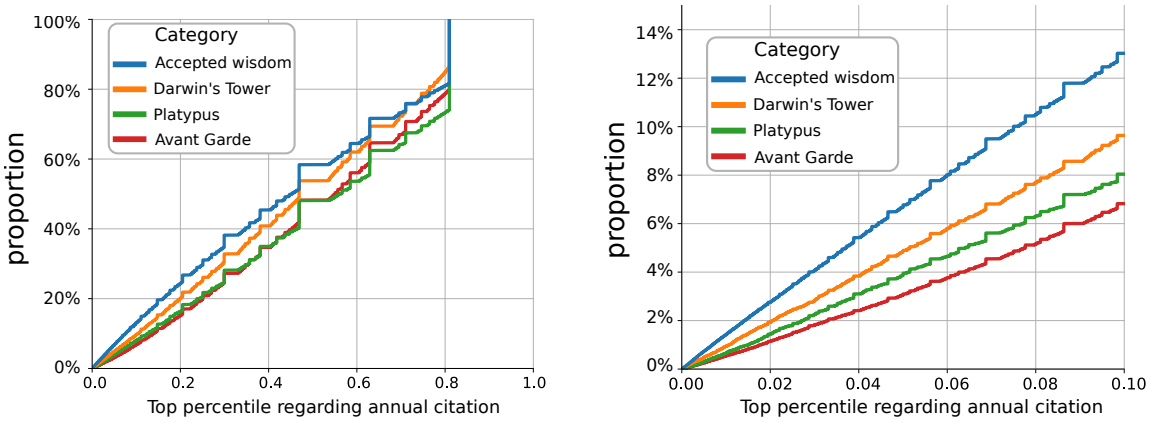
line $y = x$.

When examining Fig. 2.13b where only the top 10% impactful AI publications are considered, this research finds that being in the *Accepted Wisdom* category is related to a higher probability of a paper becoming a hit-paper (referred to as "hit probability"), because the blue line that represents *Accepted Wisdom* has a higher slope than the other three. While the orange line that represents *Darwin's Tower* is very close to the background ($y = x$), the other two (green and red) are both clearly below $y = x$. This observation is again supported by Table 2.4 that presents the hit probability of each category in a sample of four top tiers (top 10%, top 5%, top 1%, and top 0.1%). For all the four top tiers, hit probabilities in the *Accepted Wisdom* category are all higher than background probabilities (10%, 5%, 1%, 0.1%). The hit probabilities of the other three categories are all lower than background probabilities with only one exception, in which the top 0.1% hit probability of *Darwin's Tower* is 0.11%, slightly higher than the background probability (0.1%).

Table 2.4: Hit Probabilities of the Four Categories

Category	Top 10% probability	Top 5% probability	Top 1% probability	Top 0.1% probability
Accepted Wisdom	**13.03%**	**6.68%**	**1.42%**	**0.14%**
Darwin's Tower	9.64%	4.85%	0.95%	**0.11**%
platypus	8.04%	3.86%	0.66%	0.04%
Avant Garde	6.82%	3.05%	0.54%	0.05%

I further examined the time series of hit probabilities of each category. The top 10% hit probability time series is shown in Fig. 2.14, where the dashed line represents the background probability (10%). It is clear from Fig. 2.14 that the *Accepted Wisdom* category (blue) has become the category with the highest hit probability since the 2000s. It is worth noting that the decline of all four lines at the end of the series is due to the timing of the data collection (2020), which did not allow enough time for annual citations of the latest publications to accumulate to a degree comparable to



(a) (b)

Figure 2.13: Cumulative distribution of AI publications' percentiles of annual citation rankings in each category. Annual citation ranking is ranked so that the higher the annual citation of a publication is, the smaller the ranked number is, and the more likely it will be positioned towards the left. The publications with the same annual citations are ranked with the lowest in the group. Panel (a) shows the full distribution from the 0th percentile to the 100th percentile, while panel (b) shows a selected portion of the panel (a) where the percentile is from the 0th to the 10th or is in the top 10%. Panel (b) can be considered an illustration of the hit probability of the four categories.

those of earlier publications.

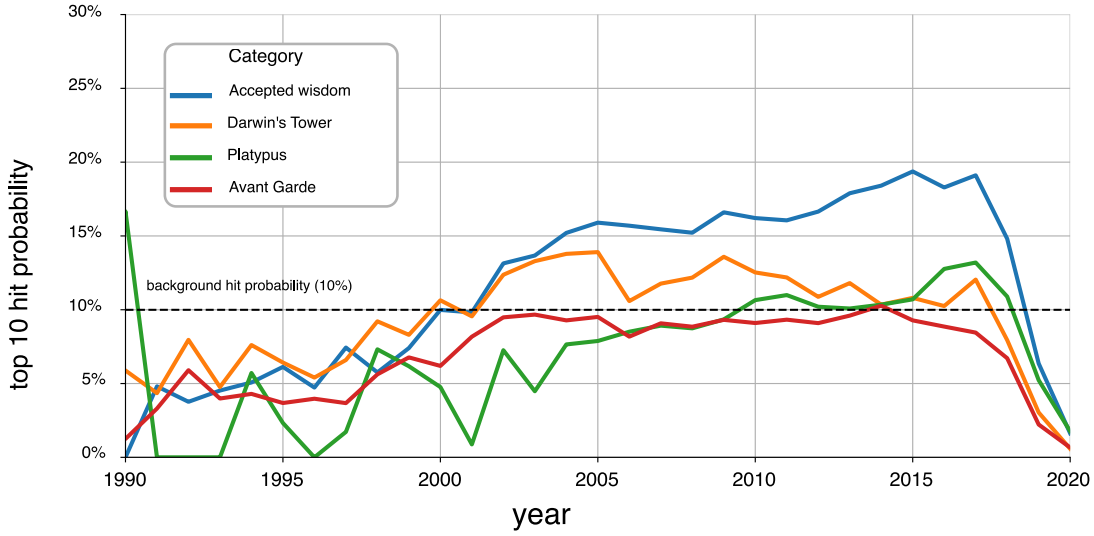


Figure 2.14: Time series of the probabilities of each category of becoming the top 10% regarding annual citations (1990-2020), i.e., top 10% hit probability. The horizontal dashed black line represents the background probability (10%).

2.5 Discussion

The results presented here provide clear evidence that AI research, so far, has mainly advanced in an incremental fashion and has not drastically deviated from the developmental experience of other scientific domains.

The finding that *Accepted Wisdom* (publications characterized by high conventionality but low novelty) is the most prevalent and most impactful category in AI research, does not align with the results from previous work by Uzzi et al. (2013) and Mukherjee et al. (2016) reporting that *Darwin's Tower* has created the most impact in science in general. There may be several reasons for this discrepancy. First, unlike Uzzi et al. (2013) or Mukherjee et al. (2016), who utilized the entire WOS database, publications in a small domain (AI) is considered. Therefore, the z-scores were computed within a different scope. A typical combination in AI may be considered novel in another domain and vice versa. Secondly, this research compared the observed frequency of a pair of journals with the expected frequency by *averaging* all frequencies cumulatively. In contrast, Mukherjee et al. (2016) compared the observed frequency of a pair with the average frequency of the same pair in ten *synthesized* networks. This difference may result in a different range of values and distributions for z-scores. However, because the two ways of computing z-scores both attempt to capture how typical a pair is, their qualitative interpretation should remain the same. That is to say, in both cases, a high z-score indicates a typical pairing, while a low z-score indicates an atypical pairing. Thirdly, instead of using the 10th percentile and the median z-scores to measure novelty and conventionality, respectively, this research chooses the 20th and 80th percentiles. This chapter has offered justifications in Section 2.3.3 regarding this adaptations to address the extreme skewness of z-score distributions. This skewness is especially substantial towards the left or "atypical"

end of the spectrum and partly resultes from how z-scores were computed in this study. As a result, the 20th and the 80th percentiles can be considered sufficiently different and enough to characterize the tendency of the two ends of a z-score distribution. Accordingly, the result of this study should be considered a complementary case study instead of a contradiction of the research by Uzzi et al. (2013) and Mukherjee et al. (2016). It should be interpreted only in the scope of AI.

AI research output has been growing faster than science output in general, as shown in Fig. 2.6. However, this study does not suggest that this observation can serve as evidence for the argument that AI is revolutionizing invention (the creation of novelty). Instead, this fast-growing trend in AI publication is likely the result of increased funding and interest that usually requires or encourages deliverable in the form of scientific publications, which in turn facilitates new journals, conferences, and submissions to established journals. Such interests may be further driven by heightened expectations of AI's economic potential or strategic significance rather than being a response to already successful implementations.

I note that the present study is based only on AI publications that went through standard academic publishing practice particularly the peer-review process. That is to say, reports, non-reviewed publications, and non-technical discussions (such as arXiv, patents, Github records, and social media discussions) are not included in the analysis. AI researchers are increasingly choosing to publish their papers on non-traditional platforms like arXiv. Organizations, individual practitioners, and hobbyists tend to publish AI codes on Github without writing papers. Those efforts are growing in scale and should not be overlooked in future assessments of AI invention.

Because of the limitation of the selection of data, this research does not dismiss the possibility that, in the conclusion that AI science aligns with "normal science," what is really "normal" here is not AI *per se*, but science. In other words, regardless of how

revolutionary particular research may be, to survive peer review and editorial scrutiny, authors are expected not to deviate from many explicit and implicit norms, including but not limited to the necessity to credit previous relevant work, how many previous works should be cited, what kind of previous works are considered credible, and so on. The norms of peer review may have amplified the extent to which conventionality in AI is observed as recorded in scientific publications. It would be interesting to implement the same analytical framework on AI invention to other types of records, such as patents and Github repositories, and compare the results.

This study is also limited by the language of choice. WOS only includes publications in English and publications in other languages that have their metadata translated into English. Therefore, WOS contains primarily English publications. So is this dataset, which is made up of 98.4% papers in English, 0.5% in Chinese, 0.3% in Spanish, and 0.8% in others. Nevertheless, China has become a significant player in AI research. Although many Chinese-speaking authors chose to publish their works in English journals (the dataset used in this research shows that authors based in China have published more AI papers than in other countries, and among the top 10 research institutions that have the most AI publications, five are located in China), publications in the Chinese language are likely to be significant in number. Therefore, it will be essential to include AI research in languages other than English in order to achieve a more global and culturally diverse perspective regarding AI research.

2.6 Conclusion

AI's ability to automate intelligent behaviors, find patterns in copious amounts of data, and extract higher-level features from data without direct human supervision has led to high hopes for its potential to revolutionize invention and the search for solutions in many domains and profoundly transform many aspects of society. However,

the question remains whether scientific research related to AI has been advancing in a fundamentally revolutionary fashion. This study presents empirical evidence that, in the main, AI is *not* progressing in radically different ways from the history of scientific inquiry in general and that it is appropriately classified as "normal science."

The evidence offered in this chapter is obtained by implementing a framework that identifies and assesses knowledge recombinations in scientific publications within a dataset containing metadata of almost 300,000 AI research publications. I identified references in each AI publication as "previous knowledge," within which the pairwise combinations are described as recombinations of existing knowledge. I computed a standardized frequency for each year's almost 8 million journal pairs. I extracted two statistical features for each publication to measure its novelty and conventionality. Based on the two features, AI publications are classified into four categories – *Accepted Wisdom*, *Avant Garde*, *Darwin's Tower*, and *Platypus*. I examined categories' composition and temporal evolution and how the categorization relates to a publication's scientific impact.

Like other normal sciences, AI research is highly skewed regarding citations. Citation counts in AI publications resemble a power law distribution. About 20% of AI publications have never been cited. More than half have no more than five citation counts. Furthermore, AI research publications can be considered a manifestation of the Pareto principle frequently demonstrated in science, where the top 20% of AI publications produced approximately 80% of all citation counts.

The result of this chapter shows that among the four taxonomic categories, the category of *Accepted Wisdom* that combines existing knowledge in a highly conventional way without introducing much novelty has become the dominant driving force in AI research. The prevalence of *Accepted Wisdom* is even more significant in the highest-impact publications (so-called "hit" papers) as its share in the top tiers is even

higher than its share in all AI publications. In addition, *Accepted Wisdom* exerts the most extraordinary scientific impact. This study found that *Accepted Wisdom* papers have created more than half of the citation counts in AI publications. A publication in the *Accepted Wisdom* category also tends to be cited more (as confirmed by regression analysis). It has a higher probability of becoming a highly-cited "hit" paper. Nevertheless, *Darwin's Tower*, the category representing both high conventionality and high novelty, is less skewed in citation counts than the other categories. A *Darwin's Tower* paper has a lower probability of never being cited.

In conclusion, the analysis presented in this research implies that AI research has been dominated by intellectual activity that exploits highly conventional combinations of existing knowledge, a manifestation of a "normal science." This type of activity also generated the most scientific impact, being diffused and adopted to the broadest range.

Chapter 3

RECOMBINING AI INVENTIONS: AN EXPLORATORY ANALYSIS ON AI PATENTS' COMBINATORIAL INVENTION

3.1 Introduction

Artificial Intelligence (AI) has been heralded as having the potential to transform how new knowledge and solutions are found in many areas of human activity.[1] Although extensive literature has used AI patents to unveil the development of AI, it is not clear how novel, diverse, and disruptive AI inventions are, let alone how similar or different AI scientific research and AI patenting are. Is the nature of invention in AI different from that which has characterized other scientific and engineering fields? In addition, to what extent have AI patents been relying on scientific knowledge created with funding from the U.S. government? This chapter investigated over 250,000 U.S. patent records related to eight AI subdomains. I found that invention rates related to AI have been increasing faster than invention rates generally. However, incremental improvements that heavily rely on existing knowledge rather than radical or highly novel ideas are increasingly the primary driver of knowledge creation in AI patents. Furthermore, AI patenting is also evidently less diverse or disruptive than inventions in general. Nevertheless, AI patents that combine existing technologies tend to have a higher impact, while those that introduce new technical capacities tend to disrupt the knowledge network to a greater extent. I also find that AI inventions rely heavily on public support, more heavily than utility inventions in general. In addition, the

[1] Part of the results in Chapter 3 will be submitted for publication. Certain parts of the manuscript are based on collaboration with Dr. José Lobo and Dr. Deborah Strumsky.

analysis in this chapter found that team sizes matter to the impact of AI patents — patents by larger teams tend to receive more citations while smaller teams tend to break with the established practice and explore new technological paths. The results presented in this chapter show that the development of AI does not mark a deviation from historical trends of conventional technological and scientific inquiry.

AI methods, techniques, and procedures are already transforming medical diagnostics, transportation planning, manufacturing processes, data analysis, automated-decision-making, speech and optical pattern recognition, software design, and even legal analysis and investment decisions (National Research Council, 1997; Matheny et al., 2019b; World Economic Forum, 2018; Mozer et al., 2019; Correia and Reyes, 2020; Chang, 2020; Chalmers et al., 2021; Surden, 2019; National Academies of Sciences, Engineering, and Medicine, 2021). AI is expected to transform the workplace like the Industrial Revolution, and electrification did (Frank et al., 2019a). Furthermore, deployment and mastery of AI are seen by governments as a cornerstone of national security and economic prosperity (State Council of the People's Republic of China, 2017; Eric and Work, 2021). "Machine-learning" and "deep learning," algorithmic and computational procedures able to learn from and recognize patterns in copious amounts of data, can now surpass human capacities in important problem areas (LeCun et al., 2015; Sejnowski, 2020). AI is thought to have the potential to not only aid the process of invention but to become a method of invention because of its ability to rapidly and effectively search high-dimensional solution spaces of the sort characteristic of many types of problems in science and technology (Agrawal et al., 2018; Cockburn et al., 2019; Crafts, 2021).

AI is both a field of scientific and engineering research and a domain of technological development, all seeking to invent, that is, to generate intellectual novelty. Here, technologies are considered ideas about rearranging matter, energy, and information

(Romer, 2010). An invention consists of ideas or technologies, either new or already in use, brought together in a way not previously observed (Schumpeter, 1934). How is the invention within the domain of AI occurring? What are the important AI technologies, and how are they advancing? Are they making progress by developing new technological functionalities and scientific understandings? Are they mainly refining existing ideas and methods or making breakthroughs? Are they inclined to engage distinctly different technologies or combine closely related fields? To what extent are AI inventions relying on knowledge created by government-funded research? What characteristics are associated with highly impactful, highly disruptive AI inventions?

Chapter 2 investigated AI as a scientific inquiry where new knowledge is recorded in scientific publications and created by authors referencing and acknowledging existing knowledge as recorded in previous literature. It showed that knowledge creation in AI scientific publications is primarily driven by combining existing knowledge in conventional ways, and such publications are also the most impactful. This chapter examines AI knowledge by selecting another format of human knowledge that records primary technical advancement rather than scientific ones — patents.

Patents are exclusive rights for inventions that are new and inventive. Patent databases contain technical information not found in any other source and thus provide early indications of technological developments and can reveal how inventions drive transitions in energy systems. Patents are meant to protect intellectual property and assets that can help attract investment, secure licensing deals, recoup R&D costs and provide market exclusivity. Patenting trends and subject matter thus provide helpful information about the extent and focus of AI invention.

This chapter investigates several patent indicators of the novelty, diversity, and disruption of U.S. patents relating to AI and its eight subdomains (also called AI component technologies) by examining their respective quantifiable metrics. Specif-

ically, the novelty of AI patents is investigated through their percentage of patents that only refine existing technologies instead of introducing new technical components and their percentage of patents that combine closely related technologies (or "narrow combinations") rather than distinctly different ones (or "broad combinations"). AI patents' technological diversity is computed as the normalized Shannon entropy regarding patent-technology-code pairs on the level of CPC subclasses (or 4-digit codes). Disruption of AI patents to subsequent technological paths is measured by the five-year disrupt index (CD_5) developed and pre-computed by Funk and Owen-Smith (2017); Funk et al. (2022). I also investigate how many AI patents are created with scientific knowledge funded by the U.S. government and how many of them are government agencies, universities, private corporations, lone inventors, and foreign inventors. Furthermore, team size and its relationship with the sources of novelty, disruption, and narrow-or-broad combinations are examined. Finally, the characteristics of AI patents and AI scientific publications are compared and insights are offered regarding what the similarities and differences between the two tell us about AI inventions and inventions in general.

3.2 Research Question

I seek to address the following research questions in this chapter.

1. What is the pace of AI patenting?
2. How novel is AI patenting?
3. How disruptive is AI patenting?
4. How is government support fueling AI patenting?

5. How do team sizes and public support affect the impact and novelty of AI patents?

6. What is different and what is similar between scientific publications and patents in AI? Moreover, what are those differences and similarities tell us about AI knowledge and inventions in general?

3.3 Data

3.3.1 AI Patent Dataset

The analysis of this chapter is based on a dataset consisting of 256,892 AI patents granted by the United States Patent and Trademark Office (USPTO) over the years from 1976 through 2020. This AI patent dataset was constructed by selecting AI-related patent numbers from a utility patent dataset. The USPTO currently does not explicitly or officially identify AI-related patents with technological classifications, although the U.S. Patents Classification (USPC) system, which the USPTO previously used to classify patents, has a code for AI. Such codes were used by scholars to compile AI patent datasets. For instance, Fujii and Managi (2018) used AI-related USPC classes to compile a dataset consisting of 13,567 AI patents and found that the priority of AI technologies has shifted from biological and knowledge-based models to mathematical models. However, since 2015, the USPTO has switched from USPC to Cooperative Patent Classification (CPC) system, which has been widely used by major patent offices around the globe, making it challenging to identify AI patents systematically.

Nevertheless, the USPTO released the AI Patent Dataset (AIPD) in 2021 with identifications of eight AI component technologies (also referred to as AI subdomains in this dissertation) for each patent using machine learning algorithms, allowing for

further investigation (Giczy et al., 2021). The AI patent dataset is compiled from two sources — the AIPD for AI patent numbers and the *PatentsView* open data platform for obtaining detailed patent information (Toole et al., 2021). It is worth noting that the AIPD provides each patent with eight scores ranging from 0 to 1, predicting how likely the patent belongs to one of eight AI subdomains, such as natural language processing and speech recognition. In this study, a patent is identified as an AI patent if its highest score is no less than 0.99. I considered AI patents granted by the USPTO a representative sample of global AI inventive activity because the USPTO had received more AI patent applications than any other patent office (WIPO, 2019).

The eight AI subdomains defined by the USPTO include knowledge processing, speech, AI hardware, evolutionary computation, natural language processing, machine learning, vision, and planning and control (Toole et al., 2020). The following brief definitions are excerpted from the *Inventing AI* report by the USPTO (Toole et al., 2020):

- **Knowledge processing:** "The field of knowledge processing involves representing and deriving facts about the world and using this information in automated systems."

- **Speech:** "Speech recognition includes techniques to understand a sequence of words given an acoustic signal."

- **AI hardware:** "Modern AI algorithms require considerable computing power. AI hardware includes physical computer components designed to meet this requirement through increased processing efficiency and/or speed."

- **Evolutionary computation:** "Evolutionary computation contains a set of computational routines using aspects of nature and, specifically, evolution."

- **Natural language processing (NLP):** "Understanding and using data encoded in written language is the domain of natural language processing."
- **Machine learning:** "The field of machine learning contains a broad class of computational models that learn from data."
- **Vision:** "Computer vision extracts and understands information from images and videos."
- **Planning and control:** "Planning and control contains processes to identify, create, and execute activities to achieve specified goals."

For further analysis, multiple supporting datasets that contain detailed information about each patent were downloaded from *PatentsView* open data platform[2], including patent grants dataset, patent-inventor dataset, patent-classification dataset, patent-assignee dataset, and patent reference dataset (Toole et al., 2021). These datasets were downloaded in March 2022. These downloaded patent data were based on the Bulk Download Database Tables on *PatentsView* updated on December 30th, 2021[3] while the CPC classifications rely on the CPC version updated on August 2021.[4]

Those patent datasets were merged based on patent grant numbers. The ultimate AI patent dataset used in this study consists of detailed bibliographic information about 256,892 patents related to AI.

[2] `https://patentsview.org/download/data-download-tables`

[3] See *PatentsView* release notes for details for this data release: `https://patentsview.org/release-notes`.

[4] See the official Notice of Changes on the CPC website for more details: `https://www.cooperativepatentclassification.org/CPCRevisions/NoticeOfChanges`.

3.3.2 Patent Disruptive Index Data

Citation counts as a measure of how many future patents reference previous patents are often considered a useful metric to assess how impactful or successful a patent is. Nevertheless, citation counts alone cannot tell how a patent affects future technological paths by disrupting or consolidating existing knowledge. The "capability destroying index," or *CD index*, a metric for measuring how a patent or a scientific publication affects the network of existing knowledge, has been developed by Funk and Owen-Smith (2017) and used by several studies to capture the impact of a patent or a publication on subsequent knowledge creation (Park et al., 2023; Wu et al., 2019).

The CD index of a patent ranges from -1 to 1. It appraises the extent subsequent patents cite a given or focal patent (forward citations) rather than the previous works the focal patent has cited (backward citations). A patent with a CD index equal to 1 signifies maximal disruption to existing citation networks for subsequent patents that cite the given patent but do not cite the previous works cited by such patent. On the other hand, if the patents that cite a patent as likely as they cite the previous patents cited by the focal patent, regardless of how many citations the focal patent receives, it can be considered having caused little disruption to future technological paths, hence associated with a CD index that equals to -1, indicating consolidating of an existing technological network without disruption. Significant novelty is associated with high disruption. Implementing the CD index on scientific publications and patents, Park et al. (2023) have reported a decline in the general level of disruption in science and technology in the past half-century.

In this chapter, the CD index is applied to measure how disruptive or consolidating patents in AI subdomains are. For this purpose, a dataset containing computed CD

indices for USPTO patents granted from 1976 to 2010 is obtained (Funk et al., 2022). Indices that reflect citation networks five years after a patent is granted (the CD_5 index) are selected. It is worth noting that this dataset only covers patents up to 2010 because computing the index requires the accumulation of citation data after a patent is granted. There is a lot of missing data for the first several years because the citation links for pre-1976 patents are not digitally available. Therefore, for measuring the CD index of AI patents, only patents granted from 1980 to 2010 are considered.

3.3.3 Patents Relying on Government-supported Research

A technical domain that heavily relies on governmental research can arguably be considered still in its infancy or not ready to attract the support and resources from the private sector where financial return is essential. Nevertheless, previous research has reported that patenting in the U.S. is increasingly fueled by scientific knowledge created with support from the U.S. government (Fleming et al., 2019b).

In this chapter, to identify AI patents produced based on research efforts funded by the U.S. government, a dataset that identifies USPTO-granted patents relying on government support (referred to as the *public-reliance patent dataset*) was acquired (Fleming et al., 2019b,a). In the dataset, a patent is identified as relying on research supported by the U.S. government funding if at least one of the following three criteria is met (Fleming et al., 2019b):

1. the patent is owned by the U.S. government, including any U.S. government agency or military institution.

2. the patent explicitly acknowledges Federal government support.

3. the patent cites another patent or a scientific publication that satisfies at least one of the two criteria above or is authored by an individual affiliated with a

government agency.

The public-reliance patent dataset also contains information about the type and country of origin of each patent's assignee(s). An assignee of a patent is the entity that owns the property right of the patent. According to U.S. patent law, a patent could be assigned to one or more government agencies, universities, private companies, other organizations, or individuals. If a patent is not assigned upon grant, the patent is owned by the inventor(s). In the public-reliance patent dataset, for patents authored by U.S. inventors, four types of assignees are identified — U.S. government agencies, universities, corporations, and lone inventors. The U.S. military academies are treated as governmental agencies. For patents authored by inventors outside the U.S., the dataset identifies whether the patents are owned by entities located in the following eight countries or regions — Germany, UK, India, Taiwan, France, Japan, Korea, and China (Fleming et al., 2019b). This dataset only covers patents up through 2017. Therefore, subsequent analysis of AI patents' reliance on government-supported science is up through 2017.

3.4 Method

3.4.1 Source of Technological Novelty

Patents are intended to describe novelty — the invention or discovery of "a new and useful process, machines, article of manufacture, or composition of matter, or any new and useful improvement thereof; a new, original, and ornamental design for an article of manufacture; and any distinct and new variety of plant" in the language of U.S. Patent Law (USPTO). The technological components responsible for a patent's novelty are classified using a technology-code system that identifies distinct technological functionalities (USPTO).

Major patent offices, including the USPTO, have adopted the Cooperative Patent Classification (CPC) system to classify the technologies responsible for the novelty of patented inventions. The CPC system has a hierarchical structure that describes three technical information levels: sections, classes,sub-classes, and groups. A class generally delineates one technology from another, while subclasses delineate processes, structural features, and functional features of the subject matter encompassed within the scope of a class. For instance, the CPC code "G06N 20/10" describes a machine learning technique that uses kernel methods, such as support vector machines (SVM). The first letter "G" represents the section of "mechanical engineering," which, together with the following "06N" (class and subclass) represents "computing arrangements based on specific computational models." The remaining symbol ("20/10") represents the "group" that specifies the most detailed level of technical information. In this case, group "20/00" of subclass "G06N" describes machine learning techniques, while its subgroup "20/10" signifies a machine learning technique characterized by kernel methods. Under the CPC scheme, every patent is assigned at least one code to specify the non-trivial technical component(s) disclosed in the patent (USPTO, 2015b).

To identify sources of technological novelty, a method that provides a taxonomy consisting of four categories — namely *Technological Novelty Taxonomy* is utilized. Such a method was introduced to classify the novelty of patents (Strumsky and Lobo, 2015a). Multiple codes are often combined by patent examiners in order to describe the technologies responsible for a patented invention's novelty. The set of n technology codes associated with a patent is referred to as the *n-tuple* of the patents. The Technological Novelty Taxonomy focuses on the pairings (binary combinations) of capabilities generated from an n-tuple and uses these pairings to assess patented inventions' technological novelty.

Each AI patent is categorized into one of the following four types:

1. *Origination*: all the technology codes used to classify a patent are new. In other words, none of the technology codes have been used previously in the patent record.

2. *Novel Combination*: the technology codes of the patent form new binary combinations with at least one new code. In other words, new code(s) and old code(s) are used in classifying a patent.

3. *Combination*: the technology codes of the patent consist of new pairwise combinations, although none of the codes are new.

4. *Refinement*: none of the codes nor pairwise combinations of codes are new.

Although a patent signifies the arrival of a new invention, the invention might constitute an improvement or refinement of existing technological capabilities. The percentage of refinement patents (referred to as *refinement rate*) in a technological domain indicates novelty in its inventive efforts. The more patents in a domain constitute refinements of existing inventions without introducing new technological capacities, the fewer breakthroughs can be assumed to occur, and the lower the level of novelty is in the domain.

Previous research has reported that, although increasingly more technical components have been introduced, refinement patents have become prevalent in patenting since the 1990s, accounting for 80% of all new patents in the 2010s (Strumsky and Lobo, 2015b; Lobo and Strumsky, 2019). In this chapter, in addition to categorizing each AI patent into the four types defined by the *Technological Novelty Taxonomy*, the percentage of refinement is computed for AI patents and patents in the eight AI subdomains. Comparing refinement rates in different domains and their time series

can inform us about the primary source of the novelty of the patents in each sector and whether they are becoming more or less novel over time. In order to compare the refinement rates of AI patents to inventions in general, the annual refinement rate for utility patents in general is also computed.

Two constraints are notable when deciding whether a technology code or a code pair is new. First, if a code or a pair in a year t has never existed in previous years, it is considered new throughout the year t. In other words, if two patents in the year t involve the same code that has never appeared before the year t, both patents, regardless of which one is granted first, are considered to have a new code. Secondly, whether a code or a pair is new is evaluated within domains. Therefore, a new code to one domain may already exist in another. This chapter considers nine domains — AI patents in general and the eight AI subdomains. (When computing the refinement rate of utility patents, utility patents are considered belonging to one domain.) This consideration is based on the convention of patent examination practice, which assesses whether an invention is new based on the perspective of "a person skilled in the art" in the context of the technical field the invention belongs to (U.S.Code, 2011). Because the temporal scope of the data is from 1976 to 2020, a technology code or a combination is considered new in a given year only if it has not been recorded in the patent record since 1976. Accordingly, all technology codes and code pairs in 1976 are considered new, resulting in a zero refinement rate in 1976.

3.4.2 Technological Diversity

The concept of *entropy* can be used to describe the informational surprise of a system (Shannon, 1948), along with others (such as the Herfindahl–Hirschman index and Inverse Simpson index), and has been used to measure the diversity of ecosystems and industries (Jacquemin and Berry, 1979; Fang and Lou, 2019; Gemba and

Kodama, 2001). The higher the entropy is, the more diverse or surprising a system's configuration is. An entropy of 0 indicates that the system has only one entity, thus has no surprise. Entropy has been used by researchers to measure the technological diversification of regions and companies as well as organizational concentration (Mewes and Broekel, 2022; Rocchetta et al., 2022; Rocchetta and Mina, 2019; Triulzi et al., 2020).

Normalized Shannon Entropy is a variation of Shannon Entropy that considers the maximum possible number of entities in the system with the benefit of obtaining a unified range of. In this chapter, normalized Shannon Entropy is computed to measure the technological diversity of patents in a field. The normalized technological entropy (θ) of a domain d can be computed using Equation 3.1:

$$\theta_d = \frac{-\sum_{c=1}^{C_d} p_c \cdot \ln(p_c)}{\ln(C_d)} \tag{3.1}$$

where C_d denotes the number of distinct CPC subclasses observed in the classifications of patents in domain d, while p_c denotes the percentage of distinct patent-subclasses pairs that are classified with subclass c. CPC subclasses are referred to as the four-digit CPC codes, for instance, "A03D." I select subclasses as the level to measure the technological diversity because different subclasses represent technologies that are distant enough to be considered to belong to different domains. There are 674 subclasses in the current CPC scheme (version 2022.05).

3.4.3 Broad or Narrow Inventions

Patents' novelty can stem from how technological functionalities are combined and how novel these combinations are. Another dimension of inventive novelty can be found in how similar or dissimilar ("distant") the combined technologies are. A more distant combination can be presumed to require more creative thinking to conceive

than one consisting of closely-related technologies, hence is more likely to be novel.

Youn et al. (2015) have categorized technological combinations into "broad" ones and "narrow" ones. Technological combinations are classified as distant if they consist of two technologies classified within the same class of the United States Patent Classification (USPC) scheme (used to classify patents before adopting the CPC schema). Classes are broad categories of technological functionalities (such as excavating or microprocessors). Using this categorization, they reported a boost in broad inventions in the U.S. following World War II but an increase in narrow inventions since around 1970.

In the CPC scheme, subclasses (also referred to as four-digit CPC codes) can be considered the equivalent of USPC classes (and the total number of CPC subclasses (674) is similar to that of USPC classes (474)). Accordingly, a pair-wise combination is considered broad if the two components are classified into two different CPC subclasses. A multi-coded patent is considered a broad invention if it has at least one broad combination. Otherwise, if a multi-coded patent's codes are all classified into the same CPC subclass, it is considered a narrow invention. Subsequently, the proportion of multi-coded patents in each AI subdomain that can be considered "narrow" (referred to as *narrow invention rate*) are computed as another indication of the technological novelty of the domain.

It is worth noting that each U.S. patent may have two types of CPC classifications — mandatory inventive classifications that represent what is novel in the invention compared to state of the art and optional classifications (referred to as *other additional information*) that represent non-trivial technical information that has already existed in previous technology (USPTO, 2015b). When applying the *Technological Novelty Taxonomy* and computing entropy, refinement rates, and technological diversity, inventive and additional classifications are not distinguished because they

both represent non-trivial technical components that are combined in the inventions. However, when computing narrow invention rates, only inventive classifications are included to tease out the combinations that appear to be broad but have been considered commonplace or obvious to make for a person of ordinary skill in the art based on the prior art.

3.4.4 Assessing Patents' Impact

As discussed in Section 2.3.4, citation counts and annual citations have often been considered reliable indicators for scientific quality and impact, especially at the highly-cited end of the distribution (Phelan, 1999). Recent studies also suggest citation percentiles as a helpful indicator to assess scientific impact (Bornmann, 2020). This chapter adopts such a line of reasoning and utilizes citation percentiles to assess a patent's impact. To address the potential bias brought by the cumulative effect of citations, each patent's citation percentile in its cohort of the same domain and the same year, referred to as the "citation percentile of the year" is computed. Furthermore, Alcácer et al. (2009) have reported different patterns between patents cited by patent examiners and applicants. The former focuses more on the competition between patents, and the latter on the previous technologies a patent is built upon (Wu et al., 2019). Therefore, in this chapter, only citations made by applicants are considered in assessing a patent's impact.

3.5 Results

3.5.1 AI Subdomains and Their Networks

The two most productive AI subdomains are knowledge processing and planning control, while evolutionary computation is the most barren. A significant technologi-

cal heterogeneity is observed in AI patents — 42% of AI patents involve multiple AI component technologies. The most connected pair exists between Knowledge processing and planning and control. Evolutionary computation and machine learning are the two most connected AI subdomains, with 98% and 84% patents involving other AI components, respectively. Two clustering communities can be detected among the eight AI subdomains — one is featured by the processing of verbal or written language, while the other focuses on representing knowledge.

The key variables computed for AI patents in general and patents in the eight AI subdomains are listed in Table 3.1. Knowledge processing, the AI subdomain that seeks to represent and derive facts about the world and use this information in automated systems, has the most significant number of patents, accounting for 61% of all AI patents. It is followed by the domain of planning and control, which "processes to identify, create, and execute activities to achieve specified goals" (Toole et al., 2020). Machine learning and computer vision have almost the same number of patents (around 41,000, accounting for 16% of all AI patents). The least fruitful AI subdomain is evolutionary computation (only 237 patents), which "contains a set of computational routines using aspects of nature and, specifically, evolution."

The eight AI component technologies can co-exist in the same patent. In such cases, the patent engages more than one AI technology. As a result, the sum of the patent count of AI subdomains presented in the second column in Table 3.1 is larger than the number of AI patents (257,214). The co-existence of AI component technologies is not uncommon. A total of 108,817 (42%) patents in the AI patent dataset are associated with more than one AI component technology. There are even 320 patents that involve seven out of the eight AI component technologies. For example, patent No. 10,847,138 granted in 2020 discloses a set of systems and methods for "deep learning internal state index-based search and classification" (Ward

Table 3.1: Key Variables of Eight AI Subdomains

Domain	Patent count	% Refinement	Normed entropy	Narrow invention	Mean CD index	Gov funded rate	Mean team size	Mean citation	Median citation
All AI	257,214	50%	0.52	44%	0.06	19%	2.97	14.59	1
Evolutionary computation	237	42%	0.60	44%	0.05	41%	2.51	20.78	7
AI hardware	29,106	41%	0.49	48%	0.06	28%	2.99	36.62	5
Knowledge processing	156,617	50%	0.51	46%	0.06	20%	2.98	11.78	1
Machine learning	41,079	39%	0.58	37%	0.08	26%	2.92	10.14	1
Natural language processing	13,823	52%	0.43	56%	0.05	27%	2.95	30.53	3
Planning	126,754	53%	0.47	46%	0.05	17%	3.10	18.57	1
Speech	18,154	51%	0.45	46%	0.05	21%	2.79	25.98	2
Vision	41,269	26%	0.52	29%	0.06	18%	2.99	19.81	0

et al., 2020). Seven of the eight prediction scores associated with this patent are higher than 0.99. This patent combines techniques from all the AI subdomains except evolutionary computation.

Therefore, the eight AI subdomains can form an "AI component technology network" where each node represents each subdomain, and each edge represents the co-existence relationship between two subdomains. The topology and characteristics of such a network can inform how different AI technologies are combined. Furthermore, the community structure of such a network can illuminate which AI components tend to form stronger interconnections and cluster together. Figure 3.1 presents such a network between eight AI subdomains in creating new AI inventions. In Figure 3.1, the weight of each edge represents the number of patents that combine the two nodes.

Using a modularity-optimization-based community-detection method developed by Blondel et al. (2008) and Lambiotte et al. (2009), two communities are identified

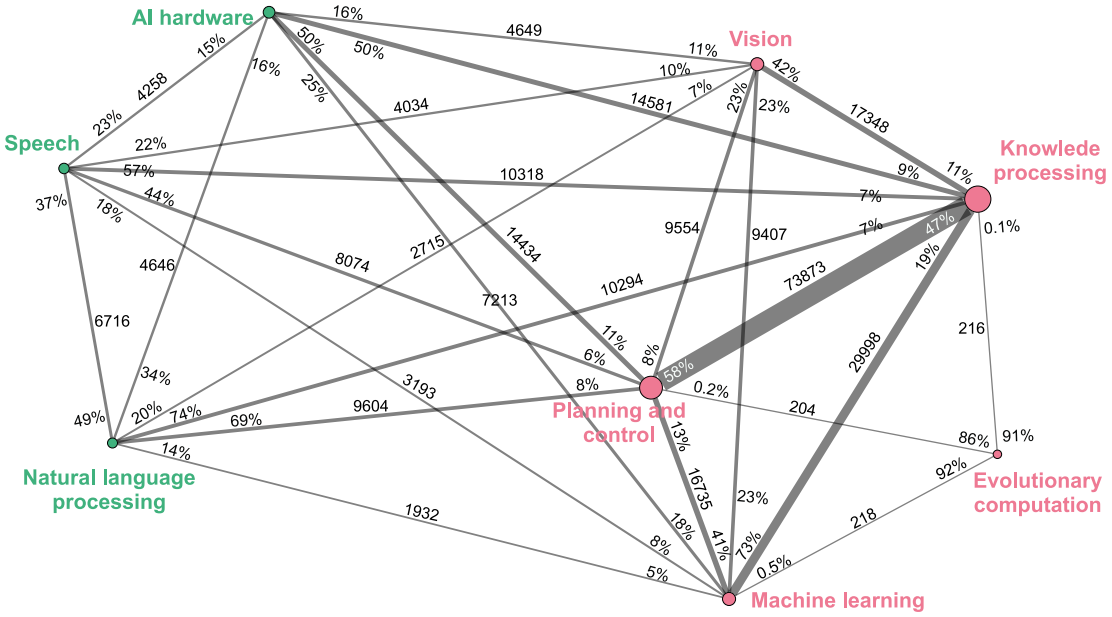


Figure 3.1: AI component technology network between pairs of AI subdomains. Each node represents one of the eight AI subdomains or AI component technology. The weight of the edge between two nodes measures the number of patents with both components. In this plot, only edges with more than 100 patents are shown. The size of each node represents the number of patents observed in the domain. The color of each node indicates its community, where nodes are strongly interconnected. Each edge is presented with three numbers — the number of patents and the percentage of patents in the two subdomains represented as the two nodes connected by the edge.

with a built-in functionality in *Gephi*, a network analysis software (Bastian et al., 2009). The two communities are colored in green and red, respectively, in Figure 3.1. The green community comprises AI hardware, speech, and natural language processing. The red community includes the rest five subdomains — vision, knowledge processing, planning and control, machine learning, and evolutionary computation. The green community features the processing of language, while the red community focuses more on representing knowledge.

As illustrated in Figure 3.1, the strongest link exists between knowledge processing and planning and control (73,873 patents). Knowledge processing and machine learning are also heavily connected (nearly 30,000 patents). The least-weighted edges are the pair between NLP and evolutionary computation and the pair between speech

and evolutionary computation, which have only one and five patents each. (These two edges are not shown in Figure 3.1, where only edges with weights higher than 100 are shown.) Speech recognition and NLP are both domains that address issues related to human language, either verbal or written. Such weak links indicate that evolutionary computation has not substantially helped process human language.

On the other hand, a large proportion of patents related to evolutionary computation are engaging with machine learning (92%), knowledge processing (91%), or planning and control (86%), indicating that although barren in helping process language, evolutionary computation is primarily successful in domains that seek to represent and process human knowledge and experience to achieve the automation of reasoning and planning. Further, there is a significant overlap between the three subdomains related to evolutionary computation. In other words, a large proportion of patents related to evolutionary computation combine machine learning, knowledge processing, and planning and control at the same time.

Regarding to what degree they are connected with or isolated from other domains, AI subdomains differ significantly. However, regardless of how isolated, at least half of the patents in each subdomain involve technologies from other subdomains. Patents related to evolutionary computation are the most connected — 98% of patents involve AI components other than evolutionary computation. This is followed by machine learning, of which 84% of patents result from combining machine learning with other AI techniques. Patents related to computer vision are the most isolated, yet still 49% of percent combine with other AI components.

3.5.2 The Pace of AI Patenting

AI patents have grown exponentially since the 1970s at a significantly faster rate than inventions in general. Among AI subdomains, knowledge processing and plan-

ning control have the fastest growth rate, while patents in evolutionary computation have almost stagnated from 1990 to 2020. Other AI subdomains are growing slightly slower than knowledge processing from the 2000s. Nevertheless, patents related to computer vision increased sharply in the late 2010s. AI inventions' knowledge components and their combinations are also growing. As AI's knowledge stock increases, the search space for possible combinations expands. Nevertheless, the possible search space's growth rate has slowed, but it is still faster than the growth of the searched proportion.

AI Patents Grow Faster than Inventions in General

Figure 3.2 illustrates the time series of the number of patents in AI and its eight subdomains. Specifically, the number of AI patents as compared to AI publications, scientific publications in general, and utility patents is shown in Figure 3.2a, and the number of patents in each AI subdomains as compared to AI patents and utility patents is shown in Figure 3.2b. Numbers in Fig. 3.2a are normalized using Equation 3.2:

$$y_t = \ln\Big(\frac{x_t}{x_0}\Big), \tag{3.2}$$

where x_t denotes the annual count in year t, and x_0 denotes the count of the initial year of each time series.

The annual number of new AI publications and granted patents are both verified as growing exponentially at similar rates (coefficients 0.17 and 0.16 for AI publications and patents, respectively, see Appendices F and G for regression details), as shown in the blue line and orange line in Fig. 3.2a.

AI publications (blue line) exhibit a substantially steeper slope than scientific publications (green line, coefficient 0.05, see Appendix H for regression details). Sim-

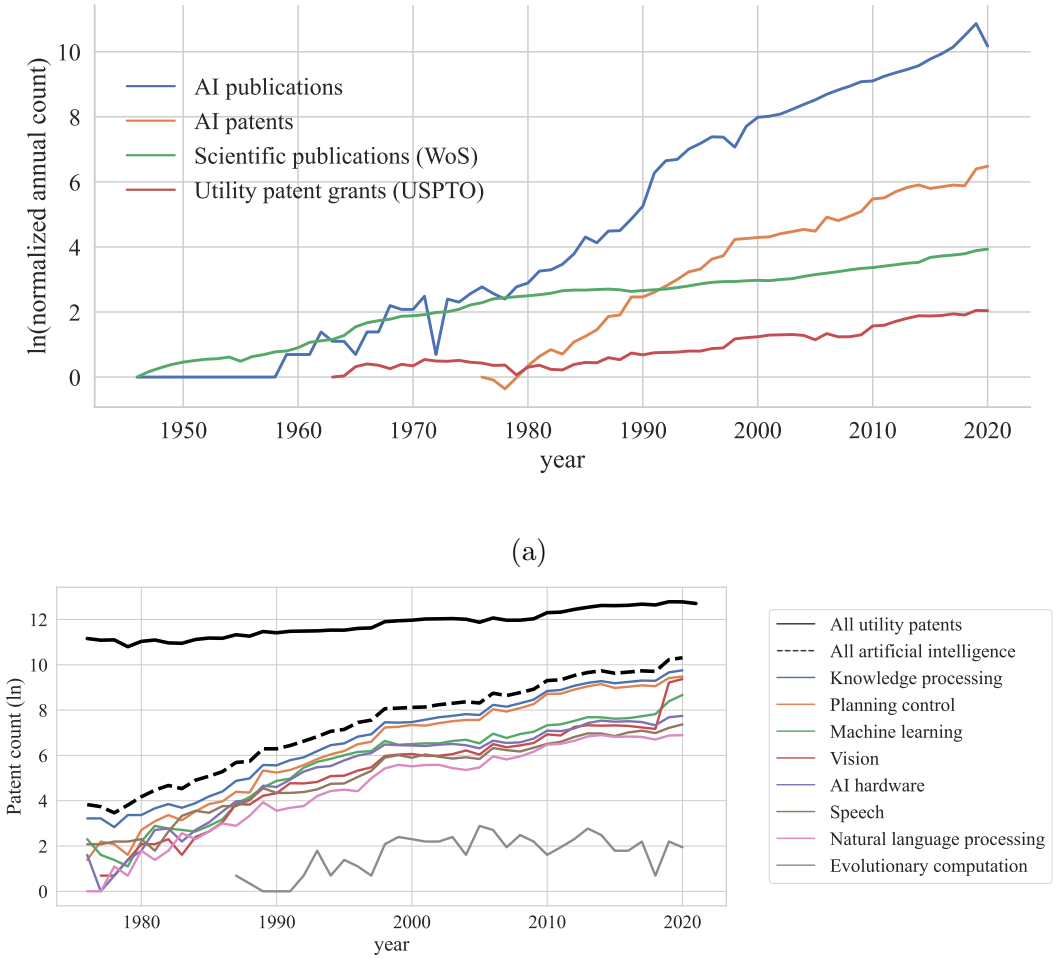


Figure 3.2: Time series of the number of AI inventions. Panel (a) presents the normalized natural logarithm of the number of AI publications (blue), AI patents (orange), scientific publications in general recorded in WOS (green), and utility patents granted by the USPTO (red). Panel (b) illustrates the time series of the natural logarithm of the number of utility patents in general (solid black line), AI patents in general (dashed black line), and the eight AI subdomains.

ilarly, AI patents (orange line) are characterized by a much steeper slope than utility patents in general (red line, coefficient 0.03, see Appendix I for regression details), implying that inventive activity in AI is proceeding at a significantly faster pace than inventive activity overall.

The Pace of Inventions in AI Subdomains

Figure 3.2b shows the time series of the number of patents in each AI subdomain as compared to AI patents in general (dashed black line) and utility patents (solid black line).

Figure 3.2b shows three clusters with different growing trends. First, new patents in knowledge processing (blue) and planning and control (orange) have invariably been more than other subdomains. The second cluster (machine learning, vision, AI hardware, speech, and NLP) featured rapid growth in the 20th century, followed by a stagnation in the 2000s that differentiated it from the first cluster. At the end of the 2000s, AI subdomains in the second cluster gradually resume momentum. Specifically, patents related to computer vision (red line), after a temporary decline in the mid-2010s, have sharply increased to the same ballpark as the first cluster in the late 2010s. Evolutionary computation (gray line) alone can be seen as the third cluster of which new patents did not grow over time. It has fluctuated or even declined throughout the years.

The Growing Knowledge Stock of AI

Although the technological components in each AI subdomain have been growing, the technical fields involved in each subdomain tend to remain stable, indicating that the growth of AI invention is driven by further segmentation of existing technologies more than expanding to new technical fields.

Figure 3.3 shows the time series of the number of CPC subclasses that are often considered technical fields (Panel A) and the number of CPC subgroups that are commonly considered technical components (Panel B). Although the technological components expanded substantially for almost every AI subdomain (see Panel B), the

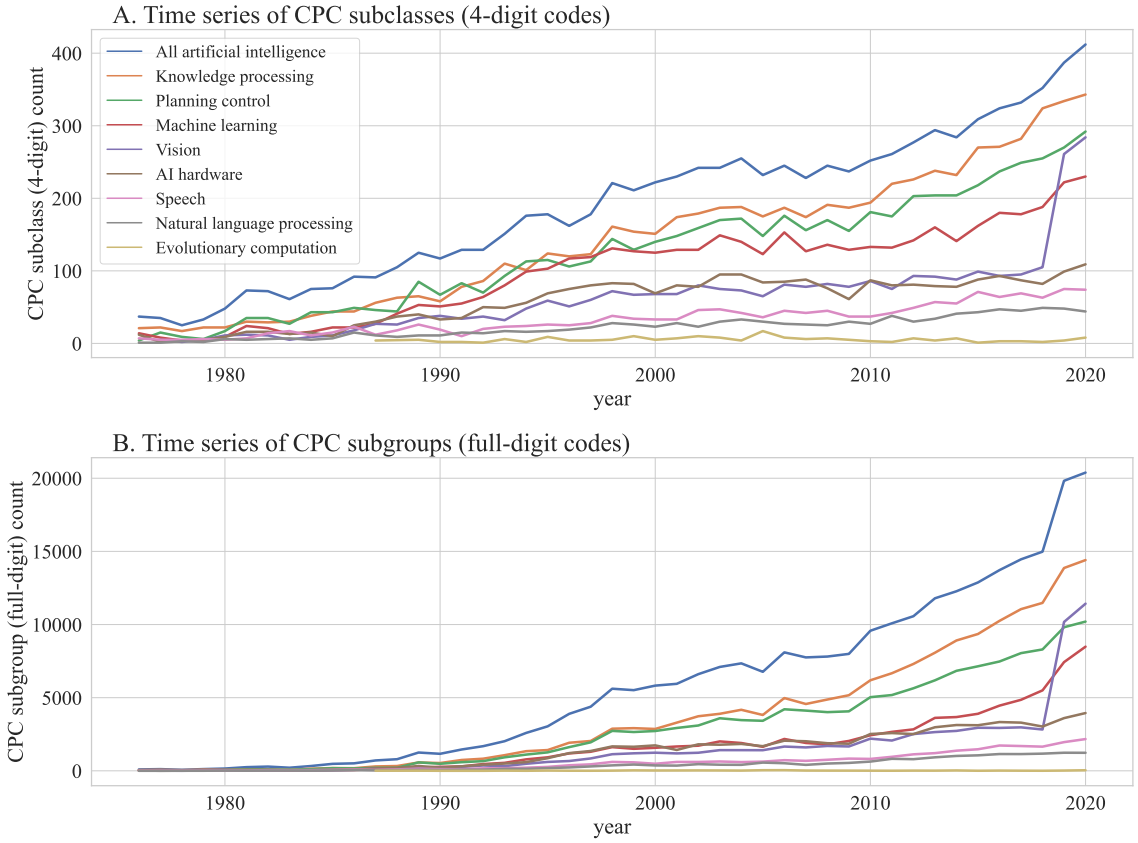


Figure 3.3: Time series of the number of CPC subclasses (also known as 4-digit codes, shown in Panel (A) and CPC subgroups (also known as full-digit codes, shown in Panel (B) in each AI subdomain.

technical fields AI inventions engage remain relatively stable (Panel A). The number of technical fields related to patents of knowledge processing, planning and control, and machine learning has been growing slightly. At the same time, AI hardware, speech, and NLP remain almost unchanged between the mid-1990s and 2020. Technical fields related to computer vision had been stable but increased drastically in the late 2010s, from 95 CPC subclasses in 2017 to almost 300 in 2020. Evolutionary computation involves very few technical fields — at most eight CPC subclasses each year.

3.5.3 How Novel is AI Patenting?

I investigated AI and its eight subdomains for three features that cover three different aspects of novelty — technological diversity, refinement rate, and narrow invention rate — and compared them with utility patents in general. Contrary to the expectation that AI is presumably and inherently novel, this chapter finds that AI patents are less novel than utility patents in general throughout the years but have become more novel than utility patents around 2019. Like utility patents, AI patents feature two time periods with different trends. From 1976 to 2012, AI patents are becoming decreasingly novel. Nevertheless, after 2012, the novelty of AI patents increased. By around 2019, AI patents have become more novel than utility patents. Among the eight AI subdomains, computer vision and machine learning are the two areas with the highest novelty, while NLP is the lowest. However, AI patents' diversity declined drastically from 1976 to 2020, and AI patents are increasingly concentrated in a few technical fields.

AI Patents are Primarily Refinements

AI patents' technological novelty primarily comes from refining existing knowledge (Refinement patents) and combining existing knowledge in new ways (Combination patents), indicating a low novelty. AI patents' refinement rate was higher than utility patents, but the former has been converging with the latter, if not becoming lower. Inventions that introduce unprecedentedly new technical components have been increasingly rare in any domain of AI. Among the eight AI subdomains, computer vision and machine learning patents have the highest proportion of novel combinations and the lowest rate of refinements, indicating a higher novelty.

Figure 3.4 illustrates the composition of the four sources of the technological

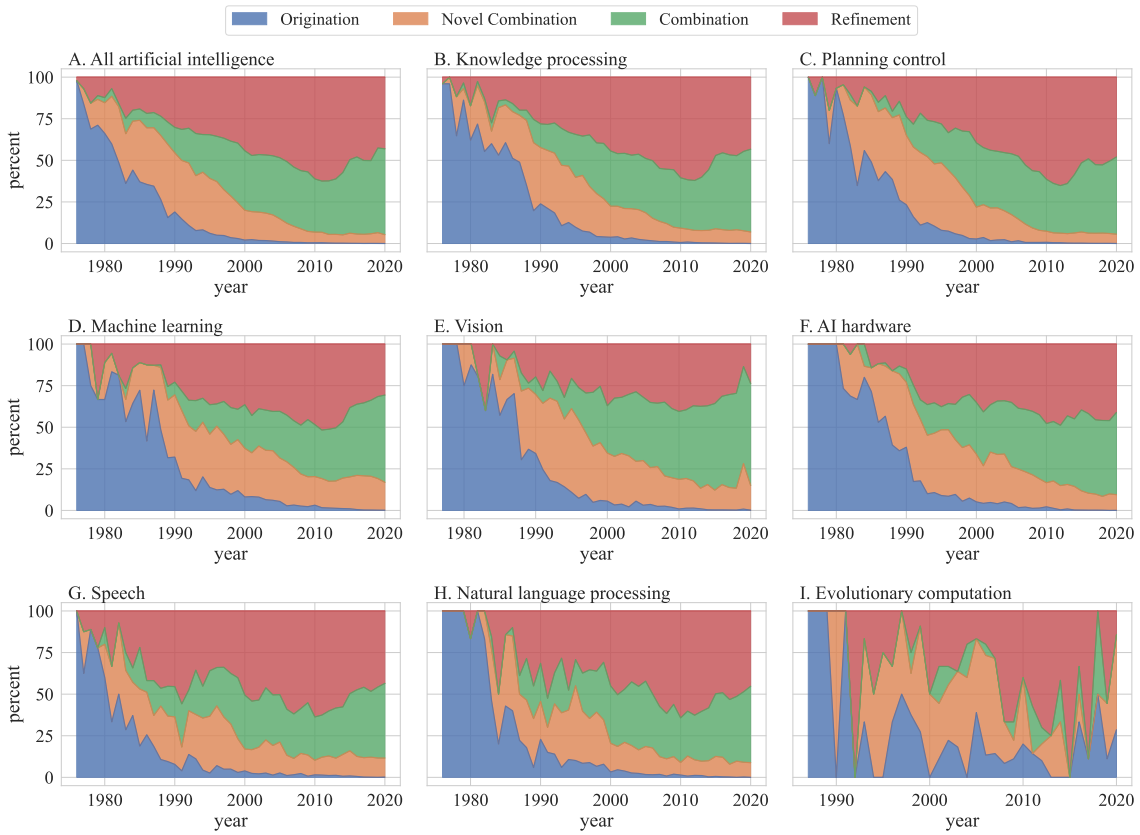


Figure 3.4: Time series of the sources of the technological novelty of AI patents and its subdomains. The four categories — origination, novel combination, combination, and refinement — are defined by Strumsky and Lobo (2015a) and described in Section 3.4.2. Each subdomain is computed separately.

novelty of patents in AI and its eight subdomains. The sources of novelty are categorized into four types — origination, novel combination, combination, and refinement, defined by Strumsky and Lobo (2015a) and described in Section 3.4.2. One trait observable across panels in Figure 3.4 is that for AI patents and patents in most AI subdomains, origination patents (blue) and novel combinations (orange) have decreased substantially since 1976, while combinations (green) and refinements (red) increased significantly since 1976. The percentage of refinements peaked in the early 2010s. After that, it has been declining. Originations almost disappear entirely from AI patents. Novel combinations account for a small proportion of all AI patents (5%). Patents related to machine learning, computer vision, and speech have a rel-

atively higher rate of novel combinations (17%, 15%, and 11% in 2020). Patents of combinations and patents of refinements are about the same proportion for most AI subdomains. However, combinations have been growing faster since the early 2010s while refinements are declining in percentage.

As discussed, the refinement rate can be considered an indication of the novelty of a field. The higher the refinement rate, the less likely that there are breakthrough inventions. In Figure 3.5, the solid red line in each panel represents the time series of refinement rate in each AI subdomain, while the dashed red line represents the refinement rate of utility patents in general as a comparison. Of the 7.2 million utility patents granted from 1976 to 2021, 38% are refinement patents, and this refinement rate grew continuously to around 43% in 2010. Since then, the refinement rate remained stable in the 2010s. Compared to utility patents, AI patents' refinement rate had been higher and peaked at 62% in 2012, almost 20% higher than utility patents, indicating a significantly low novelty. Nevertheless, since 2012, AI patents' refinement rate has been declining drastically in a way that converges with utility patents. In 2019 and 2020, the refinement rate of AI patents decreased to slightly lower than utility patents, indicating that AI patents became more novel in the 2010s. Similar trends can be observed in many AI subdomains in the panels in Figure 3.5. Patents related to machine learning (Panel D) have surpassed utility patents in novelty as measured by refinement rate. In contrast, computer vision patents (Panel E) have been almost invariably more novel than utility patents in general, with a refinement rate as low as 24%.

Half of AI Patents are Narrow Inventions

Another dimension of gauging a patent's novelty is whether the patent combines closely related or distinctly different technologies. The latter likely brings in more

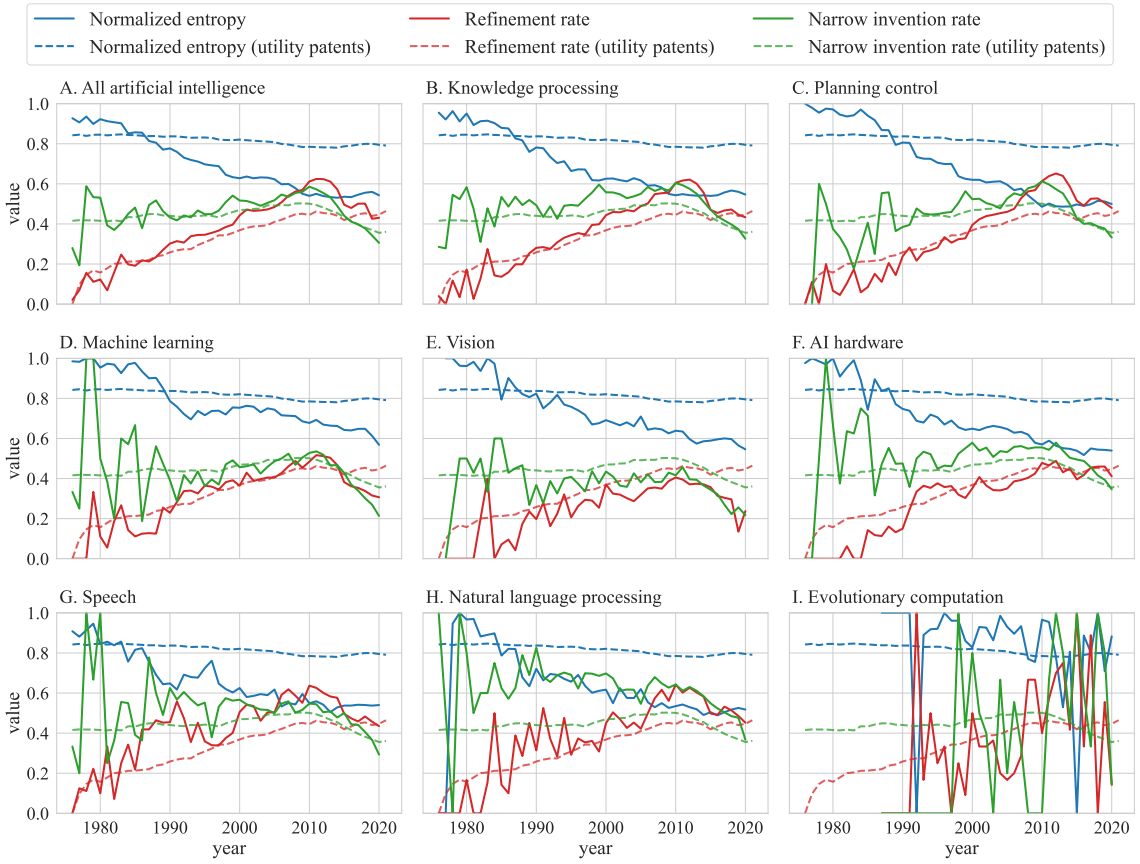


Figure 3.5: Time series of AI patents novelty features. Solid blue lines represent normalized Shannon entropy on the level of CPC subclasses (four-digit code). Dashed blue lines represent the normalized Shannon entropy of utility patents in general as a comparison. Solid green lines represent the narrow invention rate in the domain specified in the panel titles and dashed green lines represent the narrow rate of utility patents in general as a comparison. The solid red line in each panel represents the refinement rate of each domain, and the dashed red line represents the refinement rate of utility patents in general as a comparison.

novelty than the former. Therefore, the narrow invention rate defined by Youn et al. (2015) and described in Section 3.4.3 can be considered another indicator for measuring novelty in addition to the refinement rate.

AI patents share a similar narrow invention rate with utility patents — they grew from around 40% in the 1980s to more than 50% in the early 2010s and since then have declined to less than 40%. Patents related to machine learning and vision have the lowest narrow invention rates (21% and 22%), indicating that they combine distant

technologies to a greater extent than other AI subdomains, and their novelty is high. NLP patents have a higher refinement rate and narrow invention rate (37% in 2020), indicating they are low in novelty.

AI Patents are Becoming Less Diverse

As shown as the blue lines in Figure 3.5, compared to utility patents, where technological diversity tends to remain stable throughout the years and only declined slightly, AI patents' technological diversity is declining drastically from 1976 to 2020. All AI subdomains share a similar trend, except for evolutionary computation, where features fluctuate widely, and a coherent trending pattern is not observed.

I have shown in Figure 3.3 that the number of technological components in AI patents has been expanding significantly. Such an expansion did not translate into technological diversity. The drastic decline in technological diversity indicates that inventions in AI have been increasingly concentrated in a few fields. However, the number of fields engaged in AI is steadily growing, if not remaining at the same level. As shown in Panel A of Figure 3.6, the aggregated percentage of the top ten fields (CPC subclasses or four-digit CPC codes) has been multiplied from around 25% in 1976 to around 70% in 2020, although the number of fields new AI patents each year engage have increased from less the 50 to more than 400 during the same period. Notably, as shown in Panel A of Figure 3.6, technological components in AI inventions have been concentrated in fields such as electric digital data processing (G06F, accounting for 25% of all distinct patent-CPC pairs in AI, shown as blue), data processing systems or methods (G06Q, accounting for 11% of patent-CPC pairs in AI, shown as red), and transmission of digital information, e.g., telegraphic communication (H04L, accounting for 8% of patent-CPC pairs in AI, shown as pink).

Different AI subdomains differ in the technical fields in which they have been

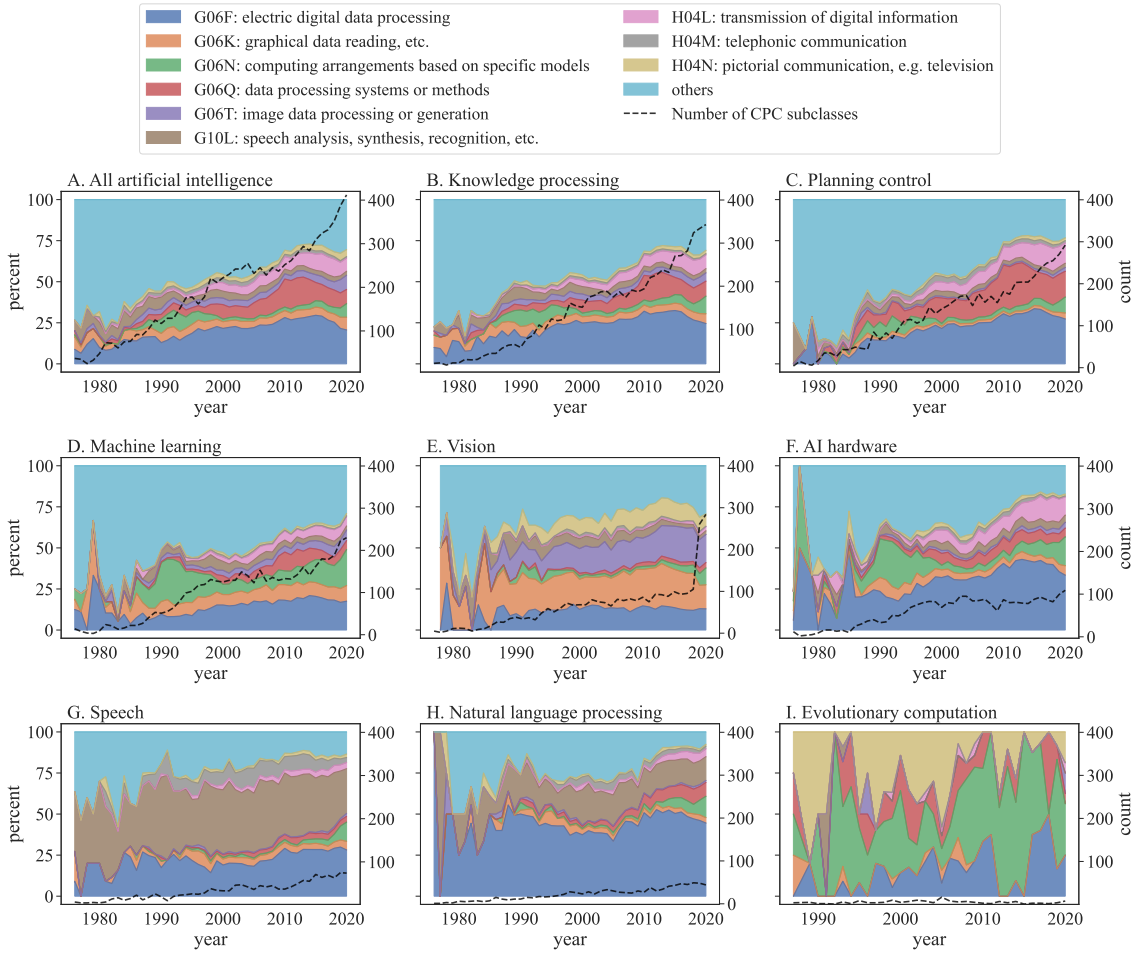


Figure 3.6: Time series of the proportion of top patent-subclass pairs in each AI subdomain. Dashed black lines (right axes) represent the time series of the number of distinct CPC subclasses in each AI subdomain.

concentrated. For example, a large proportion of vision patents (Panel E) involve technologies related to image data processing or generation (G06T, purple, 17% in 2020) or graphical data reading (G06K, orange, 15% in 2020), while patents related to speech and NLP (Panels G and H) primarily focus on "speech analysis, synthesis, and recognition" (G10L, brown, 27% and 14%, respectively) in addition to the field of electric digital data processing (G06F, blue).

Nevertheless, as the number of technical fields grows (dashed black lines in each panel), a higher proportion of patents is concentrated in a few technical fields. Figure

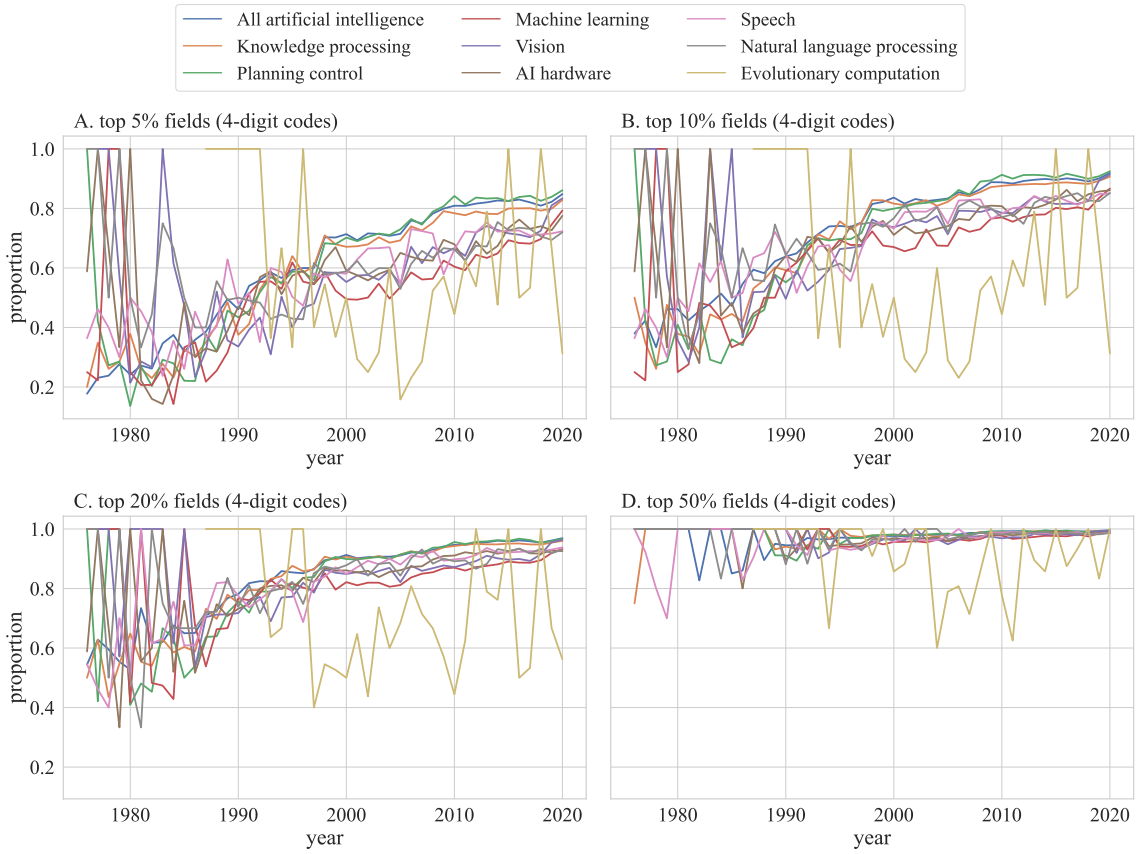


Figure 3.7: Time series of the proportion of patents classified with top 5%, 10%, 25%, and 50% CPC subclasses in each AI subdomain. The generally-observable upward trending pattern indicates that AI patents are increasingly concentrated in a few fields regardless of domains.

3.7 presents the proportion of patent-subclass pairs in each AI subdomain that involve the top 5%, 10%, 25%, and 50% subclasses. As the plots show, the proportion of patents involving top fields in each subdomain is growing. This trend is observed regardless of whether the top 5%, 10%, 25%, or 50% is considered, suggesting that patents are increasingly concentrated in a few top fields. For example, as Panel A of Figure 3.7 shows, 38% of patents related to knowledge processing (orange line) involving the top 5% CPC subclasses in 1990. In 2020, the percentage grew from 38% to 83%. Similarly, in 1990, the top 20% fields in machine learning accounted for 77% of patents (red line in Panel C), whereas in 2020, such a percentage increased to

93%. Such an increase in concentration in the top fields in AI can partly explain the decline of technological diversity in AI.

3.5.4 How Disruptive is AI Patenting?

AI patents, like utility patents in general, are becoming less disruptive to subsequent technological paths. Machine learning patents are the most disruptive among the eight AI subdomains. Among the four types of technological novelty, origination patents tend to be more disruptive, while patents of combinations that combine existing components in new ways receive more citations.

AI Patents are Becoming Less Disruptive

AI patents share a similar disruptiveness with utility patents. Like utility patents, AI patents are becoming increasingly less disruptive to subsequent technology paths from 1980 to 2010. In other words, AI patents are becoming increasingly less likely to break with historical and technological practices and generate new paths. Among AI subdomains, patents related to machine learning are the most disruptive. In addition, compared to utility patents, AI has a higher rate of moderately disruptive patents but a much lower percentage of highly disruptive patents.

As shown in Figure 3.8, despite occasional fluctuations, the disruptive index (CD_5 index) has declined steadily and drastically in each AI subdomain (blue lines). It is consistent with what has been observed in science and technology in general (Park et al., 2023). Like utility patents, the disruptive indices of many domains increased slightly in the late 2000s. Among the eight AI subdomains, patents related to machine learning disrupt future technologies to the greatest extent, demonstrated by the blue line in Panel D that is clearly above the dashed red line and the dashed green line.

Not only are patents becoming less disruptive, but highly disruptive patents also

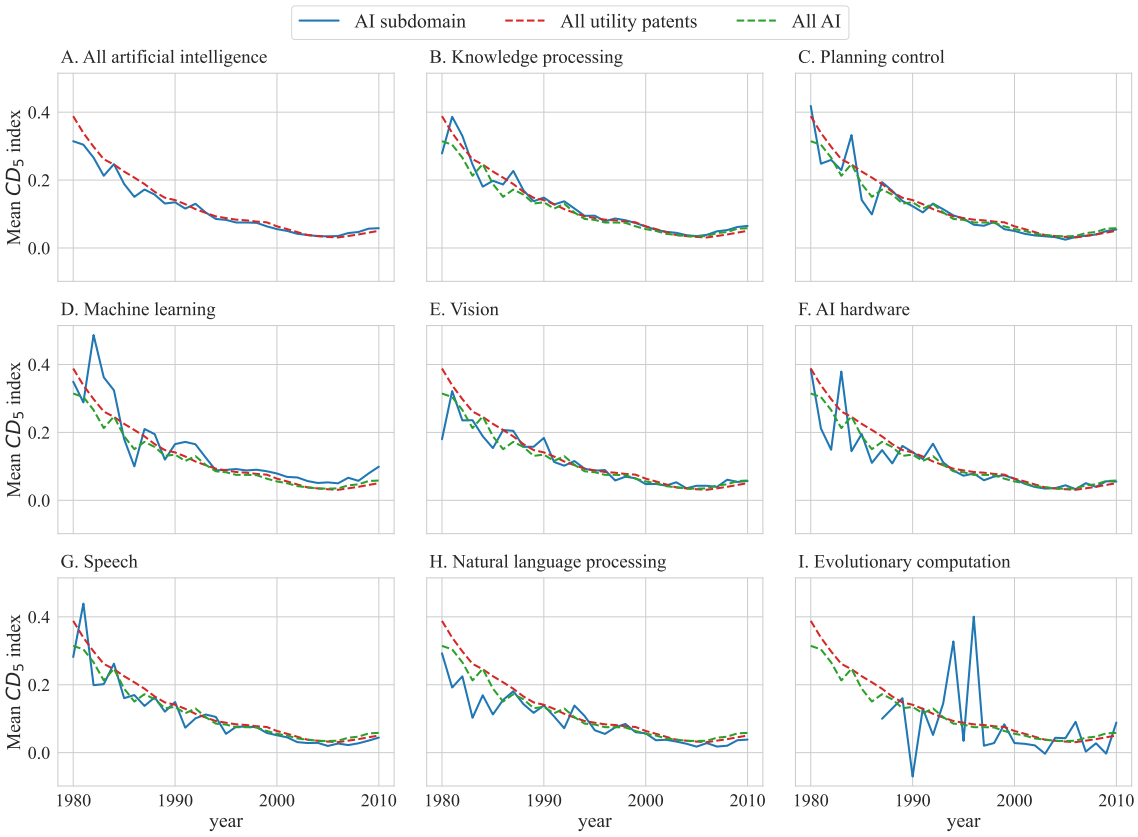


Figure 3.8: Time series of the mean CD_5 index in each AI subdomain. In each panel, the blue line represents the mean CD_5 index of the patents in the AI subdomain, while the dashed green line and red line represent all AI patents and all utility patents in general, respectively.

have significantly declined (Park et al., 2023). As shown in Figure 3.9 where the areas above the dotted black line represent the percentage of highly disruptive utility patents (as defined as patents with $CD_5 > 0.25$), highly disruptive utility patents have decreased significantly, while moderately disruptive patents (as defined as patents with $0 < CD_5 \leq 0.25$) and consolidating patents ($CD_5 \leq 0$) have both increased. A similar trend can be observed for AI patents and each AI subdomain. Nevertheless, the percentage of AI patents that are moderately disruptive is higher than utility patents, while the percentage of highly disruptive patents is invariably lower than utility patents. Only in the case of machine learning did such a highly disruptive percentage converge with utility patents in general in the late 2000s.

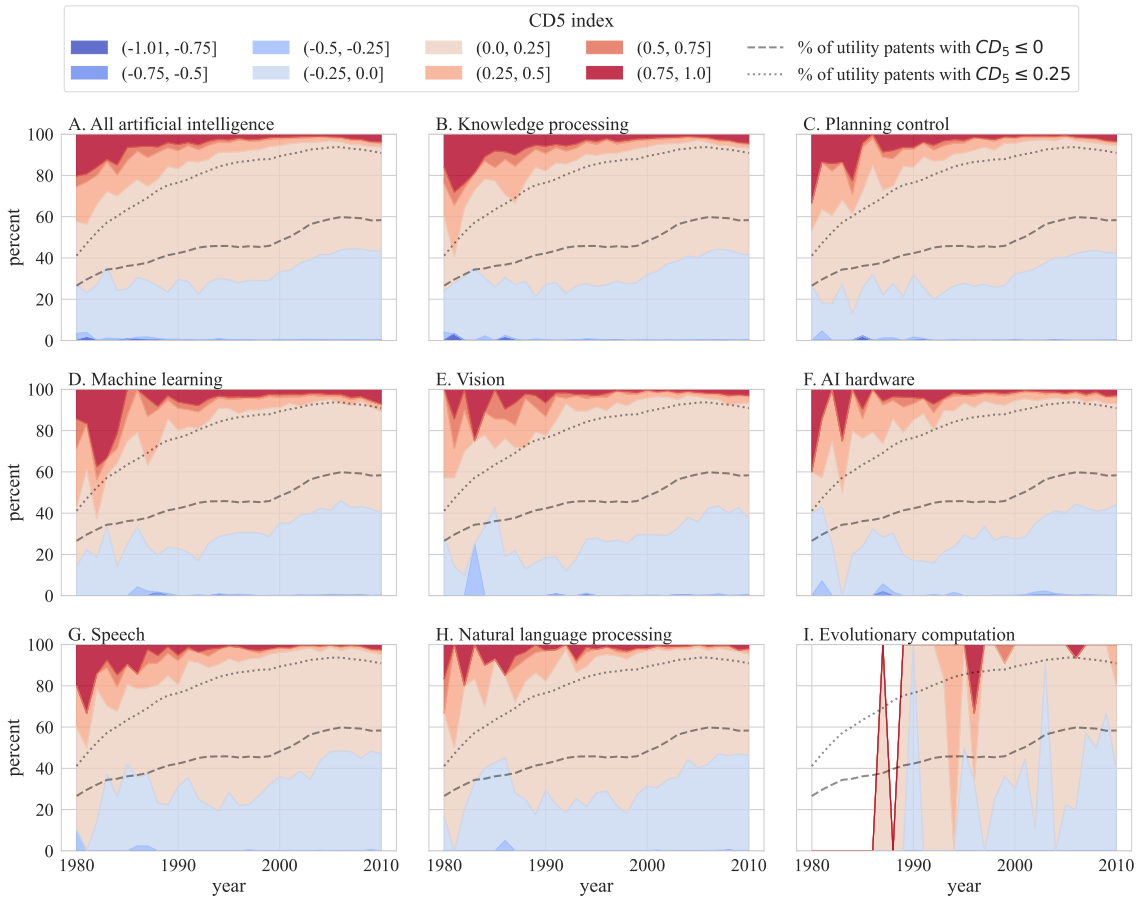


Figure 3.9: The time series of the composition of patents categorized into eight bins according to their CD_5 index in each AI subdomain. The eight bins of the CD_5 index, from maximally consolidating to maximally disruptive, are colored from the deepest blue to the deepest red. The dashed black line and dotted black line in each panel represent the percentage of utility patents with $CD_5 < 0$ and $CD_5 < 0.25$, respectively. One salient trait shared across domains is that highly disruptive patents account for an increasingly smaller proportion over the 1980-2010.

Combinations Have a Higher Impact, While Originations Disrupt More

AI patents that combine existing technologies in new ways (i.e., Combination patents) tend to receive more citations than others granted in the same year. Nevertheless, AI patents that combine new technologies with existing ones (so-called Novel combination patents) have a much higher citation hit rate. In other words, they are more likely to become top-cited patents. On the contrary, origination AI patents that only engage new technologies have a much higher disruptiveness in general and

a substantially higher hit rate of becoming the most disruptive patents.

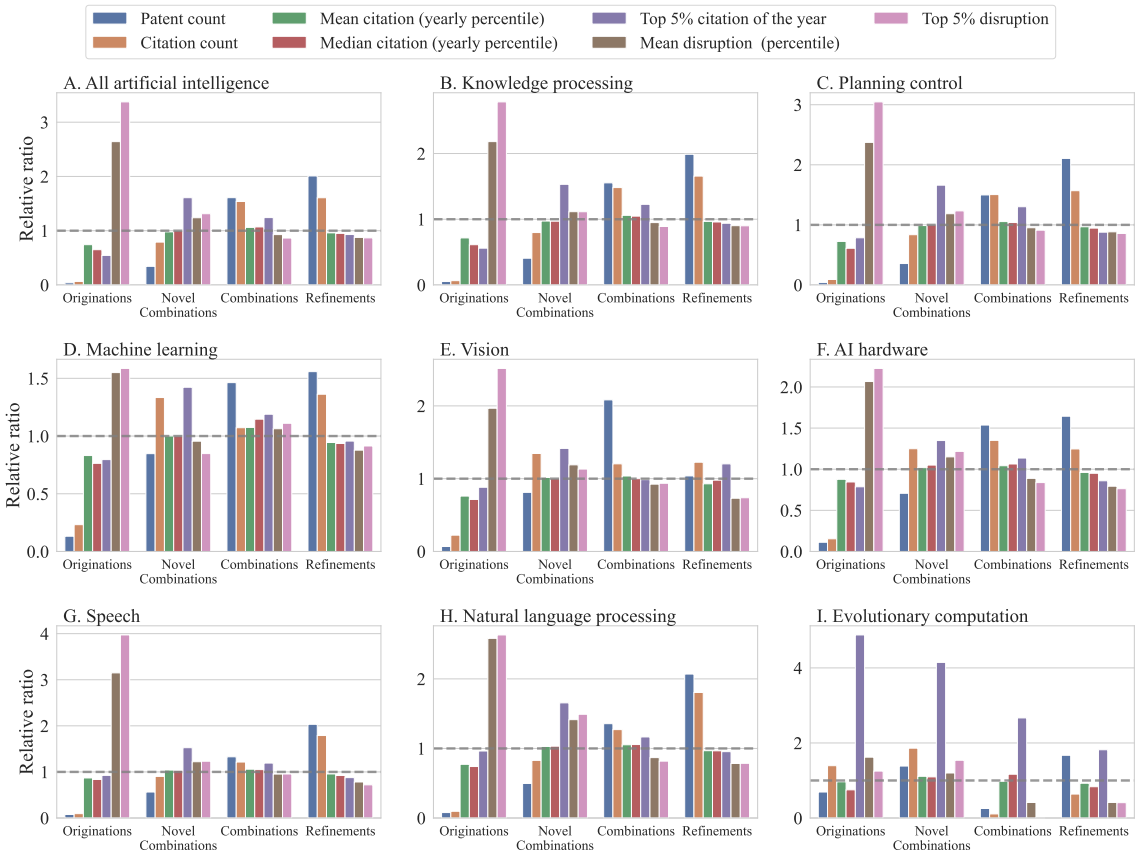


Figure 3.10: Bar plots showing the relative ratios regarding patent count (blue), citation count (orange), mean citation percentile of the year (green), median citation percentile of the year (red), top 5%-cited patents of the year (purple), mean disruption percentile (brown), and top 5% patents in terms of disruption percentile in the four categorized according to the source of their technological novelty in each AI subdomain. Each panel's horizontal dashed black line represents the fictional relative value if the variable is evenly distributed across the four categories.

Many factors may affect whether a patent is likely or not to be referenced by subsequent patents. I find that the different technological novelty sources may relate to citations received by an AI patent and how disruptive it would be in the future. Specifically, AI patents that combine existing technologies in new ways are likely cited more, and AI patents that put forward highly original ideas tend to disrupt subsequent technological paths to a greater extent.

Figure 3.10 presents the relative ratio of seven variables of four categories of

patents in each AI subdomain. A relative ratio of a variable is computed as the value of the variable in question divided by the fictional value in a scenario where the variable is distributed evenly across domains. The horizontal dashed black line at $y = 1$ in each panel indicates such a fictional value. For example, in the case of the variable of patent counts (blue bars), $y = 1$ illustrates a scenario where patents are evenly distributed across the four categories and would share the same amount of patents. A blue bar higher or lower than the dashed black line indicates that the number of patents in the category is more or fewer than one-quarter in the whole domain. If the variable in question is "top 5% hit rate in terms of citation," then the dashed black line represents 5%. Then, a bar that is taller than the dashed black line indicates that more than 5% patents in the category are ranked in the top 5% in citation among all the patents in the domain.

It can be observed that in each panel, patents of combinations and novel combinations tend to have more citations than other types of patents in the same year, demonstrated by the higher green bars and red bars in the second (novel combination) and third groups (combination) in each panel that are taller than those in the first (origination) or the fourth groups (refinement). Combination patents are inventions that introduce new pair-wise combinations, although they do not bring in new components. Novel combinations are featured by combining new components with old components. In short, the two categories either combine new technologies with existing ones or combine existing technologies in new ways. Thus, they are novel enough to deserve acknowledgment and familiar enough to subsequent inventors to be readily built upon — both could lead to reference. It can partly explain the high citations received by combination and novel-combination patents compared to origination or refinement patents. Origination patents have the lowest citations compared to other categories in the same year.

Although patents of combinations and novel combinations are similar in their impact in general, novel combinations clearly have more patents ranked in the top tier in citations. As Figure 3.10 shows, the second group (novel combination) in each panel invariably has the tallest purple bar among the four groups. Purple bars represent the relative ratio in terms of the percentage of patents in the category that are ranked in the top 5% in patents in the domain of the year in terms of citations. Recall in Chapter 2, such a percentage is referred to as the "hit rate." *Therefore, novel combinations, regardless of AI subdomains, have the highest hit rate, indicating a stronger skewness.*

On the contrary, origination patents that engage solely new components often have the lowest citation count among the four types of patents in the same year. It can be partly interpreted that originations are too novel for a person with ordinary knowledge in the field to understand and utilize, let alone improve immediately. Therefore, an origination patent is unlikely to be cited as a previous technology in subsequent patents.

Nevertheless, origination patents are illustrated in Figure 3.10 to cause the most significant disruption to future inventions. Origination patents have the highest mean CD_5 index (brown bars) and the highest hit rate (pink bar). In many AI subdomains, the mean CD_5 index of origination patents is more than twice of other categories. Originations' top 5% hit rate in disruption reaches more than 15% in all AI patents (Panel A) and almost 20% in speech recognition (Panel G). In other words, subsequent patents that cite those origination patents, although they may be small in number, tend not to cite their predecessors. Origination patents' successors are primarily built upon such origination patents. *This indicates that origination patents often break with previous technological practices, disrupt existing citation networks, and lead to new technological lineages.* This could partly be caused by the fact that origination

patents rarely cite previous patents because they are unlikely built upon previous technologies. An alternative explanation can also credit this observation to the fact that there was simply a higher percentage of origination patents in the early years when the general disruption of science and technology was much higher than today. On the other hand, patents of refinement (orange box in the fourth group in each panel) tend to be the least disruptive. In other words, refinement patents, although cited more than origination patents in most AI subdomains, represent incremental advancement and rarely alter technological practices. Their presence tends not to disrupt the citation network and knowledge flow.

Interestingly, such patterns described here are shared across AI subdomains. It may indicate that such observations may be universal in inventions across domains.

3.5.5 AI Patents' Reliance on Government-funded Science

AI Patents Increasingly Rely on Public-supported Science

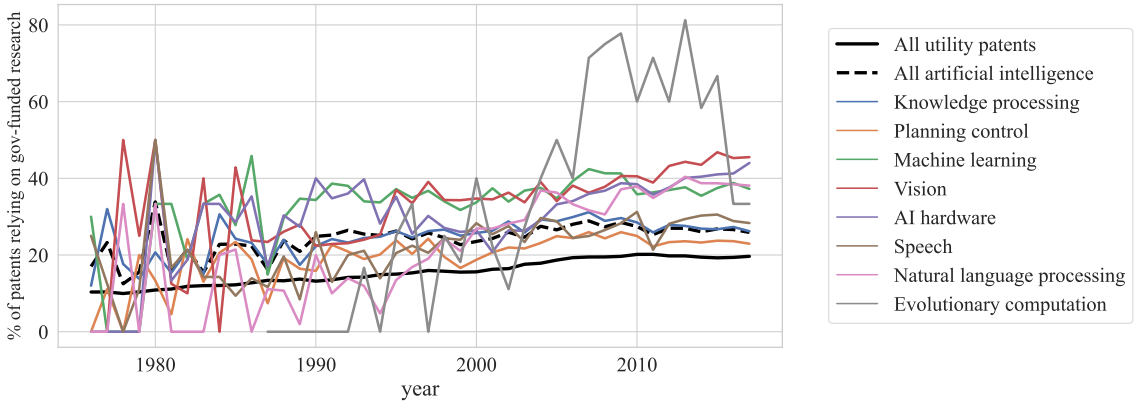


Figure 3.11: Time series of the percentage of patents relying on government-funded research in each AI subdomain. The solid black line and dashed black line represent utility patents in general and all AI patents as a comparison.

As shown in Figure 3.11, AI patents rely more heavily on knowledge created with the support from the U.S. government (19%) than utility patents in general (17%), with patents related to evolutionary computation (41%), AI hardware (28%),

NLP (27%), and machine learning (26%) associated with the strongest reliance on federally-funded knowledge.

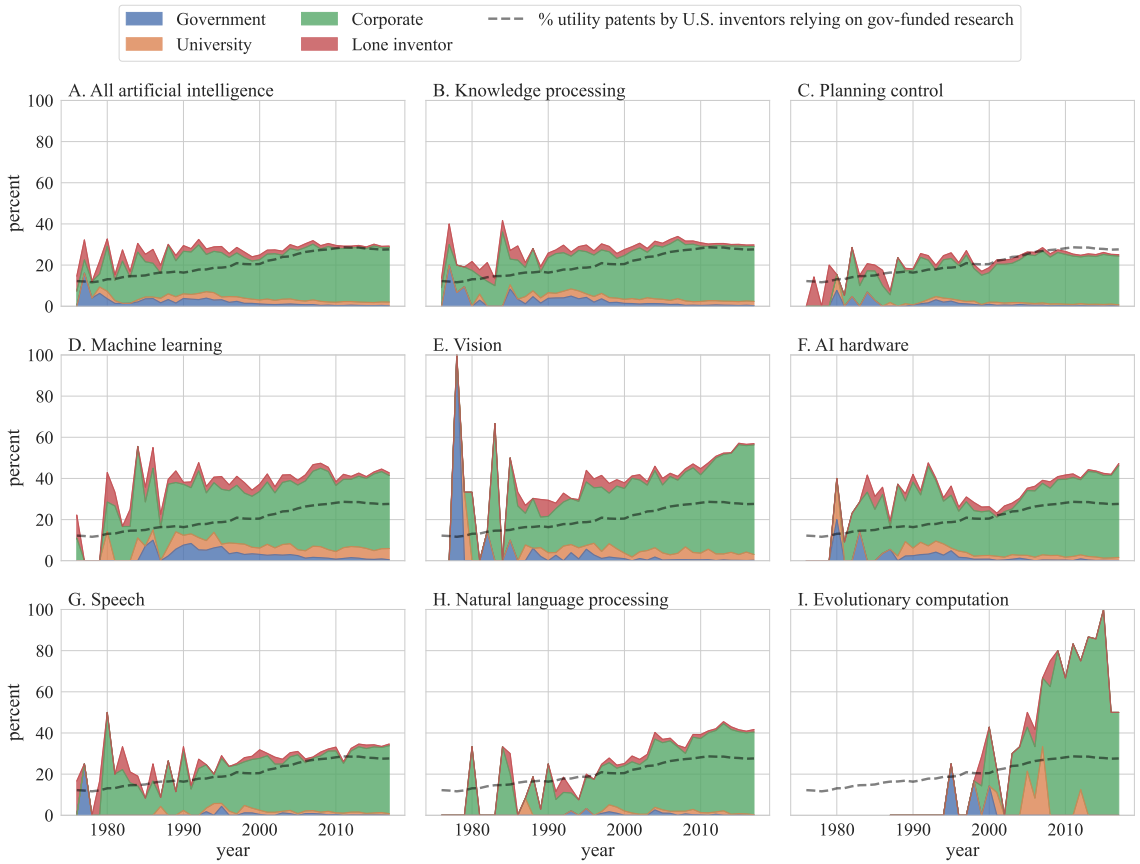


Figure 3.12: Stacked area chart of the percentage of AI patents by U.S. inventors owned by the government or rely on scientific knowledge created from government-supported research by the type of patent owner. Dashed black lines represent the percentage of utility patents by U.S. inventors that rely on government-supported research.

Like utility patents in general, private corporations account for the largest share of different types of patent owners that use government-supported scientific knowledge to create AI inventions, as shown as the green areas in Figure 3.12. Nevertheless, while the percentage of utility patents by U.S. inventors that rely on public support has grown from around 10% in the 1970s to 30% in 2017, such percentage tends to remain stable at around 30% for AI patents from 1976 to 2017. Among the eight AI subdomains, patents related to evolutionary computation, machine learning,

computer vision, and AI hardware have substantially depended on public support. Especially in the case of evolutionary computation, more than 80% of patents relied on government-supported research in the 2010s.

Figure 3.13: Stacked area chart showing the percentage of U.S. patents by foreign inventors that rely on government-supported research by country. The dashed black lines (not stacked) represent such a percentage for utility patents in general.

Fleming et al. (2019b) has observed a steady increase in the proportion of U.S. utility patents by foreign inventors who rely on U.S. government-funded research. Japan and Germany are associated with the highest reliance on U.S. public support. I observe an even higher proportion of foreign-invented government-supported patents in AI inventions, as shown in the panels in Figure 3.13. Japan (pink) is clearly the top foreign country where many inventors are utilizing knowledge created by

research funded by the U.S. government, but its proportion has declined substantially. Germany (blue) has surpassed Japan in creating AI hardware patents (Panel F) with knowledge funded by the U.S. government.

Reliance on Government Support Increases Citation but Decreases Disruption

AI patents invented with the help of scientific knowledge created by government-funded research tend to have more citations in general than other patents. However, they have a much lower probability of becoming top-cited. On the other hand, those same AI patents that depend on government-supported knowledge tend to disrupt future inventions to a significantly lower extent than others. However, they have a higher probability of becoming the most disruptive patents.

An investigation into the relationship between the reliance on government support and the citations received by and the disruption caused by the patents in each domain reveals that AI patents relying on public support tend to receive higher citations but disrupt less. In comparison, an average AI patent not relying on government support receives fewer citations but causes higher disruptions.

In each panel in Figure 3.14, the left group represents patents that do not rely on government-funded research. The right group represents patents that are owned by government agencies or invented with knowledge created by federally-funded research. The first thing observed in each panel is that the blue bar in the left group is almost always taller than the blue bar in the right group, indicating that most patents do not rely on government-supported knowledge except for evolutionary computation (Panel I). Regardless of panels, the green and red bars (mean and median citation percentile of the year) in the left group are always lower than the green and red bars in the right group, indicating that patents created with research funded by the government have

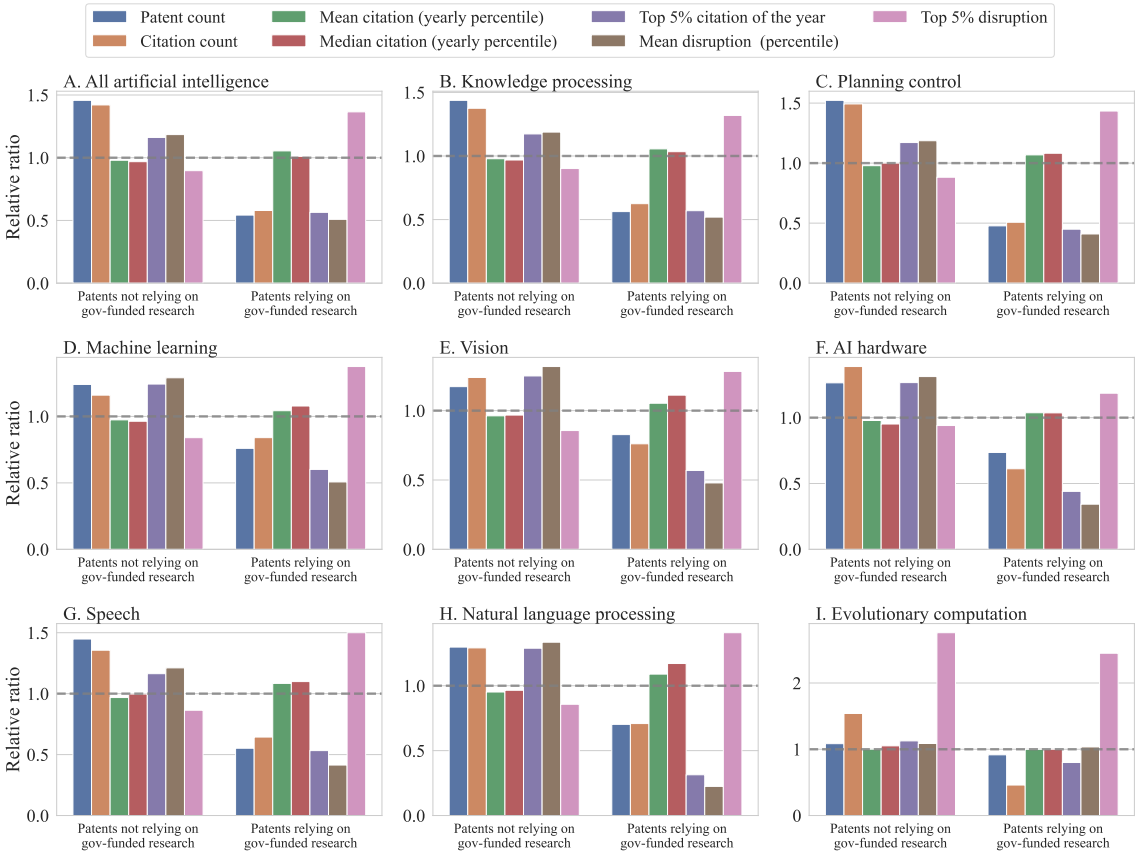


Figure 3.14: Bar plots showing the relative ratios regarding patent count (blue), citation count (orange), mean citation percentile of the year (green), median citation percentile of the year (red), top 5%-cited patents of the year (purple), mean disruption percentile (brown), and top 5% patents in terms of disruption percentile of patents classified by the reliance on government support in each AI subdomain. The horizontal dashed black line in each panel represents the fictional relative value if the variable is evenly distributed.

higher citations in general than patents otherwise in the same year. Nevertheless, extreme cases differ from general cases. The left purple bars (top 5% hit rate in patents not relying on government support) are always significantly higher than the purple bars on the right. In other words, the hit rate of patents that do not rely on public support is higher than the background rate, implying that a patent without government funding has a higher chance of becoming top-cited, although an average patent without government funding receives fewer citations than those with public support.

On the contrary, when considering disruption, an average AI patent relying on government support disrupts subsequent inventive paths to a much lower degree than an average AI patent that does not rely on public support, demonstrated by the illustration in Figure 3.14 that brown bars in the right groups are almost always much lower than brown bars in the left groups. Nevertheless, government-supported inventions have a higher hit rate — the pink bars (top 5% hit rate in terms of disruption) in the right groups are taller than the pink bars in the left groups in almost every panel except evolutionary computation (Panel I).

I note that because patents in evolutionary computation are very small in number (only 237 patents in total), it is less meaningful to attempt to interpret a general pattern in the subdomain.

3.5.6 Team Size Matters in AI Patenting

Recent research in the field of the science of science has investigated how team size affects the characteristics of the outcome of innovation. The most prominent study by Wu et al. (2019) has surveyed multiple large datasets in different domains, such as code repositories, scientific publications, and patents, and found that team size matters to the citations and disruptiveness of the works. Specifically, the findings of Wu et al. (2019) can be summarized by the title of their research — large teams develop, and small teams disrupt.

This section investigates whether such a finding is observable in AI patents, specifically, whether AI patents created by small teams differ from those invented by large teams regarding their scientific impact and technological disruption to future AI patents. In addition, to further extend such a line of rationale, this chapter also surveys how small and large teams differ in the source from which they draw technological novelty to create new AI patents.

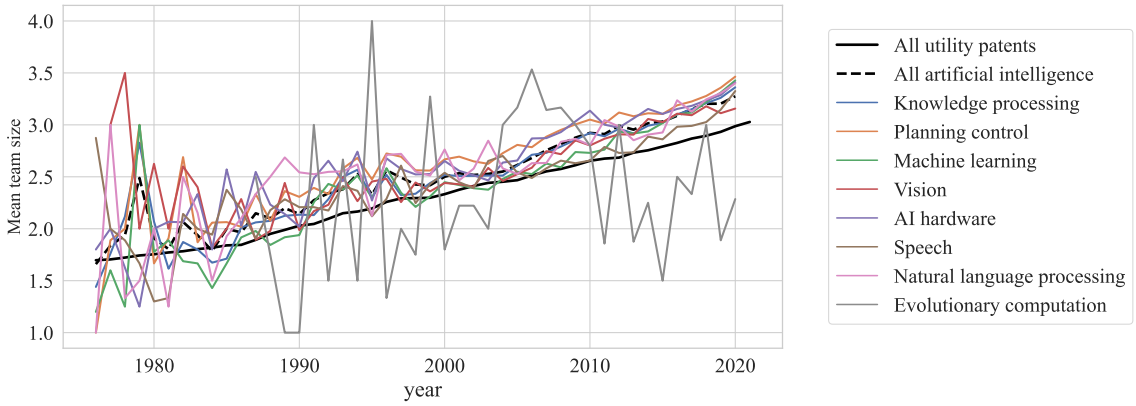


Figure 3.15: Time series of the mean team size of patents in utility patents in general (solid black line) and patents in AI (dashed black line) and its subdomains show that regardless of domains, the average patent team size is growing from 1976 to 2020. The average number of inventors involved in AI inventions appears to be larger than an average patenting team in general.

The results of this chapter suggest that consistent with the findings of Wu et al. (2019), large teams tend to create AI patents that receive more citations, while small teams create more AI patents with higher disruption to knowledge networks. Such a conclusion can be inferred from either general cases (e.g., mean values) or extreme cases (e.g., hit rates). Furthermore, team size also matters to the source of technological novelty that differentiates an AI patent from prior art — a higher level of novelty is associated with a smaller team size. Generally speaking, smaller teams tend to produce patents that are more novel. In particular, among the four categories of technological novelty, origination patents tend to have the smallest teams, while refinement patents often have the largest team.

AI Inventor Teams are Growing

Before presenting the results of how team sizes affect AI patents' characteristics, it is crucial to understand the time trends and distribution of AI patents' team sizes. As Figure 3.15 shows, like utility patents in general (solid black line), the size of teams that create AI patents has expanded significantly from 1976 to 2020. It used to take

about two inventors to create an AI patent in the 1980s. By 2020, an average AI patenting team comprises more than three people (3.3). In addition, half of the AI patents around 1980 were created by lone inventors, and since 2012, half of the AI patenting teams have more than three people.

Nevertheless, the average number of inventors involved in AI patenting appears to be always slightly larger than utility patents in general. The gap between the two has expanded, especially since the middle 2000s.

Most teams in AI patenting are small, ranging from one to three inventors. An average AI patenting team has three inventors (see the histogram distribution (Figure J.1) and statistical summary (Table J.1) of team sizes in Appendix J). Regardless of AI subdomains, over 99% of the teams are sized with no more than ten people.

Large Teams Develop, and Small Teams Disrupt AI Patenting

Figure 3.16 illustrates the relationship between team sizes (x-axes) and citations (right y-axes, red lines) or disruptiveness (left y-axes, green lines) in each AI subdomain. (Only teams with sizes of no more than ten inventors are included in Figure 3.16 for visualization purposes because over 99% AI patents are produced by teams sized no more than ten people.)

It is reasonably clear that in almost every panel of Figure 3.16, as team size grows, the red line (citation) tends to trend upwards, while the green line (disruption) declines. An AI patent created by a lone inventor would cause an average disruption higher than 43% of AI patents. In comparison, an AI patent created by a team of ten inventors would have an average disruption more substantial than about 53% of AI patents, about 10% higher than lone inventors. Contrarily, patents by lone inventors tend to receive 72% more citations than other patents. In contrast, patents created by nine inventors on average are ranked around 65% in terms of citation (about 7%

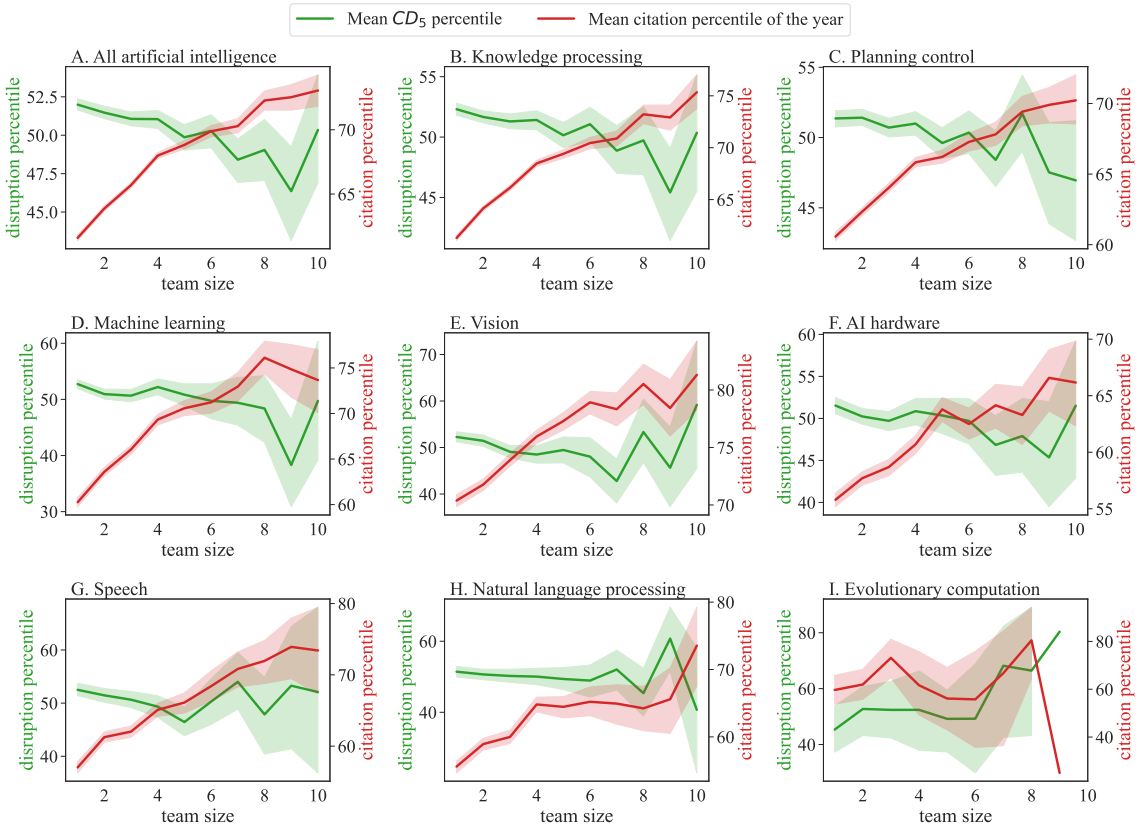


Figure 3.16: Relationship between team sizes and citation and disruption. Green lines are measured by the left y-axes, which represent the percentile of mean CD_5 indices. Red lines are measured by the right y-axes, which gauge the mean citation percentile of the year. Color bands represent 95% confident intervals. Only teams with no more than ten inventors are shown in this figure because such criteria cover more than 99% of patents in each AI subdomain.

lower than patents by lone inventors).

Nevertheless, it can be observed from several panels (A, B, D, E, and F) that while the green line does decline at the beginning, it appears to increase at the end, where the team size is equal to ten. The data presented in this chapter shows very few patents in each AI subdomain created by teams of ten (mostly 0.4%). Therefore, such drastic fluctuation at the end may result from a lack of data points.

In addition, such a "large-teams-develop-and-small-teams-disrupt-AI" phenomenon can be inferred from extreme cases where patents are ranked among the top tiers in either citation or disruption. Figure 3.17 illustrates the relative ratios of the top-5%

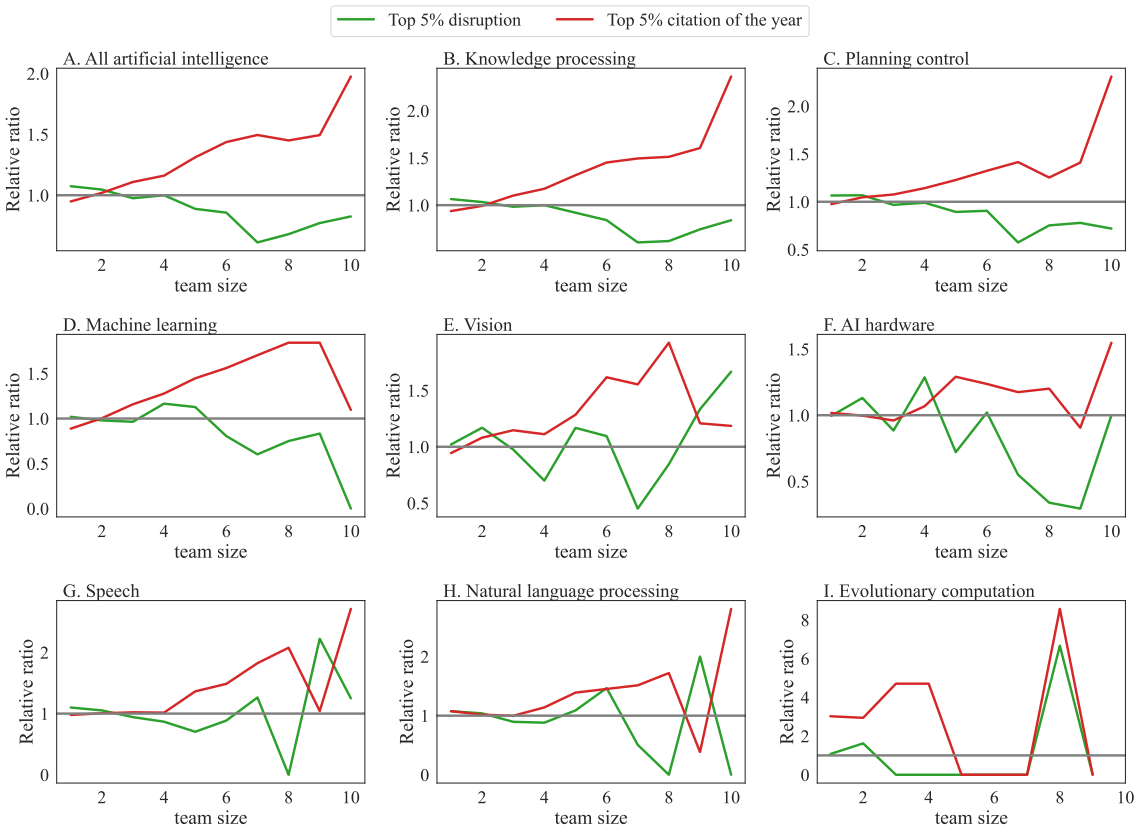


Figure 3.17: Relative ratios of the top 5% hit rates in terms of disruption (green lines) and citations (red lines) by team size. Gray lines represent relative ratios equivalent to the background, namely 5%. In other words, a relative ratio of 2 regarding disruption indicates 10% of patents are ranked in the top 5% in terms of their CD_5 indices. Only teams with no more than ten inventors are shown in this figure because such criteria cover more than 99% of patents in each AI subdomain.

hit rates in terms of citation percentile of the year (red) and disruption percentile (green). The grey lines represent the constant background hit rate, which in this case equals 5%. A relative ratio of 2 with a background hit rate of 5% can be inferred as 10%, which indicates that 10% of the patents in the category in question are ranked in the top 5% of all patents across categories. Again, a notable general pattern can be observed from the panels of Figure 3.17 — red lines tend to trend upwards while green lines decline as team size grows. For instance, Panel A indicates that AI patents created by lone inventors have a lower probability of being ranked in the top 5% among all the AI patents granted in the same year in terms of citations

but a higher probability of being ranked highly in disruption. On the contrary, an AI patent by a ten-people-team can be expected to have a 10% probability of being ranked in the top 5% (twice as the background hit rate) in terms of citation. In subdomains like speech and NLP, such citation hit rates of ten-people-teams are even higher (around 14%, almost three times the background hit rate). However, when it comes to disruption, a team with seven people would only have a 3% hit rate, much lower than the background hit rate (5%). Such an observation is generally consistent across AI subdomains, except evolutionary computation (Panel I), which has very low patent counts.

A detailed breakdown by percentiles regarding citations and disruption of AI patents by teams of different sizes can be found in Appendix K. In the two figures in Appendix K, patents in each AI subdomain are binned into twelve categories by their citation percentile or disruption percentile (see the legends). Areas colored blue represent patents ranked in the lower half, and red areas represent patents in the upper half in terms of percentile. A general and clear trend can be observed from Figure K.1 that red areas tend to expand as team size grows from one to ten, indicating that patents by larger teams tend to receive more citations. An opposite but weaker trend can be observed from Figure K.2 — blue areas tend to occupy larger space as team size grows, indicating that larger teams tend to produce less disruptive patents.

Origination Patents Have Smaller Teams, While Refinements Have Larger Teams

I find that origination patents tend to be produced by smaller teams, while refinement patents tend to be invented by larger teams.

Appendix J shows that different AI subdomains tend to have similar distributions regarding team sizes — the mean value and median value of team sizes are both

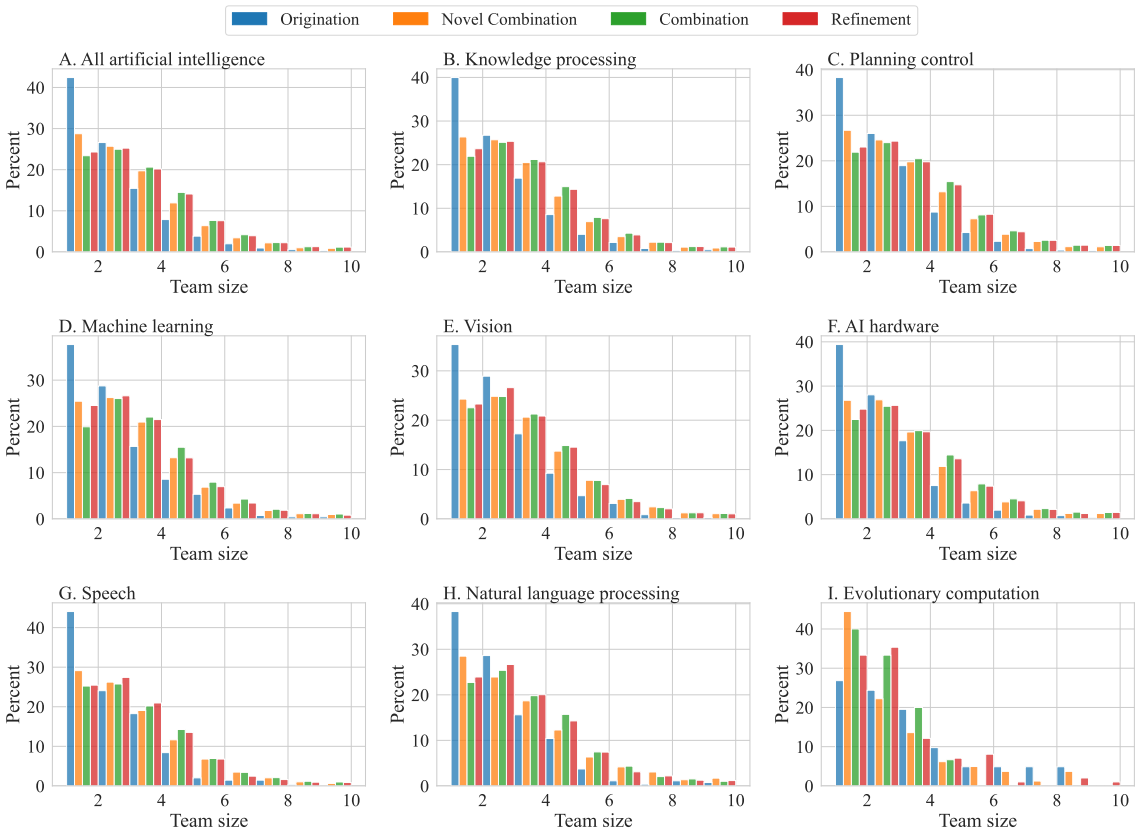


Figure 3.18: AI patents' team size distribution categorized by the sources of technological novelty.The height of the bars represents the percentage of patents in the category produced by a specific size of teams. For instance, in Panel A, the leftmost blue bar is about 42%. It indicates that about 42% of origination patents in AI are produced by lone inventors or teams of one.

around three, and the most frequent team size is two. However, different sources of technological novelty are observed with significantly different distributions in terms of team size, as shown in Figure 3.18.

In almost every panel in Figure 3.18 (except Panel I), origination patents (blue bars), which are associated with the highest novelty, are the most skewed. Lone inventors produce around 40% of origination patents, and less than 30% of origination patents have two inventors. Only 4% origination patents are brainchildren of teams with more than five people. Patents of novel combinations (orange bars), which combine new components with existing ones, are less skewed than origination patents

but more skewed than patents of combinations (green bars) or refinement patents (red bars). The percentages of lone inventors and two-people teams in novel combination patents are close to each other, around 25%. In machine learning and computer vision, two-people teams account for a higher proportion than lone inventors.

In the cases of combination patents or refinement patents, about 20% of patents are produced by lone inventors, invariably lower than two-inventor teams (around 25%). About 10% patents of combinations or refinements are produced by teams larger than five people, more than twice than that of origination patents.

In general, panels in Figure 3.18 illustrate that origination patents have a much higher percentage (nearly twice) that of lone inventors and a much lower percentage of larger teams. In contrast, patents of combinations and refinements, which are low in novelty, have a lower percentage of lone inventors and a higher percentage of patents produced by large teams.

It is also evident from Table L.1 in Appendix L that regardless of subdomains (except evolutionary computation), the team sizes of AI patents of originations (mean 2.21 and median 2) are always smaller than refinement patents (mean 2.98 and median 3).

Evidence shows that novel AI patents (originations) tend to have smaller teams, while conventional AI inventions (such as refinements) tend to have larger teams.

However, the above analyses cannot be simply translated to "smaller teams tend to originate while larger teams tend to refine AI patents" because the probability of a patent being an origination is invariably and substantially lower than other categories, although the "origination chance" of a team of one is indeed higher than a team of ten. As shown in Appendix M, especially Figure M.2, which shows the percentage of patents with four sources of technological novelty in patents produced by different sizes of teams, the composition of the four categories remains relatively stable across

different sizes of teams. For instance, in panel A, a team of one has almost the same chance as a team of ten of producing a refinement patent (about 50% probability, as shown in the red area).

I plotted the relative ratio of the percentage of AI patents belonging to the four novelty categories (see Figure M.3). A relative ratio that equals one indicates the background value if the percentage is evenly distributed across different team sizes. Most panels in Figure M.3 except for Panel I (evolutionary computation), show that the blue line (origination patents' relative percentage) declines as team size grows, indicating that smaller teams have a higher chance than larger teams of producing origination patents. For example, in the subdomain of knowledge processing (Panel B), 2% of patents created by lone inventors are categorized as origination patents. At the same time, such a percentage drops to only 0.5% in patents produced by teams of ten inventors. However, no visually obvious pattern for the other three categories can be observed.

To conclude, this chapter found that AI patents with high originality tend to be produced by smaller teams, while larger teams often create AI patents that refine existing technologies. In addition, smaller teams (particularly lone inventors) have a slightly higher chance of producing highly original AI patents. However, the probability of the other three categories remains relatively stable across different team sizes.

3.5.7 AI Patents Compared to AI Publications

Chapter 2 investigated the characteristics of AI publications and how they are related to AI publications' scientific impact. I found that AI publications are driven heavily by conventional combinations of existing knowledge, and such publications are also the most impactful. This chapter presented how novel, disruptive, diverse, and

impactful AI patents are and how team sizes may relate to those features. Again, the analysis in this chapter found that AI patents are low in novelty. AI inventions rely heavily on refining existing technical components rather than introducing radically new capacities. In short, AI's knowledge creation has not deviated from traditional practices and is not radically different from the incremental manner in which science and technology have been advancing historically.

Nevertheless, what can be found if comparing AI patents with AI publications? What can this comparison inform us about the unique characteristics of AI knowledge and invention and innovation in general? This section compares two novelty taxonomies on AI patents and AI publications (*Knowledge Recombination Taxonomy* and *Technological Novelty Taxonomy*) to characterize the way new knowledge has been created in the field of AI over the past decades.

Broader Science, Deeper Patenting

Although the number of journals referenced by AI papers (green dash line in Panel C of Fig. 3.19) has been invariably lower than the number of technology codes assigned to AI patents (solid green line), the number of combinations in AI papers (dashed orange line) has been continuously higher than technology code pairs in AI patents (solid orange line). This difference underscores a more substantial extent of knowledge combination or exploitation in AI research than in AI patenting. It is not surprising that, more journal articles are commonly referenced in a research article than the number of technology codes assigned to a patent. This observation is again demonstrated in Fig. 3.20. The left panel shows the distribution of the number of technology codes assigned to each AI patent. In contrast, the right panel shows journal articles referenced by each AI publication. Half of AI publications reference more than 23 journal articles, while half of the AI patents have less than

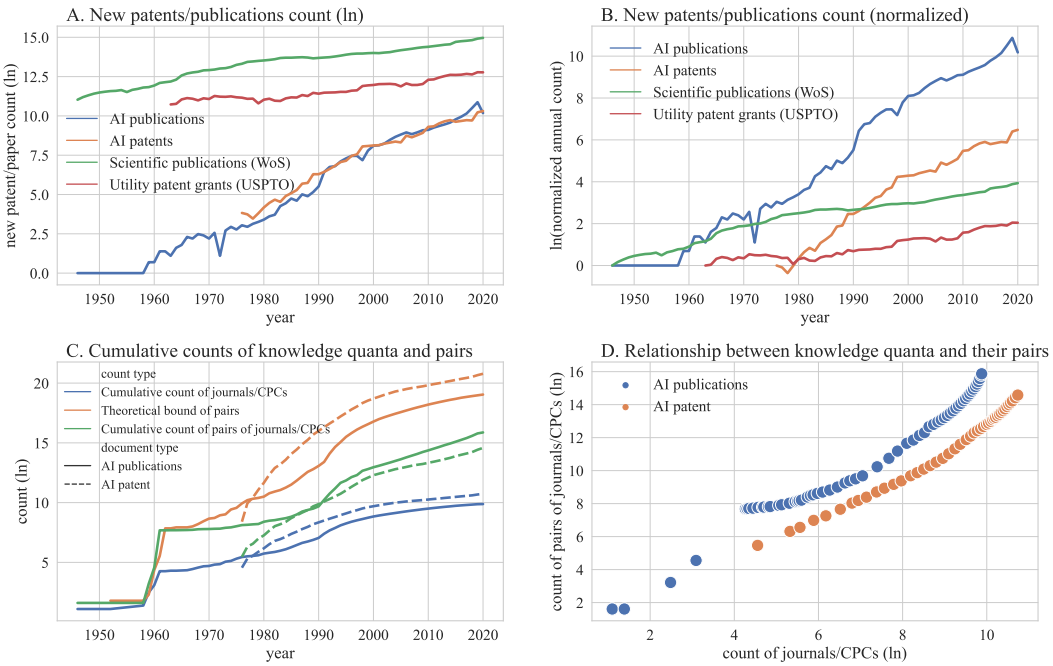


Figure 3.19: (A) Time series of the natural logarithm of the number of AI patents (orange) and AI publications (blue) compared to inventive activity in general — the green line represents scientific publications recorded in WOS Core Collection (1946–2020), and the red line represents utility patents granted by the USPTO (1963–2020). (B) Time series of AI inventions (patents and publications) compared to inventive activity in general, normalized using Equation 3.2. (C) Time series of the cumulative number of knowledge quanta, their pair-wise combinations, and their theoretical bounds (TB, computed using Eq. 3.3) in AI inventions. For patents, knowledge quanta are technology codes, while in publications, cited journals. (D) Scatter plot between the cumulative knowledge quanta and their pair-wise combinations in AI inventions.

four technology codes. Journal articles may cite any influence on an author's thinking, including articles with whose contents the author may disagree. The number of patent claims may constrain the number of its technology codes. It limits the technology codes to the number of technologies that define in the inventor's monopoly property right, a far smaller number than may be cited found in a patent's prior art citations.

The set of n technology codes classifying a patent's technologies or the set of n journal articles referenced by a publication is referred to as an *n-tuple*. The size of an n-tuple offers an estimate of how many technologies are combined in an inven-

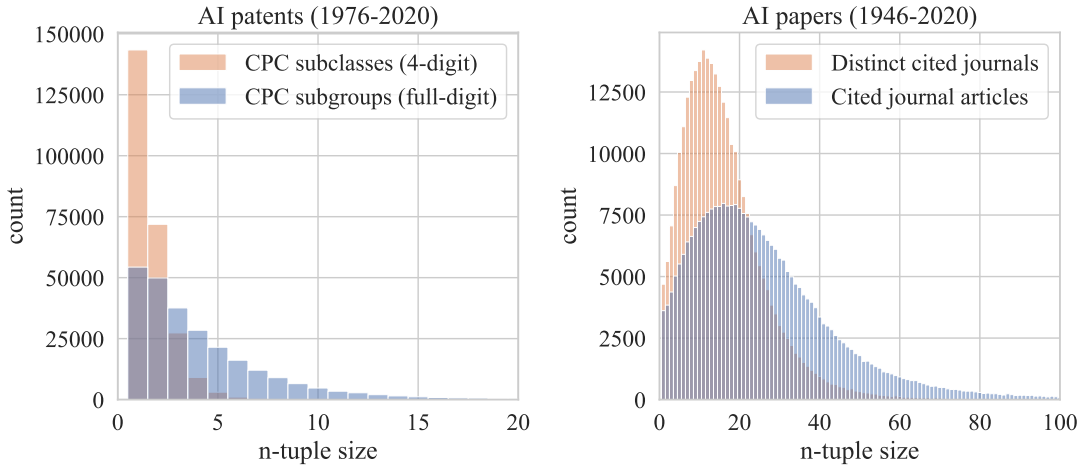


Figure 3.20: Distributions of knowledge n-tuple size of each AI patent or publication. The left panel shows the distribution of the number of technology codes assigned to each AI patent. Blue histograms represent the distribution of CPC subgroups (full-digit codes), and orange represents the distribution of distinct CPC subclasses (4-digit codes). Only inventive CPC codes are counted. Additional codes are excluded. The right panel shows the distribution of the journal articles referenced by each AI publication. Blue histograms represent the number of referenced journal articles regardless of whether they are published in the same journal(s). Orange histograms represent referenced journals that are distinct. In other words, if multiple referenced articles are published in the same journal, the are counted as one "distinct journal."

tion. Nevertheless, as discussed in Section 3.4.3, it is meaningful to distinguish broad combinations from deep combinations (Strumsky and Lobo, 2015a; Youn et al., 2015) because they differ significantly in the faculty, cost, and risk involved in the inventive process. A deep combination refers to a pairing of technologies that are close to each other and usually engages less inventive effort to combine, for instance, a keyboard and a mouse. Otherwise, a combination can be considered broad if the two components are more distant in subject matter, for instance, a cancer drug and a quantum computer. In the context of patents, two technological components that belong to two different CPC subclasses (4-digit codes) are considered distant. Otherwise, they are considered closely related to each other. In the context of publications, two referenced journal articles published in the same journal are considered as closely related

and two articles published in different journals as distant in terms of knowledge.

Figure 3.20 shows the distributions of the number of distant technologies combined in each AI patent or AI publication (orange histograms) as compared to the distributions of the sizes of n-tuples (blue histograms). In both panels, a significant extent of deep combinations can be implied because the orange histograms representing broad combinations are considerably more centered towards the left, with much higher peaks and fewer spreads. It suggests that a substantial proportion of knowledge combinations in AI are deep. Eighty percent of AI patents combined multiple technologies. However, as high as 44% of those multi-coded AI patents have all their involved technologies belonging to the same fields, indicating deep combinations. On the other hand, on average, 27.4 journal articles are referenced in AI publications, about 40% of them are published in the same journals as the rest, and only 16.1 distinct journals are cited on average. Nevertheless, the less significant discrepancy in the right panel implies that AI research tends to combine more broad knowledge than AI patents. Therefore, AI science is broader, while AI patenting is deeper.

Incremental Advances

The results of the two taxonomies implemented on the two datasets are shown in Table 3.2. I found that AI publications and AI patents have remarkably similar compositions in Knowledge Recombination Typicality Taxonomy. AI invention consists mainly of incremental advances that combine existing knowledge in conventional ways. It is demonstrated by the large share of *Accepted Wisdom* that are highly conventional and low in novelty, and it occupies almost half of AI inventions. *Avant-Garde* that is highly novel and less conventional takes slightly smaller but still considerably large proportions (over 30%). *Darwin's Tower* and *Platypus* have the smallest shares. This suggests that regarding knowledge recombination, AI inventions are driven by incre-

Table 3.2: Percentage Composition of *Knowledge Recombination Taxonomy* and *Technological Novelty Taxonomy* applied to AI patents and AI publications.

Knowledge-Recombination Taxonomy					
	AI Patents			AI Publications	
High Novelty	*Avant Garde* 32.1%	*Darwin's Tower* 11.8%	High Novelty	Avant Garde 34.7%	Darwin's Tower 11.5%
Low Novelty	*Platypus* 11.4%	*Accepted Wisdom* 44.7%	Low Novelty	*Platypus* 7.4%	*Accepted Wisdom* 46.4%
	Low Conventionality	High Conventionality		Low Conventionality	High Conventionality
Technological Novelty Taxonomy					
	AI Patents			AI Publications	
Combining New Knowledge	*Originations* 0.99%	*Novel Combinations* 8.6%	Combining New Knowledge	*Originations* 0.08%	*Novel Combinations* 6.11%
Combining No New Knowledge	*Combinations* 40.19%	*Refinements* 50.22%	Combining No New Knowledge	*Combinations* 73.55%	*Refinements* 20.26%

mental improvements built on solid foundations of conventional knowledge, followed closely by inventive activities emphasizing relatively new combinations.

Notwithstanding, AI patents and AI publications resemble each other less in Technological Novelty Taxonomy, as shown in the second part of Table 3.2. In AI patents, refinement patents take the largest share (over 50%), followed closely by combination patents (over 40%). On the other hand, in the case of AI publications, about 74% are combination patents while only 20% are categorized as refinements. This difference underscores that AI patents and publications diverge in their source of technological novelty. While an approximately equal amount of technological novelty in AI patents is from exploiting existing combinations and exploring new ways of combining existing components, AI publications are more heavily sourcing their novelty from the latter. Nevertheless, AI patents and AI publications indeed share one similarity – neither of them draws much novelty from radically new elements that have never been observed

in AI before, as indicated by the small shares of novel combinations and originations.

Conventional Knowledge Driving AI Inventions

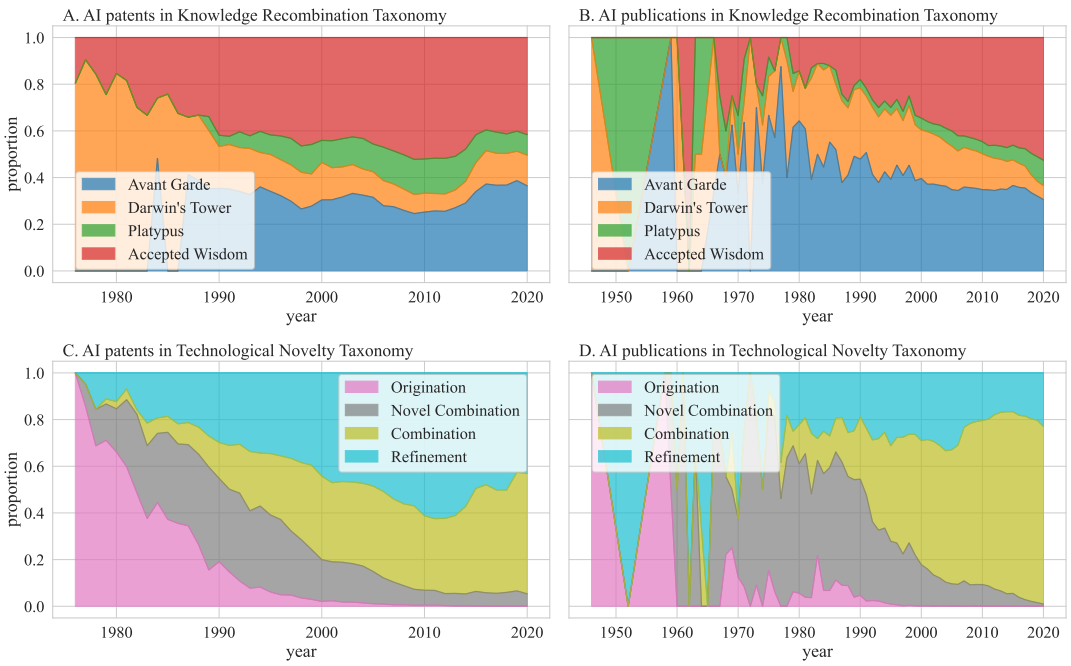


Figure 3.21: The time series of category compositions of *Knowledge Recombination Taxonomy* and *Technological Novelty Taxonomy* on AI patents and publications. The two panels on the top represent *Knowledge-Recombination Taxonomy* implemented on AI patents and AI publications, respectively. The two panels at the bottom represent *Technological Novelty Taxonomy* implemented on AI patents and AI publications, respectively.

I further explored the time series of the two taxonomies on new AI patents and AI publications published each year. The results are shown in Fig. 3.21. (I note that Panel C of Figure 3.21 is the same as Panel A of Figure 3.4 and Panel B of Figure 3.21 conveys the same message as Figure 2.10.) Inventions in *Darwin's Tower* (orange bars in Panels A and B in Fig. 3.21) that mix novel combinations with conventional pairings were suggested previously to create the most impactful scientific knowledge (Mukherjee et al., 2016). However, *Darwins' Tower* inventions appear to be contributing insignificantly to the growth of AI inventions because they have

been diminishing steadily to a small proportion since the 1980s. Instead, this research found that the two leading forces of AI inventions are *Accepted Wisdom* (red) and *Avant-Garde* (blue), of which the typicality characteristics interestingly differ from each other drastically, the former featuring low novelty and high conventionality, while the latter mixing high novelty with low conventionality. The rising of *Accepted Wisdom* is particularly substantial and appreciable in AI publications. In 2020, around half of AI publications are *Accepted Wisdom*. It is consistent with observations that many recent AI research are focusing on tweaking existing models rather than introducing revolutionary ideas, and recent development in AI has been arguably believed to neither gain sufficient robustness nor achieve substantial improvement in their performances regarding practical applications (Marcus, 2018, 2020; Hutson, 2020). The second largest category is *Avant-Garde*, primarily drawing on novel combinations with little conventional pairing. Its share has been somewhat stabilized in AI patents, if not growing, despite a slightly decreasing in AI publications. The dynamics between *Avant-Garde* and *Accepted Wisdom* may suggest that *Avant-Garde* is one of the sources of *Accepted Wisdom* as the creation of radical novelty, and its transitioning into conventional knowledge happens in a similar pace.

The result of *Technological Novelty Taxonomy* shows that Originations and Novel Combinations (blue and orange bars respectively in Panels C and D of Fig. 3.21), the two categories that involve new technological capacities have been diminishing in the recent decades to a point where they almost disappear, although they used to be a considerable force in the early days of AI. Refinements (red) and combinations (green) thrive in AI patents with approximately equal momentum. Nevertheless, in AI publications, Combinations have been taking over most new publications in the recent four decades, with Refinements remaining consistently around 25%. This result shows that technological novelty that roots in exploiting existing knowledge without

introducing new capabilities has become dominant in AI inventions. In addition, AI publications have been introducing more novel combinations than AI patents.

The results of the two taxonomies are consistent in describing exploiting and refining existing knowledge as the most significant force that drives AI invention. Much of the technological novelty, particularly in AI publications, has been increasingly coming from discovering new ways of combining existing knowledge rather than introducing radically new technological capacities.

Novelty is Better Rewarded in AI Patenting than Publications

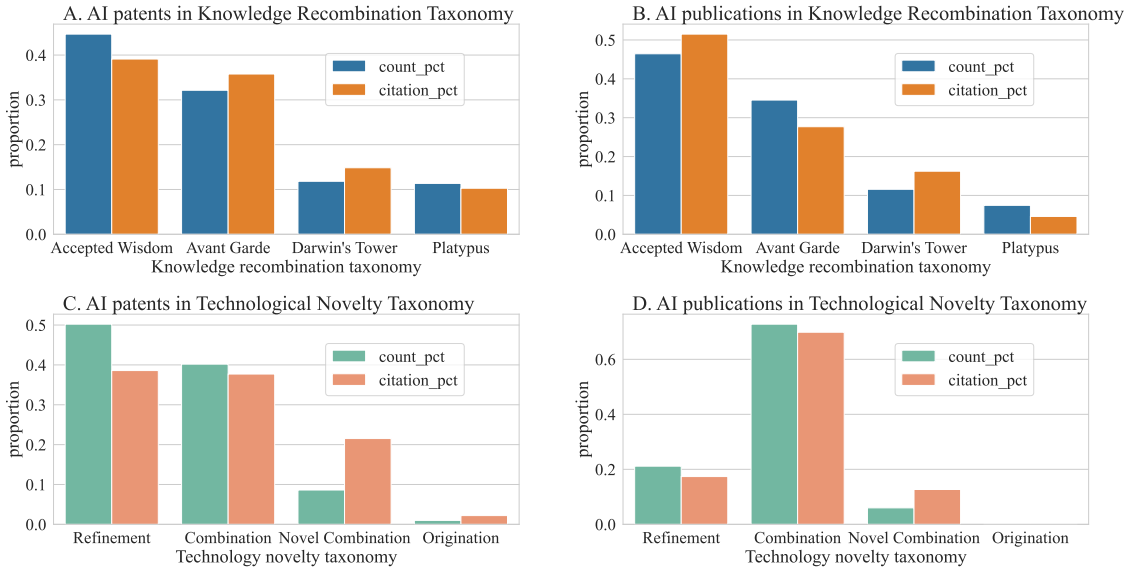


Figure 3.22: Percentage of count and citations of each category of the two taxonomies of AI patents and AI publications.

As discussed in Chapter 2, from the perspective of Knowledge Recombination Taxonomy, AI publications of *Accepted Wisdom* that tend to combine conventional pairs of knowledge receive more citations and thus are more impactful (see Panel B of Figure 3.22). However, in the case of AI patents (Panel A), although the composition of the four categories is considerably similar to AI publications, as demonstrated by the similar heights of blue bars in Panels A and B in Figure 3.22, patents of *Avant*

Garde that feature combinations of novel pairs clearly have gathered more citations on average than other three categories, demonstrated by the higher orange bar than the blue bar.

In Panels C and D of Figure 3.22, in which the results of Technological Novelty Taxonomy are presented for AI patents and publications, respectively, patents of novel combinations (the third group) have substantially more citations (pink bar) on average than other categories.

Such observation offers a shred of preliminary evidence that novelty in AI patents is likely to be better rewarded by citations than novel AI publications because it appears that categories that feature novelty tend to receive higher citations in AI patents than they would in AI publications.

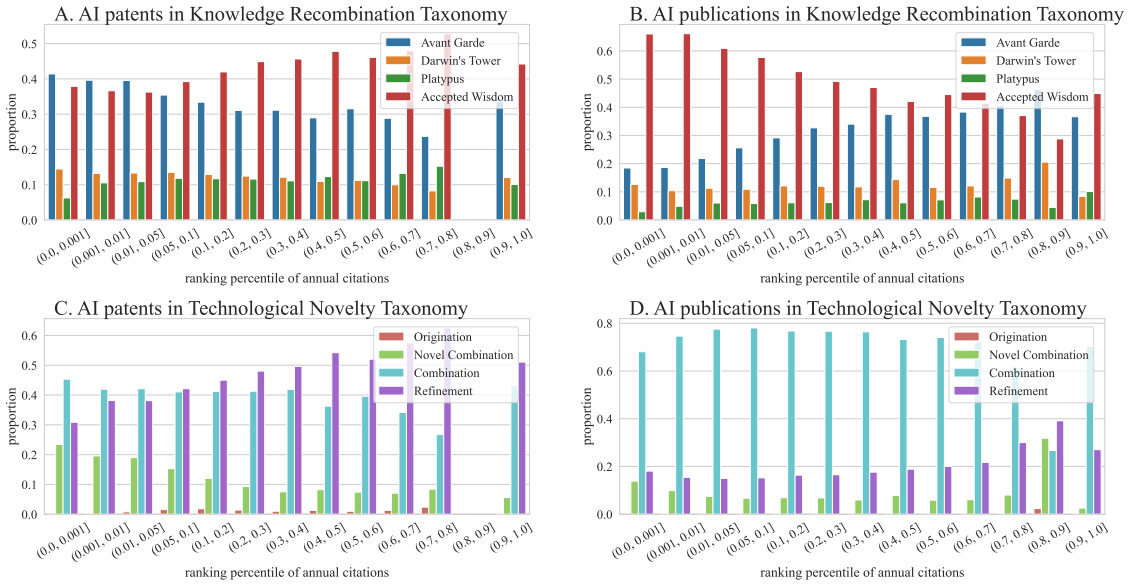


Figure 3.23: Composition of each category in different percentile groups of AI patents or publications ranked by annual citations. The x-axes represent the percentile groups of AI patents or publications ranked from the top. In other words, the more left a group, the higher it is ranked. Each bar represents the proportion of the corresponding category in the percentile group it belongs. For example, the leftmost blue bar in Panel A indicates that among the AI patents that are ranked in the top 0.1%, more than 40% are categorized as *Avant Garde*. Panel B is the same as Figure 2.11.

To examine whether the above observation is robust, it is reasonable to investigate two aspects of the relationship between taxonomic categories and citations. The first aspect is scrutinizing the category composition in each tier of AI patents or publications regarding citations. The second aspect is to observe how much percentage in each category can be ranked in the top tiers regarding citations and to compare such percentages to the background percentages.

Figure 3.23 presents the results of the first aspect. In each panel of Figure 3.23, the x-axis represents groups of AI patents or publications grouped by their percentiles in terms of annual citations (calculated by Equation 2.2). Each bar represents the percentage in the group categorized into the corresponding taxonomic type. For instance, the leftmost blue bar in Panel A indicates that among the AI patents ranked in the top 0.1% in terms of annual citations, more than 40% are categorized as *Avant Garde*. (I note that Panel B is the same as Figure 2.11.)

As Panel B of Figure 3.23 shows, the percentage of *Accepted Wisdom* publications (red) increases in higher-ranked groups while *Avant Garde* (blue) decreases, demonstrated by the red bars on the left being taller than the ones on the right. In the top 0.1% of AI publications, more than 60% are *Accepted Wisdom*, and only less than 20% are categorized as *Avant Garde*. An even smaller proportion (around 11%) of this top group of AI publications are *Darwin's Tower*. However, AI patents (Panel A) present an opposite pattern — in higher-ranked groups, *Accepted Wisdom* (red) decreases and *Avant Garde* (blue) increases. In the top 0.1% AI patents, more than 40% are *Avant Garde*, and around 36% are *Accepted Wisdom*. Similarly, in Panel C, the percentage of refinement patents (purple) decreases in higher-ranked tiers while novel combinations (green) clearly increase. However, in the case of AI publications (Panel D), no clear pattern can be observed.

This indicates that in AI patents, the taxonomic categories that feature nov-

elty (*Avant Garde* and novel combinations) often account for a higher percentage in higher-ranked groups regarding citations. On the contrary, in AI publications, highly impactful publications comprise mainly conventional combinations of previous knowledge. Novel combinations show a much weaker presence in highly-ranked groups regarding citations.

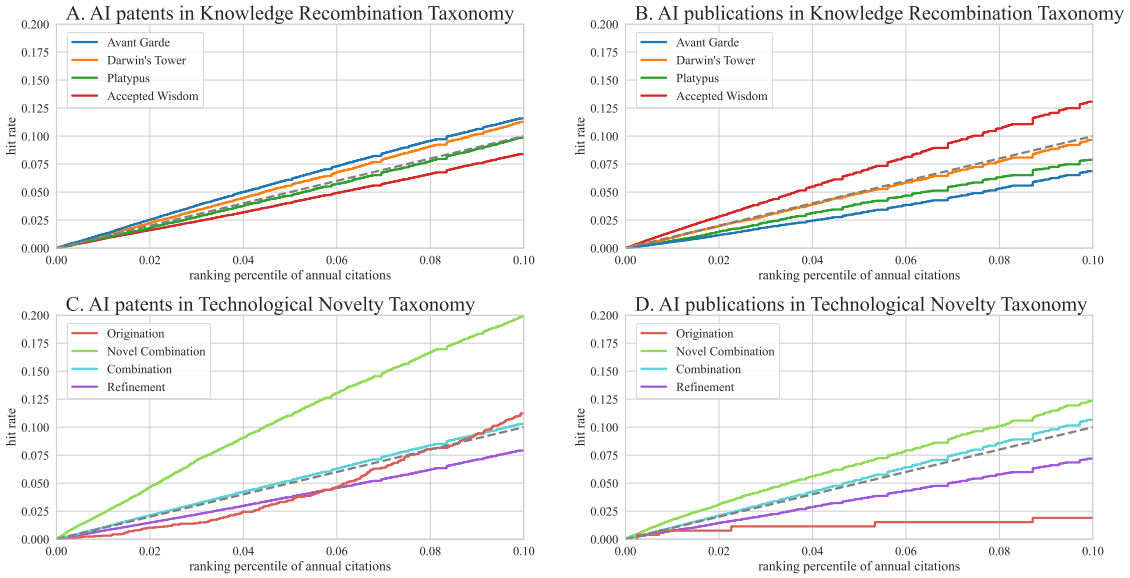


Figure 3.24: Top 0–10% hit rate of each category in AI patents and publications compared to the background. The dashed gray line in each panel represents the background hit rate that allows $y = x$.

The second aspect (hit rates) can be investigated by examining Figure 3.24, in which the hit rate of each category (y-axes) is compared with the background hit rate (x-axes) in terms of annual citations. The panels in Figure 3.24 only select the background hit rates up to 10%. The dashed gray line in each panel represents the background hit rate if hit patents or publications are distributed evenly across the four categories. (The full range hit rate plots can be found in Appendix P.)

Panel B of Figure 3.24 shows that AI publications categorized into *Accepted Wisdom* have the highest hit rate in the top 10% in terms of annual citation. In other words, by combining conventional knowledge pairs rather than novel pairs, an AI

publication has a higher chance (about 13%) of receiving a citation count that would allow it to be ranked in the top 10% among all patents granted in the same year. On the other hand, AI publications that are categorized into the other three types have much lower hit rates. Only about 7% of *Avant Garde* publications are ranked among the top 10%, lower than the background hit rate (10%). However, in AI patents, novelty appears to be better rewarded as the blue line (*Avant Garde*) is the steepest among the four. The red line (*Accepted Wisdom*) is the flattest, indicating that *Avant Garde* that features highly novel combinations have a higher hit rate than patents featuring conventional combinations.

Similarly, from the perspective of technological novelty, AI patents that feature new components combined with existing ones (novel combinations, shown as the green line in Panel C of Figure 3.24) have a significantly higher hit rate than the other three categories. Twenty percent of AI patents of novel combinations are ranked in the top 10% of annual citations. Only around 7.5% AI patents of refinement (purple line in Panel C) are ranked in the top 10%. Interestingly, the hit rate of AI patents of origination (red line in Panel C) rises from the lowest among the top 5% to the second highest among the top 10%. In the case of AI publications, novel combinations again have the highest hit rate, but it is much lower than AI patents of novel combinations. While 20% of AI patents of novel combinations are ranked in the top 10%, the top 10% hit rate for AI publications of novel combinations is only 12.5%.

Therefore, it can be inferred from previous analyses that citations better reward AI patents featuring novel combinations of knowledge than AI publications. In AI publications, novelty sometimes imposes a penalty in its subsequent impact, particularly the ones with radically new ideas. A novel AI patent is more likely to be cited highly, while a conventional AI publication has a higher chance of becoming a hit paper.

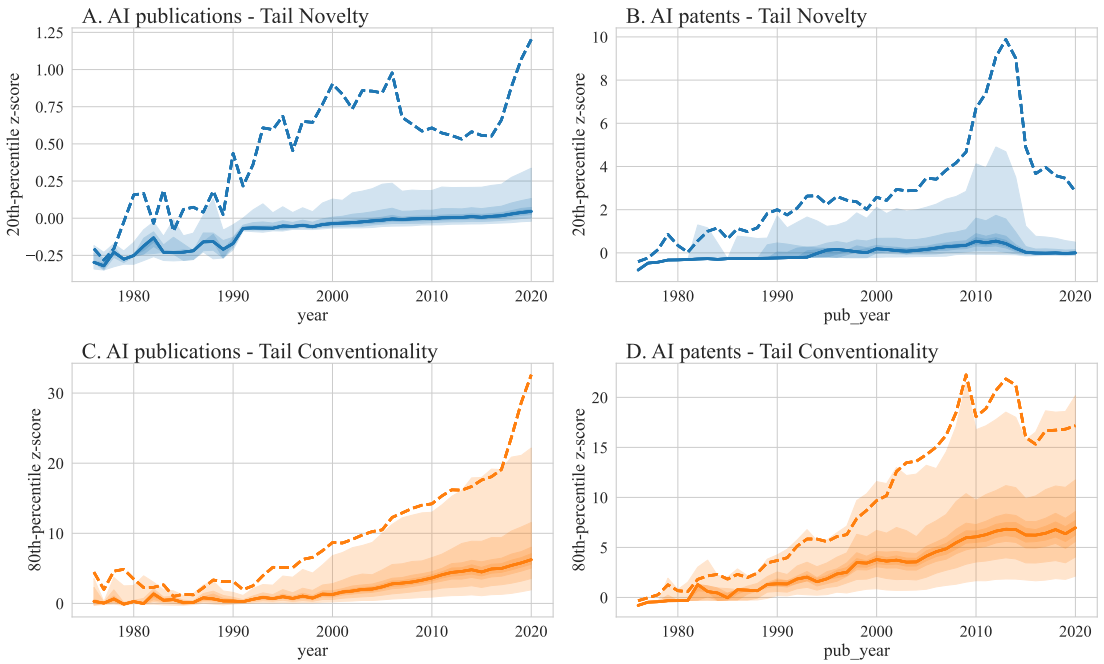


Figure 3.25: Time series of (A) TN, or tail novelty feature of AI publications, (B) TN of AI patents, (C) TC, or the tail conventionality features of AI publications, and (D) TC of AI patents. In each panel, the solid and the dashed line represent the year's median and mean values, respectively. The different shades of the bands around the solid line, from dark to light, encompass the values fifth, tenth, twenty-fifth, and fiftieth percentiles. It is worth noting that in panels A and B, a higher tail novelty feature typicality is associated with lower novelty.

As discussed in Chapter 2, Knowledge-Recombination Taxonomy offers two measures (TN and TC) for each invention to describe its tendency to combine novel or conventional pairs of previous knowledge. Fig. 3.25 shows the time series of such two measures for new AI patents and AI papers each year. It is worth noting that in Panels A and B, the higher the TN value, the lower the novelty.

In AI publications, the mean and median values of conventionality (Panel C) are continuously increasing while the median novelty (Panel A) has been invariably declining. It indicates that AI scientific research has become more conventional and less novel from 1976 to 2020.

In the case of AI patents (Panels B and D), a similar trend of increasing conventionality and declining novelty can be observed until the early 2010s, when a reversed trend began. Mean conventionality dropped with median conventionality plateaued, and novelty increased temporarily. This glitch by no means coincidentally synchronizes with the establishment of the Mayo/Alice test, a test practiced by the USPTO and the Federal Circuit that substantially increased the difficulty of AI-related patent applications to be granted, increasing invalidation and rejection for software patents (Tran, 2016; Saltiel, 2019; Toole and Pairolero, 2020). Although the Mayo/Alice test did not affect the growing trend of AI patents, it altered how AI patent applications are written and might have encouraged a certain level of novelty in AI inventions or at least prompted inventors to find new ways of drafting AI patents, bringing in an increase in novelty and a decrease in conventionality.

An alternative interpretation of the increased novelty and stagnation of conventionality in AI patents in the 2010s can be partly credited to the fact that many new machine learning techniques, such as deep neural networks, started to gain momentum in the 2010s because of the increased availability of large training data, better hardware, and improved algorithms.

Searching Slowing Down

Given the assumption that invention (the creation of new knowledge) is primarily driven by the combination of existing knowledge, a patent can be considered as a combination of technologies represented by technology codes and a scientific publication as a combination of units of knowledge represented by referenced journals. Thus, the numbers of technology codes covered by AI patents and the number of referenced journals in AI publications offer a glimpse into the expansion of scale and diversity of AI methods, techniques, and applications that have been recognized and classified to

date. Panel C of Fig. 3.19 shows the time series of the cumulative number of technology codes and referenced journals in AI (green), the cumulative number of code-pairs and journal-pairs in AI (orange), and the theoretical bound of combinations given the stock of prior codes or journals (blue). The theoretical bound is the number of possible pair-wise combinations given the number of elements, including self-pairs. It shows how big the search space is for discovering combinations of knowledge given the finite size of knowledge stock. Theoretical bound (TB) is computed through Equation 3.3:

$$TB = \frac{n \times (n+1)}{2} \tag{3.3}$$

where n denotes the cumulative number of knowledge units, in other words, the number of journals referenced by AI publications or the number of technology codes assigned to AI patents.

The three time-series for AI patents (solid lines in Panel C of Figure 3.19) and AI publications (dash lines) have progressed at very similar growth rates since the middle 1980s, implying that inventive activities in AI may share a common underlying generating process for both scientific inquiry and industrial patenting. The theoretical bound is constrained by the cardinality of the known nodes, journals, or technology codes, in the search space (solid and dashed green lines); the search space will grow as long as new nodes are introduced. Thus the theoretical bound represents the growth of the current search space of combinations (solid and dashed blue lines). The cumulative counts (solid and dashed orange lines) track how effectively researchers and inventors search that set of possible combinations. If the theoretical bound expands faster than the cumulative count, then the space of possibilities expands faster than the space can be searched. Conversely, if the lines are parallel or converging, researchers and inventors search the space at a proportional or faster rate

than the space is expanding. The plots demonstrate that although the search space's growth rate has slowed, the current search space of combinations is still expanding faster than it is being searched.

Panel D of Figure 3.19 show how AI knowledge combinations are growing with the knowledge quanta. As the cited journals in AI publications and technology codes in AI patents grow, the natural logarithms of journal pairs and code pairs increase almost linearly with the natural logarithms of the number of journals and codes. Nevertheless, the blue dots and orange dots in Panel D both appear to be trending with increasing growth rates, as demonstrated by the slight super-linearity at the end.

3.6 Discussion

AI patents have grown significantly faster than utility inventions in general since 1976. The most robust growth momentum can be observed in knowledge processing and planning, the two AI subdomains featuring symbolic representations of human knowledge. While the number of technical components involved in AI has also increased substantially, the technical fields, as recorded in the 4-digit technology codes, only grew slightly since the 2000s, if not remained stable, indicating that the growth of AI inventions is driven by the fragmentation of specific technological ingredients rather than the expansion to new technical fields.

AI patents increasingly source their novelty from refining existing technologies or combining existing technologies in new ways rather than introducing new capacities. Highly original inventions have become increasingly rare. Contrary to anticipation, inventions in many AI subdomains have a higher refinement and narrow invention rates than utility patents in general, indicating a low novelty. However, domains like machine learning and computer vision are exceptions and appear more novel than utility patents. Nevertheless, the refinement and narrow invention rates in AI

subdomains have declined since the early 2010s. It may imply that AI patents are becoming more novel in the recent decade.

Moreover, regarding the technical fields involved in the patents, AI inventions appear less diverse than utility patents in general. AI patents are increasingly concentrated in a few technical fields, such as electric digital data processing, graphical data reading, and speech analysis. By 2020, around 80% AI patents are engaging with technologies that belong to only 5% of the fields. It implies that AI inventions have not diffused sufficiently into other technical fields. It challenges the notion that AI is a general-purpose technology incorporated into almost every aspect of society. At least at this moment, AI has not become a general-purpose technology. The pattern observed in the past four decades (increasingly lower novelty and higher concentration) also indicates that AI may not quickly become such a general-purpose technology.

Regarding the disruption to the knowledge networks, AI patents display no significantly different pattern than utility patents in general. Like utility patents, AI patents are becoming less disruptive. Highly disruptive patents are rare. The highly-disruptive rate in AI inventions is even lower than in utility patents.

I find that AI patents that combine existing technologies in new ways tend to receive more citations. Nevertheless, AI patents that are highly original disrupt subsequent technological paths to a substantially greater extent. This suggests that novelty is associated with disruption, and incremental advancements are more likely to be built upon by future inventions with further incremental advancements.

I also investigated how AI patents rely on scientific knowledge created with the federal government's help. AI inventions rely more heavily on public support than inventions in general. Private corporations are the primary actors who utilize government-supported science to invent AI. Machine learning and computer vision are the two AI subdomains with the strongest reliance on public support. It has been reported

previously that government-supported patents appear to receive more citations than others. Further evidence provided in this chapter has shown that such an observation also holds in AI patents. Nevertheless, AI patents created without public support appear more disruptive than those with public support. However, public-supported AI patents have a much higher chance of becoming top-ranked in disruption.

Recent studies in the science of science have provided evidence supportive of the arguments that large teams develop and that small teams disrupt science and technology. I found this argument holds in the domain of AI. AI inventions created by larger teams tend to receive more citations. On the other hand, AI inventions produced by smaller teams, particularly those by lone inventors, although relatively low in citations, are more disruptive and more likely to become top-tier in disruptiveness.

Furthermore, this chapter found that highly original patents tend to be produced by lone inventors. Small teams also have a slightly higher chance of producing highly original patents. As team size grows, the level of novelty in the patents they create declines. A possible interpretation could be that because individual inventors' personal knowledge stocks may not completely overlap, more knowledge could be brought to the invention as team size grows. Appendix N shows that larger teams have a higher chance of producing patents that involve more technical fields or components and vice versa. However, a larger previous knowledge space is not necessarily associated with more novel ideas. On the contrary, the more people and knowledge engaged in an invention, the more difficult it is to reach a consensus on formulating radically novel ideas because it is easier to pick from the vast options from existing knowledge stock and combine or integrate them.

Moreover, this chapter compared the results of AI patents and AI publications. Based on an examination of publications and patents,this chapter finds that both publishing and patenting output related to AI have grown exponentially and markedly

faster than scientific publishing and patenting in general. Exponential growth is observed in the number of new AI patents and AI publications created each year and in the cumulative scale of their knowledge stock and combinations. AI invention (as revealed through patenting) is found to be advancing at a faster pace than inventive activities in general. Overall, the rapid progress in both AI science and AI patenting primarily depends on the conventional combination of existing knowledge. This is evident from the trend that AI invention is increasingly composed of existing knowledge combinations rather than the generation of entirely new ideas, methods, or technologies. This approach aligns with the historical evolution of science and long-standing patterns of technological change (Kuhn, 1962; Strumsky and Lobo, 2015a; Lobo and Strumsky, 2019; Ridley, 2020; Barham, 2013). I provide evidence for continuous growth in conventionality and a diminishing degree of novelty over the years, with a temporary exception in AI patents, possibly due to a change in software patenting policy. Nevertheless, radical novelty has been rare in either AI publications or AI patents, although it used to be the primary driving force in the early years of AI.

Despite these similarities, AI publications and AI patents differ in what knowledge is combined, how they are combined, and their source of novelty. AI publications often combine more pieces of distinct knowledge than AI patents, although knowledge stock in AI research is smaller in scale. In addition, the knowledge combined in AI publications tends to be more topically distant than that in AI patents, indicating combinations in AI research are broader, and those in AI patents are deeper. The findings suggest that AI research often engages in greater exploitation of existing knowledge and is less shy about connecting distant intellectual spaces. Furthermore, AI science tends to discover more new ways of combining existing knowledge, while AI patenting often finds its novelty rooted in refining conventional combinations.

Notwithstanding the transformative potential of Artificial Intelligence, the creation of new knowledge in the AI field is not significantly different from the history of other inventive activity. Despite the many promises that AI will transform the entire process of invention and innovation, AI invention is proceeding in an incremental and cumulative manner, exploiting existing knowledge and discovering new knowledge by combining existing ones in new ways.

Nevertheless, *novelty is better rewarded in AI patents than in AI publications*, demonstrated by the observations that top-cited AI patents have a higher percentage of patents that feature novel combinations and AI patents featuring novel combinations are more likely to become top-cited. At the same time, AI publications combining conventional knowledge pairs account for a higher percentage of top-cited groups of AI publications and have a higher hit rate in terms of citations. Such a discrepancy of preference in citations between AI patents and AI publications can be partly credited to the different purposes, intentions, and practices of citations between patenting and scientific publications. An invention requires a certain level of novelty that differentiates it from its prior art to be eligible for a patent. A patent cites a previous patent primarily to disclose the most closely-related technologies in previous patent record and to "inform the patent owner and the public in general that such patents or printed publications are in existence and should be considered when evaluating the validity of the patent claims" (USPTO, 2015a). The more a patent is cited, the more subsequent inventions are built upon or significantly resemble it. Therefore, a citation received by a patent is less a recognition of the invention's merit or a direct aid in creating a new invention. Rather, it signifies the patent's potential to inspire future improvements. It is not uncommon that an inventor is not aware of, let alone that they are inspired by, the patent cited as a prior art, particularly in cases where the citation is made by the patent examiners. Therefore, a patent that

opens up new technological paths and thus is novel is likely to be cited. On the other hand, a citation made by a publication to a previous publication often indicates the latter is recognized as impactful to the production of the new knowledge contained in the new publication, or at least it is worth mentioning or acknowledging due to its merit. An article, of course, requires a certain level of novelty to be appreciated by the reviewers during the publication process. However, the novelty required for a "publishability" does not necessarily pertain to the same patenting standards. A journal often publishes different types of articles, including but not limited to original research papers, research notes, discussion papers, analyses, reviews, and perspective articles. Different reviewing and editing processes apply to different types of articles. Different journals may have different criteria regarding the number and format of citations. Therefore, there is an incentive for authors to cite specific journals or articles to ensure conformity to accepted norms of the research community of the fields — to "cite inside the box." Accordingly, this discrepancy between AI patents and AI publications can also be considered generalizable to patents and publications in general, regardless of domain.

The present study has several limitations. Firstly, the scope of the AI publication dataset is limited by the keywords searched on the Web of Science and includes only publications that either are in English or have their metadata translated into English. AI research in non-English-speaking countries has been increasing, most notably in China. Although the dataset serves as a representative sample of AI publications worldwide, it can be improved by incorporating additional keywords, such as those that capture the latest emerging progress in searching and including publications in languages other than English. Secondly, the typicality and novelty of AI invention are assessed solely within the domain of AI. However, a knowledge combination considered typical in AI can be considered atypical in other domains and vice versa. Therefore,

the findings of this study can be broadened by implementing the same methods in other inventive domains and comparing them with AI.

The findings of this study advance our understanding of the generative processes driving the development of a field proclaimed as having the potential to transform technological and societal development. Perhaps, the way novelty has been generated in AI is very much consistent with how inventive novelty has been created in many other fields and over the history of human invention. Therefore, it is not foolhardy to expect AI's continued development to conform to how other technological domains have matured. This understanding can inform and temper expectations regarding what to expect in the continued progress of Artificial Intelligence.

Chapter 4

AN ANALYTICAL FRAMEWORK TOWARDS BETTER UNDERSTANDING OF FIRMS' COMBINATORIAL INVENTION IN KNOWLEDGE ACCESS AND KNOWLEDGE DISCOVERY

4.1 Introduction

Inventions have been conceptualized as generating new and useful combinations of existing knowledge. The search for combinatorial components is often classified into local and distant searches. There are organizational and industrial aspects to such local and distant search decisions. Nevertheless, it is unclear how organizational and industrial boundaries affect those searches and how knowledge is combined. Furthermore, how the distinction between exploitation and exploration is represented and reflected in such organizational combinatorial inventions is inadequately documented nor quantified across these domains. To bridge those epistemological and empirical gaps, this chapter develops an analytical framework — the **C**ombinatorial **E**xploitation and **E**xploration (CEE) framework. The CEE framework provides mathematical formulations and quantifies organizations' combinatorial exploitation and exploration in three dimensions — knowledge access, inventive outcome, and knowledge discovery. The CEE framework measures organizational inventive activities with eight variables that assess the extent to which organizations have exhausted their possible combinations rather than merely counting the number of patents. This chapter explores the application of the CEE framework through a case study using patent records in artificial intelligence (AI) granted by the USPTO. The case study addresses how the CEE framework can be used and visualized to improve our understanding

of organizational inventions and industrial development and inform decision-making. Through this application of the framework, this chapter found that AI organizations tend to prefer exploitative over exploratory invention. The CEE parameters clearly illustrate the preferences of different organizations over time. The CEE framework represents one of the first attempts to conceptualize and quantify organizational exploitative and exploratory inventions. It opens up opportunities for future research agendas.

Maintaining an appropriate balance between exploitation and exploration has long been considered critical to the long-term inventive success of organizations. How are exploitation and exploration reflected in the combinatorial search for new inventions? Specifically, how do researchers in an organizational setting combine existing knowledge to create new inventions given the exploration-exploitation trade-off? A better understanding of *organizational search* would help researchers assess organizations' inventive activities and choices and inform policymakers and investors' decision-making.

Technological inventions have been conceptualized as combinations of existing technologies (Becker, 1982; Edquist and Johnson, 1997; Youn et al., 2015; Uzzi et al., 2013). The process of creating new knowledge can be considered as searching in the space of possible combinations to discover new and valuable combinations (Wagner and Rosen, 2014; Kauffman et al., 2000; Maynard, 2020). New knowledge can be created *de novo*, but this is considered very infrequent in the history of science and technology (Arthur, 2009; Wootton, 2015).

Nevertheless, the mechanism of combinatorial invention occurring in an organizational setting is not well elucidated, nor is the phenomena of search as an organizational phenomenon well documented. Despite rich bodies of literature that delve into the exploitation and exploration of organizational inventions and treat invention as a search for combinatorial possibilities, very few studies have integrated the two

perspectives. The theoretical and empirical gaps have resulted in an inadequate understanding of *combinatorial organizational search* and an insufficient measurement toolkit for quantifying organizational inventions beyond merely counting the number of outputs (such as publications or patents). Addressing such gaps would be helpful for researchers in evolutionary economics, industrial innovation, strategic management, and policymakers in organizations and governments.

The present discussion aims to connect two of the most prevalent perspectives on technological search by developing an analytical framework — **C**ombinatorial **E**xploitation and **E**xploration (CEE) Framework — that measures organizations' knowledge access, inventive outcome, and knowledge discovery in the context of combinatorial inventions. The proposed CEE framework distinguishes two types of knowledge boundaries: the organizational (or local) level and the industrial (or global) level. Measurements of inventive activities provided by the CEE framework are constructed beyond merely counting the number of certain types of patents. The framework considers two measurement issues — to what extent an organization has exhausted its possible knowledge combinations and how combinatorial search by different organizations can be compared over time.

To illustrate the usefulness of the proposed CEE framework and what novel insights it can offer, it is applied in the context of inventions in artificial intelligence (AI), a highly innovative and diverse technical field. Preliminary findings are presented regarding the inventive preferences of several organizations engaged in AI patenting. Specifically, I quantified the degree to which each AI organization has explored its potential combinatorial space. Intertemporal and inter-organizational comparisons are drawn. The case study in AI patenting shows that AI organizations engage in exploitative search much more than exploratory search. AI organizations tend to exploit local knowledge more than the global knowledge pool. The CEE framework is

also found to be able to illustrate individual organizations' long-term inventive preferences — the exploitation of local knowledge often slows down as an organization remains active in inventing. In contrast, exploratory activities tend to grow gradually if not stabilized. Inventions that broaden the global knowledge boundary are the least produced. Organizations dedicated to scientific research (such as universities and research institutes) demonstrate stronger preferences for exploration than for-profit companies. In addition, the quantifiability of the CEE framework allows it to be adaptable for studying the geography of inventive activities. Several preliminary investigations are presented to show how organizations in different states in the U.S. and different countries worldwide differ in their tendencies for exploitative or exploratory inventions.

4.2 Research Question

I set out to address the following research questions.

1. Research Question 1: How can organizations' exploitative and exploratory search be measured to reflect the extent to which organizations have exhausted the space of their possible combinations of knowledge across organizational and industrial boundaries?

2. Research Question 2: How can the measures constructed in *RQ1* be visualized to better inform organizations' inventive competence and policy-making?

3. Research Question 3: How can the proposed framework be tested with empirical data? What insights can be obtained?

4.3 Literature Review

4.3.1 The Combinatorial Nature of Invention

Inventions have been considered as new re-combinations of existing product factors that drive economic development (Schumpeter, 1934; Becker, 1982; Edquist and Johnson, 1997). As Brian Arthur (2009) put it, "technologies are combinations," and novel technologies are constructed from existing ones. Evidence from the patent record and scientific publications have been reported to support "combinatorics" being a fundamental nature of invention and innovation and an essential source of novelty in technological change (Youn et al., 2015; Uzzi et al., 2013; Strumsky and Lobo, 2015a; Mukherjee et al., 2016).

Nevertheless, as the pool of technological knowledge grows, the number of possible combinations can increase exponentially. Among the sheer number of possible combinations, only the ones believed to be somewhat valuable would be created, adopted, and diffused (Maynard, 2020). Therefore, the process of inventions can be conceptualized as solving an optimal search problem over a space of possible combinations of technologies constrained by the resources available. This search process involves value judgment and cost assessment (Kauffman et al., 2000; Macready et al., 1996; Wagner and Rosen, 2014).

4.3.2 Exploitative and Exploratory Search in Organizational Inventions

James March (1991) documented two types of organizational search — "the *exploration* of new possibilities and the *exploitation* of old certainties" — between which an appropriately-maintained balance is essential for the survival and prosperity of an organization (Gong et al., 2021). March (1991) explained exploration as "things captured by terms such as search, variation, risk-taking, experimentation, play, flexi-

bility, discovery, innovation," and exploitation involves things like "refinement, choice, production, efficiency, selection, implementation, execution."

Local search or exploitative search, defined by March (1991), is the type of knowledge creation that strongly depends on and builds upon the organization's previous knowledge. Local search has been considered and empirically observed as one of the most fundamental mechanisms of organizational R&D (Nelson and Winter, 1982; March, 1991; Stuart and Podolny, 1996; Rosenkopf and Nerkar, 2001). Organizations' preference for local search over non-local search can be explained by organizational routines that function to produce similar responses to frequently encountered situations (Nelson and Winter, 1982). Furthermore, local search likely results in a higher success rate in domains where the organization has accumulated experience and prior knowledge (Stuart and Podolny, 1996). Nevertheless, exploratory search that advances beyond local search and brings in new knowledge across boundaries has been suggested to be essential for organizations' sustainable competitive advantage (Kogut and Zander, 1992; Henderson and Cockburn, 1994).

A rich body of literature was established, and a consensus has emerged that a balance between exploitative and exploratory search should be maintained to improve the performance of organizations (Benner and Tushman, 2003; McGrath, 2001; Gupta et al., 2006). Some scholars extend the original model, which is agent-based with either further simulation (Gong et al., 2021), theoretical development like ambidexterity (Andriopoulos and Lewis, 2009), or characteristics observed in the real world, such as network structures (Fang et al., 2010; Lazer and Friedman, 2007), interpersonal learning, and tacit knowledge (Miller et al., 2006). Others aim to apply such models in an empirical context, such as the IT industry and oil industry (Wang and Hsu, 2014; Dixon et al., 2007; Geiger and Makri, 2006).

There have been intentions to capture how organizations reconfigure their knowl-

edge bases through local and non-local searches. Kogut and Zander (1992) investigated how firms apply current and acquired knowledge across organizational boundaries. Henderson and Cockburn (1994) examined "architectural competence" that enables firms to acquire knowledge across organizational and disciplinary boundaries. Kauffman et al. (2000) formalized a quantitative notion of technology distance and simulated firms' strategy of searching for optimal improvement given their technology landscape. Rosenkopf and Nerkar (2001) introduced a typology that recognizes firms' local or non-local search tendencies. They defined two dimensions of boundaries — organizational and technological boundaries — upon which four types of search are formulated.

4.3.3 Knowledge Production Function

The classic knowledge production function in Economic studies can be summarized as a statement that researchers use the existing knowledge base to create new knowledge, and new knowledge, in turn, is incorporated into the knowledge base and serves as an ingredient for future discoveries (Romer, 1990; Jones, 1995; Agrawal et al., 2018). This function is illustrated in Figure 4.1 as the two solid boxes at the bottom — Existing Knowledge Base (A) (excluding the internal dotted box) and New Knowledge ($\dot{A}$) — and the two solid arrows between them. Mathematically, new knowledge can be expressed as a function of existing knowledge (Equation 4.1).

$$\dot{A} = f(A) \tag{4.1}$$

However, this knowledge production does not capture the combinatorial nature of inventions. Therefore, Agrawal et al. (2018) extended this model by adding an intermediary step that illustrates the possible new combinations of the existing knowledge base. This intermediate step is shown as the box at the top of Figure 4.1 with the

text "Potential Combinations (Z)" (excluding the four dotted internal boxes). This extended model is completed by linking "Potential Combinations" with the existing knowledge base and new knowledge using dashed arrows as well as the two new functions (Equation 4.2 and 4.3).

$$Z = g(A) \tag{4.2}$$

$$\dot{A} = h(Z) \tag{4.3}$$

The knowledge production model of Agrawal et al. (2018) recognizes the fact that as the knowledge base grows, only a fraction of potential combinations is considered valuable and hence being explored and potentially incorporated into the future knowledge base, allowing for further investigations into the extent to which a society has exhausted its potential combinations and the pace the space of potential combinations is traversed.

In their model, Agrawal et al. (2018) further defined a knowledge access parameter ϕ to control the amount of knowledge each individual researcher has access to, which is denoted as A^{ϕ} where A represents the total stock of knowledge of the society (presumably $0 < \phi < 1$). Then, the total number of potential combinations of a given individual research i can be computed as:

$$Z_i = \sum_{a=0}^{A^{\phi}} \binom{A^{\phi}}{a} = 2^{A^{\phi}} \tag{4.4}$$

where a denotes the number of components combined at a time and $a = 0, 1 \ldots A^{\phi}$. Agrawal et al. (2018) also extended the model by considering research teams in addition to individual researchers.

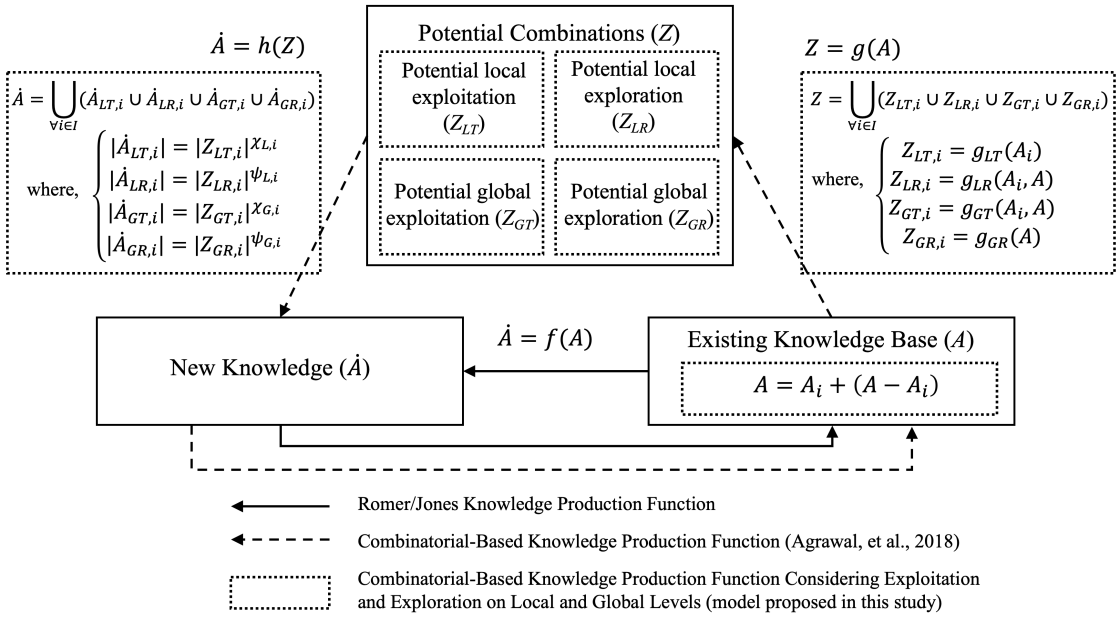


Figure 4.1: Combinatorial knowledge production models. Solid boxes and arrows represent the classic Romer/Jones knowledge production function. Dashed arrows represent the components extended by Agrawal et al. (2018) with the consideration of potential combinations of existing knowledge. The model developed in this chapter (**C**ombinatorial **E**xploitation and **E**xploration, or CEE framework) further extends the previous models with the parts encapsulated in dotted boxes. The CEE framework extends the existing knowledge production function by considering two levels of knowledge boundaries (organizational or local boundaries and industrial or global boundaries) and two types of knowledge discovery (exploitation and exploration).

4.4 Framework Development

To further our understanding of organizational knowledge access and knowledge discovery and characterize organizational exploitative and exploratory invention to a finer granularity, an analytical framework that extends existing models of knowledge production functions is proposed. Our proposed framework is named CEE framework, acronym for "**C**ombinatorial **E**xploitation and **E**xploration."

The CEE framework complements previous knowledge production models in the following ways, and they are illustrated in Figure 4.1 with dotted boxes. First, this research distinguishes two levels of boundaries for "existing knowledge base" — organizational or local knowledge base and global knowledge base, which are laid out

in the dotted box inside the solid box at the lower right corner with the text "Existing Knowledge Base (A)"). Depending on the analysis scale, global knowledge could refer to the knowledge pool of an industry, a region, or a country. Secondly, the CEE framework categorizes "Potential Combinations" into four types according to the boundary (local or global) where the combinatorial components locate. The four types of potential combinations are shown in the four dotted boxes inside the solid box at the top with the text "Potential Combinations (Z)." How the four types of potential combinations are computed and how they relate to previous models can be found in the dotted boxes at the right of Figure 4.1. Thirdly, the CEE framework distinguishes four types of new knowledge discoveries described in the dotted box on the left in Figure 4.1.

Three domains are quantitatively measured by the CEE framework: knowledge access, inventive outcome, and knowledge discovery. A total of eight measurements for these three domains are constructed. They are described in detail in the following sections. A summary of the three domains and their eight measurements can be found in Table 4.1.

4.4.1 Organizational Knowledge Access

Patents, Technology Codes, and Code Pairs

Patents have been considered an indication of innovation and invention. Thus, patent data is often used to study innovation in organizations and societies, given that it contains the information of invention itself and the combinatorial characteristics of inventions (Youn et al., 2015; Strumsky and Lobo, 2015a; Lobo and Strumsky, 2019).

While a patent *per se* represents a stand-alone invention, it can be considered a combination of smaller technological components (Arthur, 2009). In the language of

patents, those components are often embodied in the technology codes assigned to each patent for classification purposes.

Patent offices around the globe have formulated several classification systems to identify the technical components found in each patent. Those classification systems usually consist of symbols or codes, each describing a specific technological component and collectively forming a nested, hierarchical structure. These symbols are referred to as "technology codes." One of the most widely-used classification systems is the Cooperative Patent Classification (CPC) system developed by the European Patent Office (EPO) and the USPTO, and it has been used by many major patent offices around the world (USPTO, 2021b). Each full-digit CPC code is a symbol containing several levels of hierarchical information, including section, class, group, and subgroup. For instance, the CPC code "G03H 1/0011" consists of section "G" (physics), class "03H" (holographic process or apparatus), main group "1/00" (holographic processes or apparatus using light, infra-red or ultra-violet waves for obtaining holograms or for obtaining an image from them), and subgroup "1/0011" (for security or authentication) (USPTO, 2022a). Therefore, collectively, the code "G03H 1/0011" describes a technical component of holographic process or apparatus using light for security or authentication.

Each patent is assigned at least one CPC code to specify its technical content. The first one in the code sequence of a patent or the primary code is selected to "most completely cover[s] the technical subject matter" (USPTO, 2015b). A patent has at least one mandatory "inventional" CPC code that describes novel technical component(s) disclosed in the patent. A patent could also have one or more optional "additional" CPC codes to describe the non-trivial components that are not necessarily novel (USPTO, 2015b). This study does not differentiate between inventional and additional codes, nor does it distinguish primary codes from the rest because they all

describe essential technical components included in patents.

Researchers often use technology codes to study the technical information of patents and their combinatorial characteristics because they represent the building blocks of each patent (Youn et al., 2015; Strumsky and Lobo, 2015a; Lobo and Strumsky, 2019). Building upon such reasoning, this chapter considers three levels of analysis "quanta" or units of knowledge recorded in patent record — the invention level that concerns each patent as a unit, the technological component level that recognizes each technology code (e.g., CPC code) as a unit, and the combinatorial level that considers each combination of technology codes as a unit (this study only considers the pair-wise combinations or combinations of two). A combination is formed when two technology codes are listed in the same patent.

Assume a patent p is associated with a set of technology code TC_p (which refers to "**T**echnology **C**ode"), then, the set of code combinations TCP_p (which refers to **T**echnology **C**ode **P**air) can be described with Equation 4.5:

$$\begin{aligned} TCP_p &= \binom{TC_p}{2} \\ &= \big\{\{a, b\} \mid a, b \in TC_p\big\} \end{aligned} \tag{4.5}$$

Note that a set is unordered, and each element is unique. Therefore, assumptions $\{a, b\} = \{b, a\}$ and $a \neq b$ apply to Equation 4.5. I assume a patent with only one code does not combine technologies and has zero code combinations.

Suppose we have a set of patents denoted using the capital letter P, and each patent is associated with a set of technology codes. Then, TC_P or the set of technology codes associated with the patent set P can be described as the union of all the technology code sets associated with all the patents in the set P (Equation 4.6):

$$TC_P = \bigcup_{\forall p \in P} TC_p \tag{4.6}$$

where TC_p denotes the set of technology codes associated with the patent p ($p \in P$). Similarly, the set of technology code pairs associated with set P can be expressed as:

$$TCP_P = \bigcup_{\forall p \in P} TCP_p \tag{4.7}$$

where TCP_p is described in Equation 4.5.

Three Levels of Knowledge Access

I consider an organization as the boundary that sets the limit for the knowledge freely accessible to researchers within the organization. That is to say, presumably, a researcher is considered to have full and unrestricted access to the knowledge contained in the patent pool owned by the organization.

Therefore, the three levels of knowledge unit described in Section 4.4.1 correspond to three levels of knowledge access for researchers in an organization — the access to patents, the access to technology codes, and the access to code pairs. The access to patents describes how many inventions a firm has been granted with patents by the patent office, indicating the accumulative *outcome* of its previous inventive efforts. The access to technology codes represents the technological *capacities* a firm has accumulated previously. Together with patent access, technology code access can tell us how broad an organization's technologies are. For instance, a firm with a hundred patents and only three technology codes has a very focused and relatively narrow direction when it comes to R&D. On the contrary, a firm that only has ten patents yet with a hundred technology codes noticeably has been mobilizing a relatively wide range of technical capabilities. In addition, the access to technology code pairs illustrates a firm's previous efforts of creating new knowledge through combinatorial invention. For example, let us assume two firms (A and B) with ten patents and ten technology codes. For firm A, each of its patents has only one code,

while for firm B, each of its ten patents is assigned five codes. By definition, the two firms would share the same degree of access to patents and technology codes. However, firm B has higher access to code pairs and combines technologies to a greater extent than firm A.

Although researchers have often utilized patent count as a proxy to describe organizations' inventive activities and competence, technology codes and their combinations are less explored regarding their explanatory power in describing and characterizing organizational invention. Nevertheless, the number of patents alone is not informative enough regarding how its inventive activities evolve compared to other firms in the same industry. It tells us nothing about how the organization combines existing knowledge to create new ones. It is necessary to redefine measurements that gauge organizations based on their previous capacities against their cohorts and capture the combinatorial behaviors in inventing.

Building upon the knowledge production model extended by Agrawal et al. (2018) where a *knowledge access parameter* of an individual ϕ is defined as the power of the total knowledge stock (as in $A_i = A^{\phi}$), three knowledge access parameters are defined in similar ways to measure the three levels of knowledge access. For the organization i in the year t, the three knowledge access parameters are described as follows:

- ϕ_P, or *patent access parameter* satisfies Equation 4.8, where $P_{i,t-1}$ denotes the set of previous patents granted to organization i up to but not including year t while P_{t-1} denoting the set of previous patent grants globally (or industry-wide, depending on the scale of the analysis) up to year t:

$$|P_{i,t-1}| = |P_{t-1}|^{\phi_{P,i,t}} \tag{4.8}$$

- ϕ_{TC}, or *technology-code access parameter* satisfies Equation 4.9, where $TC_{i,t-1}$ denotes the set of previous technology codes accumulated in organization i up

to year t (i.e., on the local scale) and TC_{t-1} denotes that on the global scale:

$$|TC_{i,t-1}| = |TC_{t-1}|^{\phi_{TC,i,t}} \tag{4.9}$$

- ϕ_{TCP}, or *technology-code-pair access parameter* satisfies Equation 4.10, where $TCP_{i,t-1}$ denotes the set of previous technology code pairs accumulated in organization i up to year t and TCP_{t-1} denotes the set of previous technology code pairs accumulated up to year t on the global scale:

$$|TCP_{i,t-1}| = |TCP_{t-1}|^{\phi_{TCP,i,t}} \tag{4.10}$$

Then, for an organization i in year t, its patent access parameter $\phi_{P,i,t}$, technology code access parameter $\phi_{TC,i,t}$, and code-pair access parameter $\phi_{TCP,i,t}$ can be computed using Equations 4.11, 4.12, and 4.13 respectively. The plus one transformation in logarithms is necessary to avoid the undefined $\log 0$.

$$\begin{aligned} \phi_{P,i,t} &= \frac{\ln\left(|P_{i,t-1}|+1\right)}{\ln\left(|P_{t-1}|+1\right)} \text{ , where} \\ P_{i,t-1} &= \bigcup_{n=1}^{t-1} \dot{P}_{i,n} \text{ , and} \\ P_{t-1} &= \bigcup_{m=1}^{t-1} \dot{P}_m = \bigcup_{m=1}^{t-1} \bigcup_{\forall i \in \dot{P}_m} \dot{P}_{i,m} \end{aligned} \tag{4.11}$$

where $\dot{P}_{i,n}$ denotes the set of new patents the organization i have been granted in the year n; $P_{i,t-1}$ denotes the set of patents the organization i has accumulated in its patent pool up through the year $t-1$, or up to but not including the year t, and $|P_{i,t-1}|$ denotes the cardinality of the set $P_{i,t-1}$, in other words, the number of accumulative patents of organization i up to year t. $\dot{P}_m$ denotes the set of new patent grants of the entire industry in the year m, P_{t-1} denoting the set of patents the industry has accumulated up through the year $t-1$. Note that the global scale does not necessarily

mean worldwide; it could mean different scales depending on the scope of analysis, for instance, an industry or a geographic area. $\dot{F}_m$ denotes the set of firms in the industry that have patent grants in year m.

Likewise, technology-code access parameter $\phi_{TC,i,t}$ can be computed as:

$$\begin{aligned}
\phi_{TC,i,t} &= \frac{\ln\left(|TC_{i,t-1}|+1\right)}{\ln\left(|TC_{t-1}|+1\right)} \text{ , where} \\
TC_{i,t-1} &= \bigcup_{n=1}^{t-1} \dot{TC}_{i,n} \text{ , and} \\
TC_{t-1} &= \bigcup_{m=1}^{t-1} \dot{TC}_m = \bigcup_{m=1}^{t-1} \bigcup_{\forall i \in \dot{F}_m} \dot{TC}_{i,m}
\end{aligned} \tag{4.12}$$

where $TC_{i,t-1}$ denotes the set of distinct technology codes the organization i has accumulated in its patent pool up through the year $t-1$, while TC_{t-1} denoting the accumulative set of technology codes on the global scale up to but not including the year t. Therefore, $|TC_{i,t-1}|$, or the cardinality of the set $TC_{i,t-1}$, denotes the number of distinct technology codes of i up to year t. At the same time, $\dot{TC}_{i,n}$ denotes the set of technology codes that are involved in $\dot{P}_{i,n}$, or the set of new patent grants of organization i in the year n, while $\dot{TC}_m$ representing the set of codes found in $\dot{P}_m$, which is the set of new patent grants of all firms in the year m.

Similarly, technology-code-pair access parameter $\phi_{TCP,i,t}$ can be computed as:

$$\begin{aligned}
\phi_{TCP,i,t} &= \frac{\ln\left(|TCP_{i,t-1}|+1\right)}{\ln\left(|TCP_{t-1}|+1\right)} \text{ , where} \\
TCP_{i,t-1} &= \bigcup_{n=1}^{t-1} \dot{TCP}_{i,n} \text{ , and} \\
TCP_{t-1} &= \bigcup_{m=1}^{t-1} \dot{TCP}_m = \bigcup_{m=1}^{t-1} \bigcup_{\forall i \in F} \dot{TCP}_{i,m}
\end{aligned} \tag{4.13}$$

where $TCP_{i,t-1}$ denotes the set of distinct pair-wise combinations of technology codes the organization i has accumulated in its patent pool up through the year $t-1$, while

TCP_{t-1} denoting the accumulative set of code pairs on the global scale up through the year $t-1$.

I note that the accumulative patent count up to a particular year equals the sum of annual new patent counts of previous years, as expressed in Equation 4.14:

$$|P_{i,t}| = |\bigcup_{n=1}^{t} \dot{P}_{i,n}| = \sum_{n=1}^{t} |\dot{P}_{i,n}| \tag{4.14}$$

where $\dot{P}_{i,n}$ denotes the set of new patents granted to organization i in year n.

However, the equation does not hold for technology codes and code pairs because a technology code or a code pair as a component of inventions rather than an invention itself may repeatedly appear in future inventions after it is introduced to an organization. In other words, the intersection of two sets of technology codes of the same organization in two different years is unlikely to be empty. Thus, the cardinality of the generalized union of these sets should be smaller than the sum of the cardinality of each set. This re-use of previous technologies is considered exploitation of existing knowledge, and it is a common practice and highly prevalent in invention (Lobo and Strumsky, 2019). The number of the accumulative technology codes or code pairs should be equal to or smaller than the sum of the numbers of annual technology codes or code pairs (as expressed in Equations 4.15 and 4.16 respectively).

$$\begin{aligned} |TC_{i,t}| &= |\bigcup_{n=1}^{t} \dot{TC}_{i,n}| \\ &= \sum_{n=1}^{t} |\dot{TC}_{i,n}| - \sum_{n=1}^{t-1} \sum_{m=n+1}^{t} |\dot{TC}_{i,n} \cap \dot{TC}_{i,m}| \\ &+ |\bigcap_{n=1}^{t} \dot{TC}_{i,n}| \leq \sum_{n=1}^{t} |\dot{TC}_{i,n}| \end{aligned} \tag{4.15}$$

where $\dot{TC}_{i,n}$ denotes the set of distinct technology codes assigned to firm i in year n. Similarly, Equation 4.16 can be inferred:

$$|TCP_{i,t}| = |\bigcup_{n=1}^{t} T\dot{C}P_{i,n}| \leq \sum_{n=1}^{t} |T\dot{C}P_{i,n}| \tag{4.16}$$

where $T\dot{C}P_{i,n}$ denotes the set of distinct pair-wise combinations of technology codes found in $\dot{TC}_{i,n}$.

All three knowledge access parameters should be smaller than one and greater than zero. Only in rare cases where the local knowledge set is equal to the global knowledge set can the knowledge access parameter be equal to one. Suppose an organization has no previous knowledge (patents), for instance, in the cases of start-up firms or an established firm that just entered a new market. In that case, the knowledge access parameters are equal to zero.

Distinguishing the three levels of knowledge access provides a more holistic description of an organization's previous knowledge stock beyond merely counting the number of patents. An organization with more patents does not necessarily have more technology codes or code pairs representing technological capacities. The three knowledge access parameters complement each other and could be helpful when assessing the depth and breadth of organizations' previous inventive activities.

4.4.2 Patent Grants as Inventive Outcome

The second domain of the CEE framework addresses how many new patents are granted to an organization during a given year. Although critical about merely using the number of new patents to assess an organization's innovation competence, this research recognizes that new patent grants are an informative indicator for inventive outcomes. Here, the patent grant parameter $\rho_{i,t}$ of firm i in year t is defined in Equation 4.17:

$$|\dot{P}_{i,t}| = |\dot{P}_t|^{\rho_{i,t}}, \text{ where } 0 \leq \rho_{i,t} \leq 1 \tag{4.17}$$

$\dot{P}_{i,t}$ again denotes the new patent grants of firm i in the year t, while $\dot{P}_t$ denotes the new patent grants of the entire industry in the year t. Then, $\rho_{i,t}$ can be computed as:

$$\rho_{i,t} = \frac{\ln\left(|\dot{P}_{i,t}| + 1\right)}{\ln\left(|\dot{P}_t| + 1\right)} \tag{4.18}$$

The patent grant parameter describes how many new patents a firm has been granted in a year compared to the whole industry, capturing the relative inventive outcome of an organization.

4.4.3 Organizational Knowledge Discovery

Inventive knowledge discovery can be conceptualized as a search process in the space of possible combinations to identify valuable and new knowledge. Nevertheless, this space of possible combinations is neither homogeneous nor isotropic to a given organization at a given time. A technology that locates in one organization's existing knowledge stock may be considered hard to reach for another firm even if they are in the same industry. For instance, "*dropout*," a regularization method for training deep neural networks, is described in a patent owned by Google (Hinton et al., 2016). Hence, as a technical component, dropout is within Google's local technological pool but not in IBM's knowledge stock. The resources (hardware, software, engineers, and license) needed to implement dropout and combine it with other components to create new inventions are more readily available for a Google researcher than a researcher at IBM.

The resources required to combine combinatorial components differ significantly depending on each component's location in the technical landscape. In the example of dropout, combining dropout with other technologies would be relatively more straightforward for a Google researcher than for an IBM researcher.

Therefore, it is necessary to distinguish the knowledge discovery based on where

the technical components are located and whether they are within or outside the local or global boundary. Our CEE framework categorizes technology code pairs formed by the patents granted to an organization during a period into four types — local exploitation, local exploration, global exploitation, and global exploration, as illustrated in Figure 4.2 and described as follows:

1. a local exploitation (LT) if both of the components exist in the organization's local knowledge pool;

2. a local exploration (LR) if one of the components exists in the local pool while the other is in the global pool;

3. a global exploitation (GT) if both of the components exist in the global pool but not in the local pool;

4. a global exploration (GR) if one of the components exists does not exist in the local nor global pool, meaning it is completely new.

Each of the four types of code pairs can be described by a knowledge discovery parameter that ranges between zero and one and that can be computed by considering to what extent the space of theoretically possible combinations is searched. The four parameters will be described in more detail in the following sections.

Local Exploitation: Utilizing What is Known

At any given time, the R&D team at an organization would likely exploit the previous inventions in its local knowledge stock and the technical components contained in those inventions. To combine locally existing components to create new inventions, in other words, to search in the space for possible combinations of local technical components, is relatively easy, given that the resources (for example, equipment,

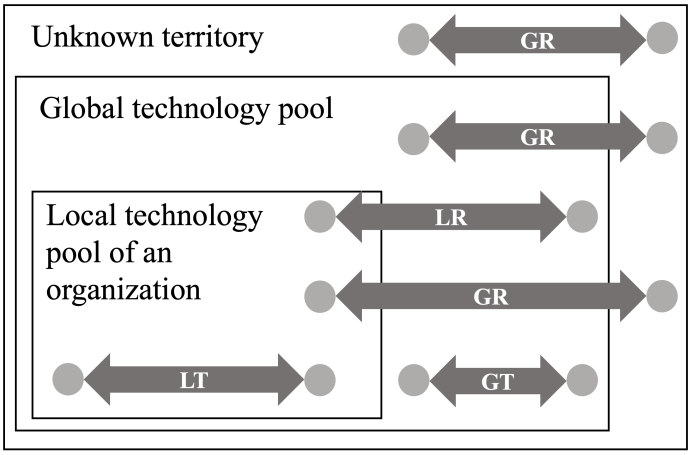


Figure 4.2: Four types of knowledge discovery — LT, LR, GT, and GR represent local exploitation, local exploration, global exploitation, and global exploration, respectively. A gray circle represents a piece of technology or a technology code in the context of patents. A double arrow represents a connection between two technologies formed through a co-existence in a patent.

human resources, and technical know-how) required for producing each component likely exist in the organization already and are readily available to the researchers. Much of the efforts would lie in estimating whether the possible combinations would create additional values.

The proposed CEE framework identifies such a knowledge creation process by searching the local knowledge pool as *local exploitation*, or LT for short. The new knowledge created through local exploitation in year t, or the set of technology code pairs made of previous knowledge in the local pool in the context of patents, is denoted as $T\dot{C}P_{LT,i,t}$.

Local exploitation can create new technologies upon existing ones by refining previously successful products or expanding existing functionality to adapt to the new market and business environment, improve product performance, and enhance customer satisfaction to maintain or expand current market share and profit.

Local exploitation is bounded by the theoretical upper limit of possible combinations of existing technology codes. Let $Z_{LT,i,t}$ denote the set of theoretically possible

combinations, then,

$$\begin{aligned} Z_{LT,i,t} &= \binom{TC_{i,t-1}}{2} \\ &= \big\{\{a,b\} \mid a,b \in TC_{i,t-1}\big\} \end{aligned} \tag{4.19}$$

where $\{a,b\}$ is an unordered pair or a set of two distinct elements.

Therefore, $T\dot{C}P_{LT,i,t}$ can be considered the intersection of the set of code pairs of organization i in year t and the set of its theoretically possible combinations of global exploitation:

$$T\dot{C}P_{LT,i,t} = T\dot{C}P_{i,t} \cap Z_{LT,i,t} \tag{4.20}$$

Then, the cardinality $|Z_{LT,i,t}|$ can be seen as the theoretical bound of the knowledge discovery through local exploitation. Because this research considers combinations of two, such a theoretical bound equals the number of possible pair-wise combinations between technology codes already existing in the local technology stock up to year t. $|Z_{LT,i,t}|$ can be computed using Equation 4.21:

$$\begin{aligned} |Z_{LT,i,t}| &= \binom{|TC_{i,t-1}|}{2} \\ &= \frac{1}{2} \cdot |TC_{i,t-1}| \cdot (|TC_{i,t-1}| - 1) \end{aligned} \tag{4.21}$$

where $TC_{i,t-1}$ denotes the set of previous local technology codes of organization i up through the year $t-1$.

I define $\chi_{L,i,t}$, or the *local exploitation parameter* of firm i in the year t to describe to what extent an organization i has exhausted the space of possible combinations (Equation 4.22):

$$|T\dot{C}P_{LT,i,t}| = |Z_{LT,i,t}|^{\chi_{L,i,t}} \tag{4.22}$$

Where $0 \leq \chi_{L,i,t} \leq 1$. Then, we have:

$$\chi_{L,i,t} = \frac{\ln\left(|T\dot{C}P_{LT,i,t}| + 1\right)}{\ln\left(|Z_{LT,i,t}| + 1\right)} \tag{4.23}$$

where $|Z_{LT,i,t}|$ can be computed using Equation 4.21 and $0 \leq \chi_{L,i,t} \leq 1$. (Again, the plus one transformation is necessary to avoid the undefined situation of $\log 0$.)

Local Exploration: "Polishing My Jade with Your Stones"

In addition to local exploitation, researchers at an organization would often keep track of the patents of their industrial cohorts, not only to maintain pace with state-of-the-art techniques and obtain inventive inspiration ("stones from other hills may serve to polish the jade of this one"[1]), but also to better anticipate future competitive products and market trends and help make middle- or long-term organizational decisions. A researcher well-informed about industrial-wide patenting activities is likely able to combine local components with technologies found in other organizations' knowledge pools. This type of knowledge discovery through combining a local component with a component found in another organization's patent pool is defined in the CEE framework as *local exploration*, or *LR* for short. The set of code combinations created from local exploration of firm i in year t is denoted as $T\dot{C}P_{LR,i,t}$.

Let $Z_{LR,i,t}$ denote the set of theoretically possible combinations of local exploration, then:

[1] "他山之石可以攻玉." This quote is from *Lesser Court Hymns, Shi-jing*, or *Classic of Poetry* (诗经), the oldest existing collection of Chinese poetry, dating from the 11th to 7th centuries BCE. The meaning of this quote can be understood as that other people's opinions can help us correct my shortcomings and improve ourselves. It is quoted here to mean that technologies from other organizations can be used to combine with ours to create new knowledge.

$$Z_{LR,i,t} = \{\{a, b\} \mid a \in TC_{i,t-1} \wedge b \in (TC_{t-1} - TC_{i,t-1})\} \tag{4.24}$$

It can be inferred that:

$$T\dot{C}P_{LR,i,t} = T\dot{C}P_{i,t} \cap Z_{LR,i,t} \tag{4.25}$$

Then the theoretical bound of local exploration $|Z_{LR,i,t}|$ can be computed as the product of the number of local codes and the global stock of technology codes excluding local codes. It can be expressed in Equation 4.26:

$$|Z_{LR,i,t}| = |TC_{i,t-1}| \cdot |TC_{t-1} - TC_{i,t-1}| \tag{4.26}$$

where $TC_{i,t-1}$ denotes the set of technology codes accumulated in organization i up to but not including the year t, and TC_{t-1} denotes the set of global technology codes in the industry up to the year t.

Therefore, a *local exploration parameter* $\psi_{L,i,t}$ can be defined as:

$$\psi_{L,i,t} = \frac{\ln(|T\dot{C}P_{LR,i,t}| + 1)}{\ln(|Z_{LR,i,t}| + 1)} \tag{4.27}$$

where $|Z_{LR,i,t}|$ can be computed using Equation 4.26 and $0 \leq \psi_{L,i,t} \leq 1$.

Global Exploitation: Inventing with Others' Knowledge

It is possible but often costs more than local exploitation or local exploration for a researcher to combine two pieces of existing knowledge, neither of which is in the local stock. I define this type of knowledge discovery as *global exploitation* or GT for short. The combinations in the new patents granted in the year t that satisfy global exploitation can be aggregated into a set denoted as $T\dot{C}P_{GT,i,t}$.

Let $Z_{GT,i,t}$ denote the set of theoretically possible combinations of global exploitation:

$$\begin{aligned} Z_{GT,i,t} &= \binom{TC_{t-1} - TC_{i,t-1}}{2} \\ &= \big\{\{a,b\} \mid a,b \in (TC_{t-1} - TC_{i,t-1})\big\} \end{aligned} \tag{4.28}$$

It can be inferred that:

$$T\dot{C}P_{GT,i,t} = T\dot{C}P_{i,t} \cap Z_{GT,i,t} \tag{4.29}$$

Then its cardinality or the theoretical bound of global exploitation can be computed as the number of possible pair-wise combinations of the difference between the global set and local set, as expressed in 4.30:

$$\begin{aligned} |Z_{GT,i,t}| &= \binom{|TC_{t-1} - TC_{i,t-1}|}{2} \\ &= \frac{1}{2} \cdot |TC_{t-1} - TC_{i,t-1}| \cdot (|TC_{t-1} - TC_{i,t-1}| - 1) \end{aligned} \tag{4.30}$$

Likewise, the *global exploitation parameter* of firm i in year t can be defined as:

$$\chi_{G,i,t} = \frac{\ln\left(T\dot{C}P_{GT,i,t} + 1\right)}{\ln\left(Z_{GT,i,t} + 1\right)} \tag{4.31}$$

where $0 \leq \chi_{G,i,t} \leq 1$.

Global Exploration: Reaching Uncharted Territory

The researchers at a firm may explore uncharted technical territories unknown to themselves or their cohorts in the industry (or beyond). As a result, a brand new technology may be combined with existing technology, or two new technologies may be combined. I define such a combination that contains globally new component(s) as *global exploration*.

Occasionally, global exploration may result from serendipity, such as the discovery of penicillin (Copeland, 2019). Notwithstanding, in most cases, global exploration is motivated and planned, requiring a considerable amount of effort and resources but may lead to expanding the knowledge boundary of the industry.

Let $Z_{GR,i,t}$ denote the set of theoretically possible combinations of global exploration of organization i in year t. In this model, $Z_{GR,i,t}$ cannot be derived directly from previous knowledge stock because it is largely impossible to predict what globally new knowledge will be created given existing knowledge. A workaround would be to retrospectively approximate such a boundary with all possible combinations containing at least one new technological component introduced in the year t (Equation 4.32). The set of globally new technology codes introduced in year t can be expressed as $TC_t - TC_{t-1}$ or $\dot{TC}_t - TC_{t-1}$.

$$Z_{GR,i,t} = \big\{\{a, b\} \mid a \in (TC_t - TC_{t-1}) \wedge \ b \in TC_t\big\} \tag{4.32}$$

Similarly,

$$\dot{TCP}_{GR,i,t} = \dot{TCP}_{i,t} \cap Z_{GR,i,t} \tag{4.33}$$

$Z_{GR,i,t}$ can be considered as the union of two disjoint sets, the first of which contains the combinations of one within the global technology pool while the other is a new technology. The second set consists of combinations, each of which both codes are new. Therefore, as an alternative to Equation 4.32, $Z_{GR,i,t}$ can also expressed as Equation 4.34,

$$Z_{GR,i,t} = \big\{\{a, b\} \mid a \in (TC_t - TC_{t-1}) \wedge \ b \in TC_{t-1}\big\} \cup \binom{TC_t - TC_{t-1}}{2} \tag{4.34}$$

Therefore, the theoretical bound of global exploration, or the cardinality of $Z_{GR,i,t}$ can be computed as the sum of the cardinalities of the two sets, as described in

Equation 4.35.

$$
\begin{aligned}
|Z_{GR,i,t}| &= |\{\{a,b\} \mid a \in (TC_t - TC_{t-1}) \wedge \ b \in TC_{t-1}\}| + \binom{|TC_t - TC_{t-1}|}{2} \\
&= |TC_{t-1}| \cdot |TC_t - TC_{t-1}| + \frac{1}{2} \cdot |TC_t - TC_{t-1}| \cdot (|TC_t - TC_{t-1}| - 1)
\end{aligned}
\tag{4.35}
$$

It is worth noting that because in the model, the scope of global exploration is irrelevant to the boundary of local knowledge stock, the theoretical bound of global exploration should be the same for every firm in the industry (Equation 4.36).

$$
Z_{GR,i,t} = Z_{GR,j,t} \quad (\forall\ i, j \in F) \tag{4.36}
$$

Therefore, the *global exploration parameter* can be computed using Equation 4.37.

$$
\psi_{G,i,t} = \frac{\ln\left(|\dot{TCP}_{GR,i,t}| + 1\right)}{\ln\left(|Z_{GR,i,t}| + 1\right)} \tag{4.37}
$$

where $0 \leq \psi_{G,i,t} \leq 1$.

4.4.4 Characterizing Organizations' Inventive Activity

To recap, this chapter has defined ϕ_P, ϕ_{TC}, and ϕ_{TCP} as the three knowledge access parameters describing an organization's accumulative stock of three levels of knowledge units — patents, technology codes, and technology code pairs. In addition, this chapter has defined ρ as the patent grant parameter to describe organizations' inventive outcomes of a given year. As for knowledge discovery, depending on where the combinatorial components locate, two analytical dimensions (exploitation vs. exploration, local vs. global) have helped identify four different types of knowledge discovery — local exploitation $\dot{TCP}_{LT}$ of which the cardinality is bounded by $|Z_{LT}|$, local exploration $\dot{TCP}_{LR}$ bounded by $|Z_{LR}|$, global exploitation $\dot{TCP}_{GT}$ bounded by $|Z_{GT}|$, and global exploration $\dot{TCP}_{GR}$ bounded by $|Z_{GR}|$. A knowledge discovery

parameter is defined for each of the four types of knowledge discovery — local exploitation parameter χ_L, local exploration parameter ψ_L, global exploitation parameter χ_G, and global exploration parameter ψ_G. A summary of the eight parameters of the proposed framework and their symbols and definitions can be found in Table 4.1.

Table 4.1: Three Domains and Eight Measurements, and Their Definitions and Equations Defined by the CEE Framework.

Domains	**Measurement**	**Symbol**	**Eq.**	**What does it measure?**
Knowledge access	Patent access parameter	$\phi_{P,i,t}$	4.11	Patents an organization i has accumulated previously ($P_{i,t-1}$) normalized by that of the whole industry (P_{t-1}).
	Technology code access parameter	$\phi_{TC,i,t}$	4.12	Technology codes an organization i has accumulated previously ($TC_{i,t-1}$) normalized by that of the whole industry (TC_{t-1}).
	Technology code-pair access parameter	$\phi_{TCP,i,t}$	4.13	Technology code-pairs an organization i has accumulated previously ($TCP_{i,t-1}$) normalized by that of the whole industry (TCP_{t-1}).
Inventive outcome	Patent grant parameter	$\rho_{i,t}$	4.18	New patent grants of an organization i in a certain year t ($\dot{P}_{i,t}$) normalized by all the new patent grants of the industry ($\dot{P}_t$).
Knowledge discovery	Local exploitation parameter	χ_L	4.23	Combinations of which both components exist in local pool ($\dot{TCP}_{LR,i,t}$) normalized by the theoretical bound ($Z_{LT,i,t}$ defined in Eq. 4.21).
	Local exploration parameter	ψ_L	4.27	Combinations of which one is in local pool, one in non-local global pool ($\dot{TCP}_{LR,i,t}$) normalized by the theoretical bound ($Z_{LR,i,t}$ defined in Eq. 4.26).
	Global exploitation parameter	χ_G	4.31	Combinations of which both components exist in non-local global pool ($\dot{TCP}_{GT,i,t}$) normalized by the theoretical bound ($Z_{GT,i,t}$ defined in Eq. 4.30).
	Global exploration parameter	ψ_G	4.37	Combinations that contain at least one globally new technology ($\dot{TCP}_{GR,i,t}$) normalized by the theoretical bound ($Z_{GR,i,t}$ defined in Eq. 4.35).

For organization i in the year t, the set of possible combinations $Z_{i,t}$ can be expressed as the union of the four types of possible combinations (Equation 4.38),

which are disjoint.

$$Z_{i,t} = Z_{LT,i,t} \cup Z_{LR,i,t} \cup Z_{GT,i,t} \cup Z_{GR,i,t} \tag{4.38}$$

Meanwhile, the set of combinations that are actually created in year t can also be seen as the union of the four types of realized combinations (Equation 4.39), which also should be disjoint.

$$\dot{TC}P_{i,t} = \dot{TC}P_{LT,i,t} \cup \dot{TC}P_{LR,i,t} \cup \dot{TC}P_{GT,i,t} \cup \dot{TC}P_{GR,i,t} \tag{4.39}$$

To illustrate an organization's inventive behavior, to compare multiple organizations in the same year, or to compare an organization's behavior in different years, a type of visualization that can show multiple variables simultaneously would greatly help convey the information obtained via the CEE framework. A multi-dimensional radar chart (also known as the spider chart) with eight axes can be a great option. Radar charts have been commonly used for visualizing multivariate data in two-dimensional space. They can overcome the problems many traditional graphic methods have faced in displaying more than three variables (Porter and Niksiar, 2018; He et al., 2021). Section 4.5 will demonstrate how to generate such radar charts to convey the eight CEE parameters and how they can help illustrate organizations' inventive behavior and preferences.

4.5 An Empirical Case: How AI Organizations Create Combinatorial Inventions

To test and validate the applicability of the proposed CEE framework and demonstrate how it can help us understand organizations' inventive behavior and inform decision-making, the CEE framework is applied to patents and organizations in artificial intelligence (AI), a field that has been developing with considerable momentum

and that has produced significant inventive outcomes. I selected AI to test the framework not only because many AI inventions have been created but also because of the rich diversity of the understandings, philosophies, and schools about what AI is, whether it should be semantic or connectionist or hybrid, what the future direction of AI should be, how it can become more robust, and how it can generate profits without harming the society. The continuous investment from private and public sectors and a lucrative market prospect have attracted many organizations, old and new, big and small, in AI patenting. I believe these aspects of AI welcome innovation and have created a testbed for various innovation studies.

AI can be considered a collection of techniques to learn insights from data and make decisions automatically. Four general approaches of AI can be summarized — thinking humanly, thinking rationally, acting humanly, and acting rationally, each involving distinct philosophies and different techniques (Russell and Norvig, 2010). The term AI was coined at the Dartmouth Conference in 1956, although its origin can be traced back to earlier attempts such as the Turing Test (McCarthy et al., 1955; Turing, 1950). AI in the 1960s and 1970s was featured with symbolic approaches that attempt to mimic humans' logic reasoning (Newell and Simon, 1976; Haugeland, 1985). In the 1980s, knowledge-based expert systems began to show promise in solving real-world problems. They had attracted significant funding until their failure to achieve satisfactory performance led to a deep disappointment followed by a period referred to as the "AI winter" in the late 1980s and early 1990s (Kaplan, 2016; Russell and Norvig, 2010). In the 2010s, connectionism, an AI school rooted in pioneering experiments by Frank Rosenblatt at Cornell University in the late 1950s, gained fresh momentum because of the new development of computing hardware, better algorithms, and the availability of large training data (Bishop, 2006). Modern connectionist machine learning relies on artificial neural networks to learn insights

from training data by minimizing loss functions. The neural networks with many layers are called deep learning, and it has surpassed humans in many tasks, such as recognizing faces, and playing Go (Lu and Tang, 2014; LeCun et al., 2015; Silver et al., 2016).

Throughout its history, AI has been drawing knowledge and insight from various areas, such as mathematics, cybernetics, psychology, biology, and cognitive science. Meanwhile, a rich diversity of ideas and application scenarios can be observed in AI. AI has been considered a "new method for invention" that would help generate knowledge in many fields and a potential general-purpose technology expected to affect every aspect of society (Agrawal et al., 2018; Cockburn et al., 2019; Brundage, 2019).

4.5.1 Data Collection

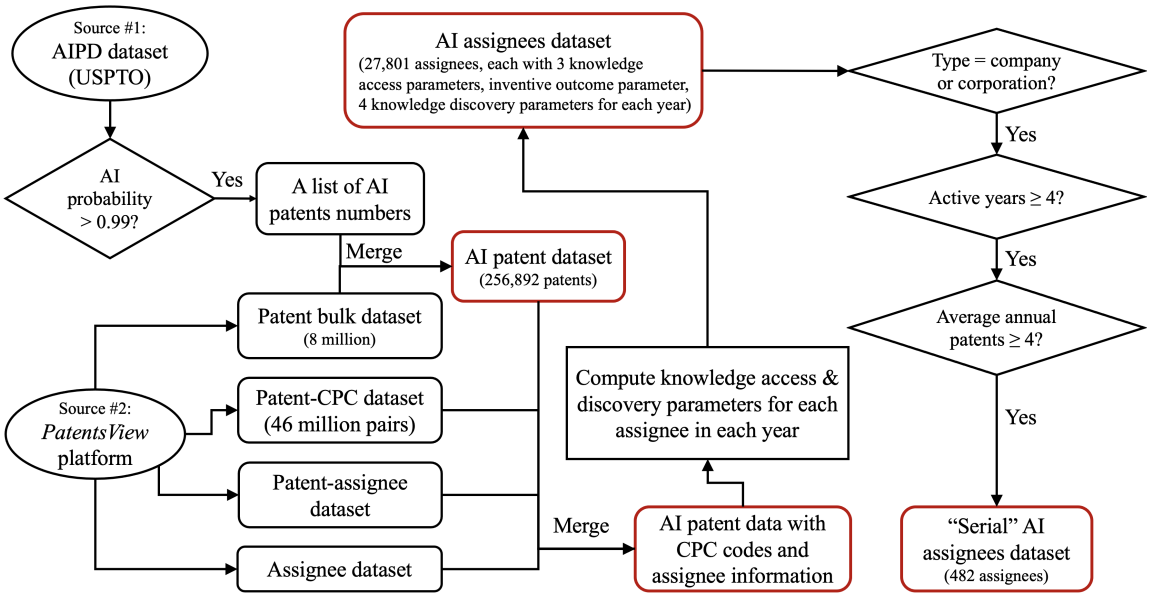


Figure 4.3: The process of constructing the AI patent dataset and serial AI assignee dataset used in this study.

I constructed a dataset of AI patents using two sources — AI Patent Dataset

(AIPD) released by the USPTO in 2021 (Giczy et al., 2021) and *PatentsView* open data platform (Toole et al., 2021), an initiative supported by the Office of the Chief Economist at the USPTO and that offers visualization, analysis, download, and other utilization of patent data of USPTO (Toole et al., 2021; USPTO, 2022b). The AIPD dataset was utilized to identify the patent numbers of AI patents, while *PatentsView* was utilized for collecting bibliographic information of patents and information about their assignees. An assignee is a person or organization that receives patent ownership rights. The construction of the dataset was described as follows (see Figure 4.3).

First, the AIPD dataset was used to identify a set of patent numbers of AI patents. In the AIPD dataset, each patent was assigned eight scores that ranged between zero and one by a machine learning algorithm to predict how likely the patent can be considered to belong to one of the eight AI subdomains, including machine learning, natural language processing, and computer vision. In their subsequent research, researchers at the Office of the Chief Economist of USPTO identified a patent as an AI patent if at least one of its eight scores is higher than 0.5. That is to say, the probability of the patent being an AI patent is higher than 50% (Toole et al., 2020). This study sets the threshold of AI "likelihood" as 0.99 instead of 0.5. In other words, this dissertation identifies a patent as an AI patent if at least one of its eight scores is higher than 0.99. In total, 256,892 AI patents were identified.

Then, the bibliographic information of all patents granted from 1976 to 2021 was bulk downloaded from *PatentsView*[2]. Such data contains nearly eight million patents (including 7.2 million utility patents) and detailed information, such as patent number, patent type (e.g., utility or design patent), date, abstract, title, and the number of claims. Then, the patents corresponding to the patent numbers in the AI patents dataset compiled in the previous step were selected. In addition, the patent-assignee

[2]`https://patentsview.org/download/data-download-tables`

and assignee datasets were downloaded from *PatentsView*, the former of which provides the unique IDs of assignees associated with each patent. At the same time, the latter matches the unique ID of each assignee with detailed information, including the categorization, country, organization names, and others. The unique IDs of assignees were generated by the *PatentsView* team through a disambiguation method (Monath et al., 2021). The categories of assignees include U.S. individuals, foreign individuals, U.S. companies or corporations, foreign companies or corporations, U.S. governments, foreign governments, and others.

4.5.2 Data Description

The AI patent dataset constructed as described in Section 4.5.1 contains 256,892 AI patents, their bibliographic information, including technology codes, and assignees' information, such as names, organizations, and cities, if applicable. In total, 27,801 assignees have owned at least one AI patent from 1976 to 2020, of which 17,677 (64%) are U.S. companies or corporations, while 8,911 (32%) foreign companies or corporations. In addition, there are 660 (2%) U.S. individuals and 318 (1%) foreign individuals. Government agencies in the U.S. and other countries also produce AI patents. The summary statistics of the main types of assignees can be found in Appendix Q.

Figure 4.4 shows the growing trend of the number of assignees ($\dot{F}_t$), patents ($\dot{P}_t$), technology codes ($\dot{TC}_t$), code pairs ($\dot{TCP}_t$), and the theoretical bound of pairs of existing codes (Z_t) in AI from 1976 to 2020. Here Z_t can be considered the union of the sets of possible knowledge discovery (except global exploration) of all AI firms up to but not including the year t. It can be defined in Equation 4.40.

$$
\begin{aligned}
Z_t &= \binom{TC_{t-1}}{2} \\
&= \left\{ \{a, b\} \mid a, b \in TC_{t-1} \right\} \\
&= \bigcup_{n=1}^{t-1} \bigcup_{\forall i \in \dot{F}_n} \left(Z_{LT,i,n} \cup Z_{LR,i,n} \cup Z_{GT,i,n} \right)
\end{aligned}
\tag{4.40}
$$

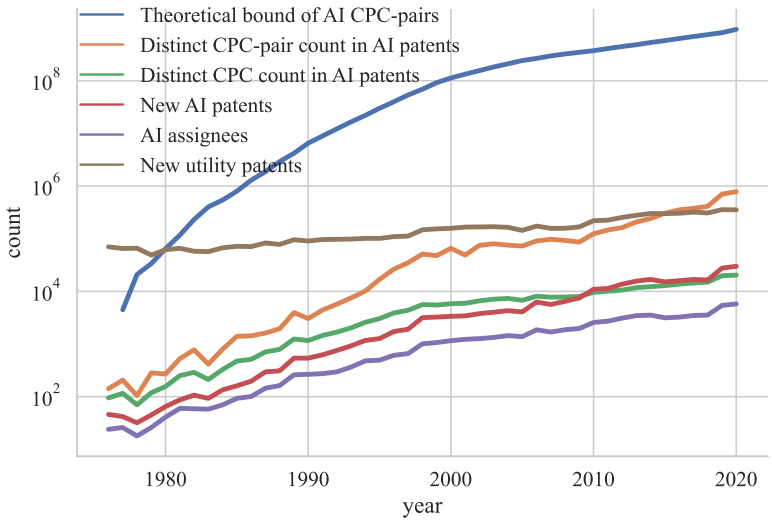


Figure 4.4: Time series (semi-log) of the number of assignees (purple), patents (red), CPC codes (green), CPC code-pairs (orange), and the theoretical bound of CPC code-pairs (blue) in AI. As a comparison, the brown line shows the time series of the number of new utility patents.

As shown in Figure 4.4, the numbers of assignees (purple line), patents (red line), technology codes (green line), and code pairs (orange line) have all been growing with significantly faster speed than utility patents in general, as shown as the brown line in the figure that is relatively flat. One salient trait illustrated in the figure was the slowing down of the growth of technology codes (green line). Since 2010, the number of distinct or different technology codes involved in new AI patents has been outnumbered by the number of AI patents. However, on average, each AI patent has almost 6 CPC codes (see Appendix R, the summary statistics of CPC codes assigned to each AI patent). It implies that new knowledge creation in AI in the recent decade

or so has been relying more heavily on combining existing technological components than introducing new ones.

By 2020, AI patents have included 46 thousand distinct CPC subgroups (i.e., full-digit codes), covering 566 CPC groups. The most frequent CPC codes include G06Q30/02 (systems or methods of marketing research and analysis, survey, promotion, and advertising), G06Q10/10 (office automation), G06N20/00 (machine learning), G06F16/9535 (search customization based on user profiles and personalization), G06F16/951 (indexing and web crawling techniques). The CPC groups are one level higher than the CPC subgroups, the lowest or the most granular level in the CPC hierarchical scheme. The CPC groups that have observed the most AI patents include G06F (electric digital data processing), G06Q (data processing systems or methods specially adapted for administrative, commercial, financial, and related purposes), H04L (transmission of digital information), G06K (graphic data reading). A total of 2.1 million CPC pairs are identified in the AI patent dataset of this study. The most-combined pairs include the combination of G06N3/0454 and G06N3/08, which represent "neural network models using a combination of multiple neural nets" and "learning methods" respectively, referring to machine learning methods using artificial neural networks. Another frequently combined pair consists of code G06N20/00 (machine learning) and G06N7/005 (computing arrangements based on probabilistic networks), referring to methods of Bayesian machine learning.

On the assignee level, it is notable that, like many other previously-reported endeavors in science and technology, AI is a largely skewed domain (Albarrán et al., 2011; Wang et al., 2022b). In other words, most assignees have very few AI patents, while very few have been granted many AI patents. Our data shows that 16,212 (58%) assignees have only one AI patent. On the other hand, the top ten assignees[3]

[3]The top ten AI assignees recorded in the data include IBM, Microsoft, Google, Amazon, HP,

(only accounted for 0.04% of all AI assignees) collectively own 63,810 (25.5%) of all AI patents.

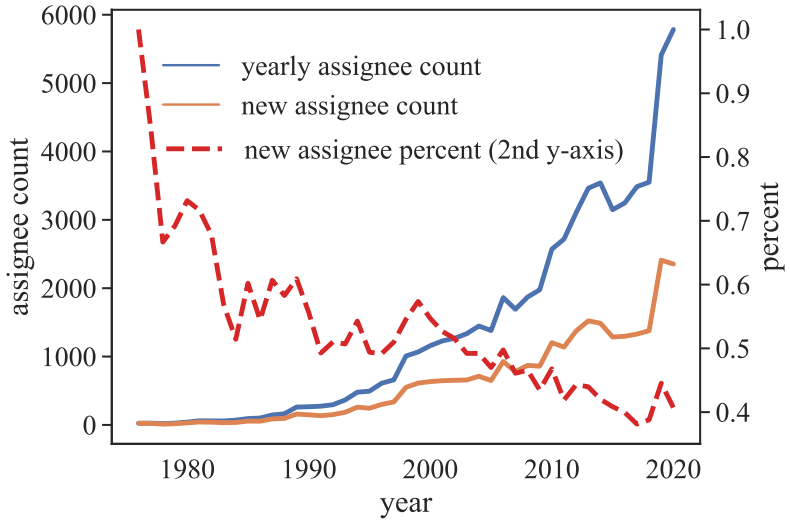


Figure 4.5: Time series of the number of assignees that are granted AI patents in each year (blue line, left y-axis), the number of new assignees (orange line, left y-axis), and the percent of new assignees in all assignees of the year (red dashed line, right y-axis). The same plots for utility patents in general and other technologies can be found in Appendix T.

The number of assignees granted AI patents has been continuously growing since 1976. The decline in assignees in the mid-2010s has likely resulted from the establishment of the Mayo/Alice test, a test practiced by the USPTO and the U.S. Court of Appeals for the Federal Circuit that, in effect had caused an increase in invalidation and rejection of software patent applications and after that that had been adjusted (Tran, 2016; Toole and Pairolero, 2020; USPTO, 2014). After this brief decline, assignees have been experiencing a sharp boost. Nevertheless, this rise in assignees appears to be driven by existing assignees that stay active rather than *new assignees* that have never patented in AI before. This is evidenced by the fact that the percentage of new assignees has been steadily decreasing, from 70% in the 1980s to around

Oracle, Samsung, Facebook, Intel, and Apple.

40% in 2020. However, compared to utility patents in general and several other technological domains that represent the frontier of science and technology (i.e., climate change mitigation and adaptation technology or green technology, biotechnology, and nanotechnology), AI has had the highest percentage of new assignees throughout the years (see Appendix T).[4] This implies that AI is still appealing to entrepreneurs more than other domains, either through attracting existing organizations to invest or establishing through new ventures.

4.5.3 Serial AI Assignees

In order to narrow down the dataset to a collection of organizations continuously investing considerable resources into AI development, assignees of companies or corporations are selected. It is worth noting that universities and research institutes are categorized as companies or corporations. I select "serial AI assignees" in subsequent analysis with the selected assignee dataset. A firm is considered a serial AI assignee if it satisfies the following two conditions. First, its active year is equal to or longer than four years. (The active year of an assignee is computed as the difference between its first year and last year in the dataset plus one.) Secondly, its average annual AI patent count is no less than four, computed as the total AI patent count of the assignee divided by its active year. The flow chart in Figure 4.3 shows the data collection and pre-processing.

The ultimate "serial AI assignee" dataset contains 482 organizations, including

[4]Green patents are identified by selecting the following CPC codes from utility patent dataset: B60L, H02S, H01M, F03D, Y02A, Y02B, Y02C, Y02D, Y02E, Y02P, Y02T, Y02W, Y04S. The CPC codes used to identify biotechnology patents include A01G, A01H, A61K, A61P, A61Q, B01F, B01J, B81B, B82B, B82Y, C05, C07, C08, C09, C11, C12, C13, C25, C40, G01N, and G16H (USPTO, 2019). Nanotechnology patents are identified using the CPC code B82Y.

332 U.S. and 150 foreign organizations. Regarding the types of organizations, there are 466 for-profit companies, of which 325 locate in the U.S. In addition, there are 16 universities or research institutes,[5] of which seven are in the U.S. The summary statistics of selected variables of AI assignees are shown in Appendix U. By 2020, half of serial AI assignees had more than 104 AI patents covering more than 264 distinct CPC codes and forming more than 1,656 code combinations. IBM owns the most AI patents (nearly 27 thousand). IBM also has AI patents involving the most CPC codes (118 thousand codes in total and nearly 8,000 distinct codes).

Serial AI Firms' Knowledge Access and Discovery

Subsequently, the eight CEE parameters for each assignee in each year are computed. The summary statistics of the eight parameters of serial AI assignees in 2020 can be found in Appendix W. The distributions of such eight parameters are shown in Figure 4.6. Appendix Y shows the joint distributions of each pair of the eight CEE parameters.

As a comparison, the distributions of the counts of patents, technology codes, or code pairs corresponding to the eight parameters can be seen from Appendix X.

It can be observed from Appendix X that the distributions of counts *per se* are incredibly skewed. Transforming the counts into the corresponding parameters (Figure 4.6) normalizes their distributions, allowing for more intuitive visual inspection and further detailed investigation. It is worth noting that local exploitation parameters are much more spread out and normal than the other three knowledge discovery parameters. In addition, comparing the mean and median values of the four types of knowledge discovery illustrates that in individual organizations' inventive activities,

[5]An organization is identified as a university or a research institute if its name contains any of the following terms — university, school, college, institute.

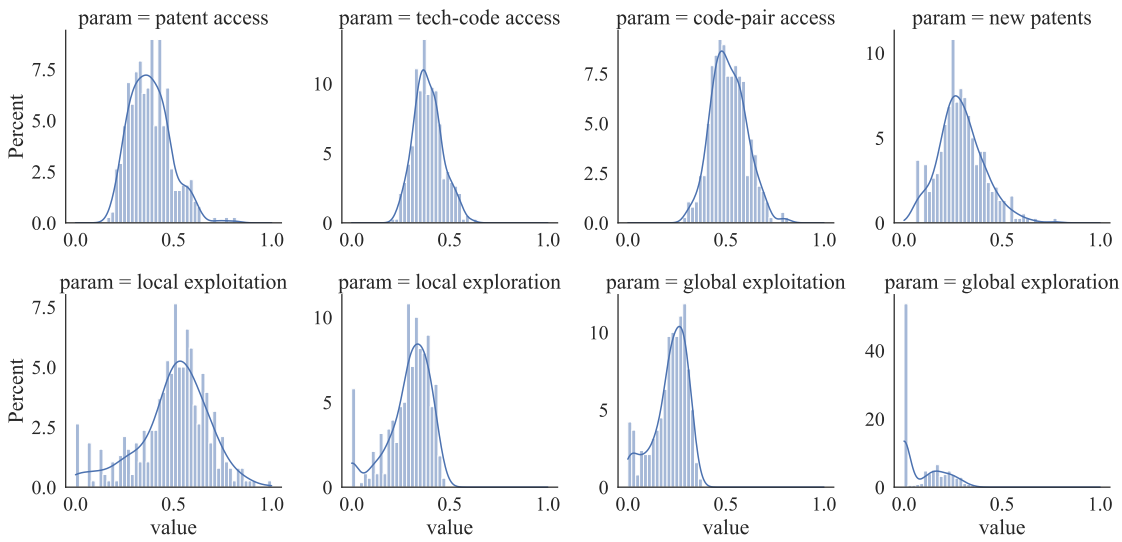


Figure 4.6: Distribution of the three knowledge access parameters, inventive outcome (new patents) parameter, and four knowledge discovery parameters of "serial" AI firms in 2020.

local exploitation tends to be the most extensive among the four. It is followed by local exploration and global exploitation. Global exploration is the inventive path that is less traveled, demonstrated not only by the mean and median values of global exploration parameters being the smallest among the four types of knowledge discovery, but also by the significant proportion of organizations (about 54% serial AI assignees in 2020) having their global exploration parameters equal to zero. That is to say, more than half of AI firms do not engage in any new technological components previously unknown to the field. This implies that exploiting the local technology pool is the most common practice in AI development, while exploring uncharted territory is rare, although not entirely absent.

Exploitative inventions often cost less than exploratory ones, especially in the short term (Greve, 2007). Despite researchers' suggestions that a balance between exploitation and exploration must be maintained to improve the long-term performance of organizations, it is not too hasty to infer that organizations that aim to

maximize profit while reducing cost would likely demonstrate a stronger preference for exploitation over exploration. Therefore, the extent to which organizations exhaust exploitative possibilities is likely more remarkable than their pursuit of exploration. I found that CEE parameters of AI organizations are consistent with such an inference — among the four knowledge discovery parameters, the local exploitation parameter tends to be larger than local exploration, which is greater than global exploitation, which is greater than global exploration (see the summary statistics in Appendix W).

The time series of the eight CEE parameters of serial AI assignees is shown in Panel A of Figure 4.7. Panel B of Figure 4.7 shows the relationship between the eight parameters and how long an assignee has been in the AI industry, in other words, the year since the first year each assignee started AI patenting. The two panels' solid lines represent the corresponding parameters' mean values. The color bands represent a 95% confidence interval. The dashed red lines in the two panels show the number of assignees.

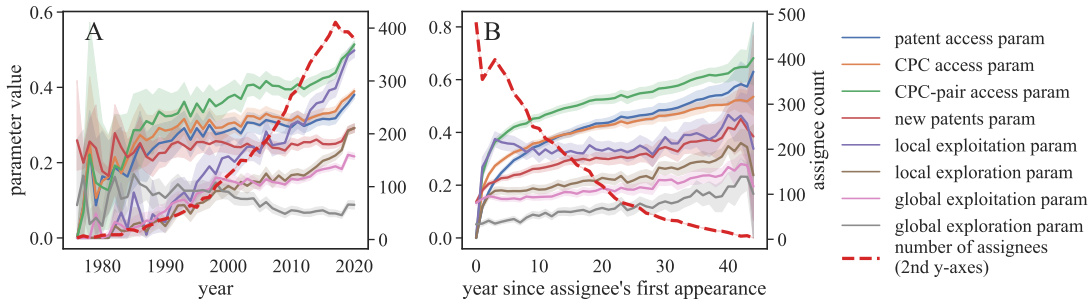


Figure 4.7: Times series of the eight knowledge access and discovery parameters of serial AI assignees from 1976 to 2020. The solid lines represent the mean values of the corresponding parameters. Their error bands represent 95% confidence intervals. The dashed red line represents the number of assignees (measured by the secondary y-axes). The x-axis of Panel A is measured by the year from 1976 to 2020. The x-axis of Panel B is measured by the year since the first year the assignee appears in the AI patent dataset.

Panel A of Figure 4.7 illustrates that as more organizations (dashed red line) invest in AI R&D, a typical organization has increasingly more access to AI knowledge, and

this is true on all the three levels of knowledge units – access to patents (blue), access to technology codes (orange), and access to code pairs (green). In addition, such accessibility has been increasing even faster in the 2010s. Also, the extent to which organizations exploit their local knowledge pool (purple line) has increased drastically. However, the number of new patents owned by a typical AI organization has remained relatively stable (red). AI organizations are tending to explore less brand-new knowledge, as shown in the downward olive line.

Panel B of Figure 4.7 illustrates the relationship between the years since an organization's first AI patent and its eight CEE parameters. It is noticeable that although the number of AI organizations is growing over time, as described in the last paragraph, the longer an organization remains active in AI patenting, the higher the chance that it drops out, as shown in the downward-trending red dashed line in Panel B. Regarding knowledge access, the three access parameters increase as the organization remains active (blue, orange, green). Nevertheless, the growth rates of the three knowledge access parameters tend to be fast during the first five years and then slow down drastically, subsequently entering an almost linear growth (a Γ-shape curve). Patent access parameters tend to grow slower than the other two initially, but outpace technology code access parameters in a decade or two. This indicates a non-trivial proportion of re-use of technology codes in creating new patents.

Contrarily, AI organizations' new patent grant parameters (solid red line) are relatively stable with a modest growth rate. A similar trend can be observed for local exploration, global exploitation, and global exploration parameters. However, it is not surprising that in terms of the extent of exhausting the search space, local exploitation is more extensive than local exploration, global exploitation, and global exploration. However, the local exploitation parameter of a typical AI organization (purple line) tends to decline after the rapid increase during the first several years,

indicating an inventive behavior favoring exploratory combinations. A slow increase in local exploitation parameters can be observed after twenty to thirty years of remaining active. This implies that when a firm first enters the AI industry, it often enters with a relatively focused vision and often exploits its local knowledge pool to the greatest possible extent during the first several years. After that, if it remains active in AI patenting, it prefers to engage in more exploratory activities, such as combining local knowledge with components found in other firms' repositories.

4.5.4 Visualizing the CEE Framework

This section offers three examples of how the CEE framework can inform individual organizations' inventive activities. The first example in Section 4.5.4 demonstrates that radar charts with eight dimensions can present multiple sets of CEE parameters of different organizations at a given time to compare their inventive behaviors. The second one in Section 4.5.4 exemplifies how to use radar charts to track short-term changes in the inventive activities of individual organizations. The third case (Section 4.5.4) offers a possible visualization method to compare multiple organizations across a more extended period.

Comparing Inventive Behavior between AI Organizations

I utilize a radar chart with eight dimensions to illustrate the eight CEE parameters of an organization at a given time. The eight CEE parameters are arranged from the top in counter-clockwise order.

Figure 4.8 is an example of the radar chart with CEE parameters of three organizations in 2020 — Marvell[6] (blue), Korea Advanced Institute of Science and Technology

[6]Marvell International Ltd. is a company that develops and produces semiconductors and is based in California.

(KAIST)[7] (red), and Alibaba[8] (green). The three organizations are randomly selected from the dataset.

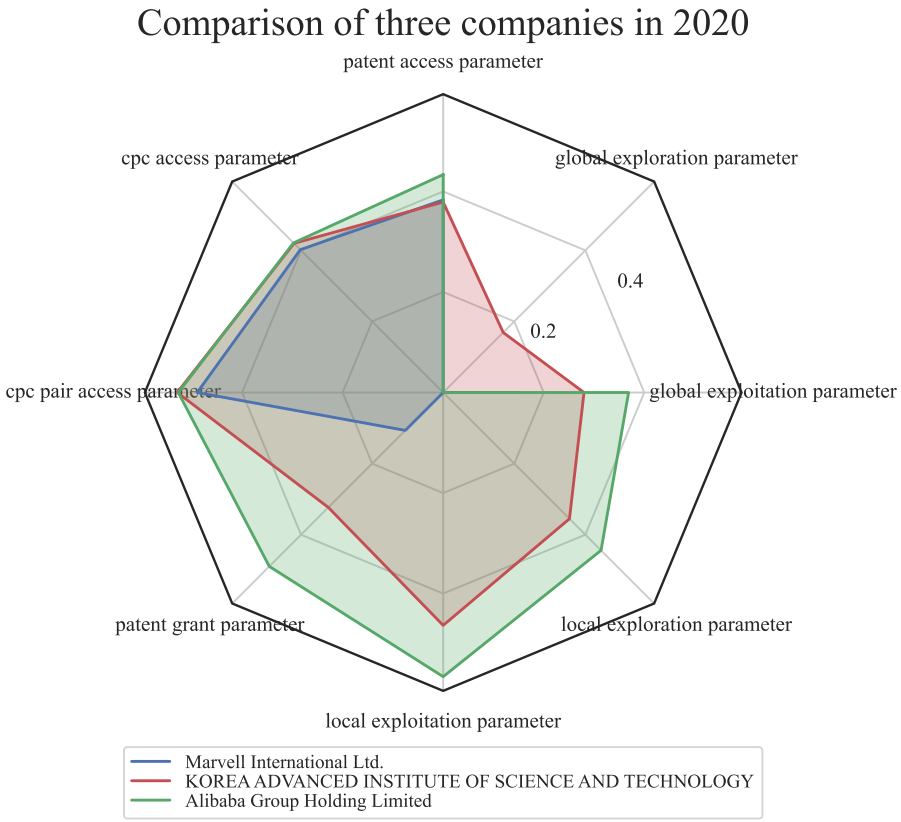


Figure 4.8: Radar charts of the eight CEE parameters of three organizations in AI patenting — Marvell (blue), Korea Advanced Institute of Science and Technology (KAIST, red), and Alibaba (green) in 2020. It can be seen clearly from the graph that Alibaba has the highest patent grants, and it exploited the local technology pool the most thoroughly. Nevertheless, KAIST, with much fewer patent grants in AI than Alibaba, has a significantly higher global exploration parameter. In other words, KAIST introduced more globally new knowledge into AI than the other two.

As the chart in Figure 4.8 shows, prior to 2020, Alibaba has accumulated the

[7]KAIST is a South Korea research university established by the Korean government in 1971 as the nation's first public, research-oriented, science and engineering institution

[8]Alibaba Group Holding Limited is a Chinese multinational technology company specializing in e-commerce, retail, Internet, and technology.

most AI patents among the three because it has the largest patent access parameter (top axis). Nevertheless, Alibaba and KAIST have similar access to technological components and their combinations. In addition, in 2020, Alibaba was granted many more new patents than KAIST or Marvell. It exploits its local knowledge pool and explores its global knowledge pool more than KAIST or Marvell. However, when it comes to exploring uncharted territory (indicated by the global exploration parameter at the upper right axis), KAIST outnumbered Alibaba and Marvell significantly, which did not engage in any global exploration in 2020. This contrast is not surprising because KAIST has conducted advanced research in AI, especially on autonomous arms (Haas, 2018).

This example demonstrates the descriptive advantage of the CEE framework, which measures inventions beyond merely "patent count," and appreciates different aspects of organizational inventive activities, namely exploitation and exploration.

Tracking the Change Over Time of Individual Organizations in Inventive Activities

The proposed CEE framework can help visualize how the inventive behavior of an individual organization evolves.

Figure 4.9 shows the biannual CEE parameters of Facebook, Inc. (currently known as Meta) from 2012 to 2020. Blue, red, green, purple, and yellow represent the CEE parameters of 2012, 2014, 2016, 2018, and 2020, respectively.

Some visible trends can be observed from Figure 4.9. First, the three knowledge access parameters are continuously growing, indicating that Facebook has accumulated inventions, technological components, and components' combinations. The patent grant parameter is also growing, although the growth seems slowed as the gaps between hexagons decrease. This observation does not indicate that the growth rate

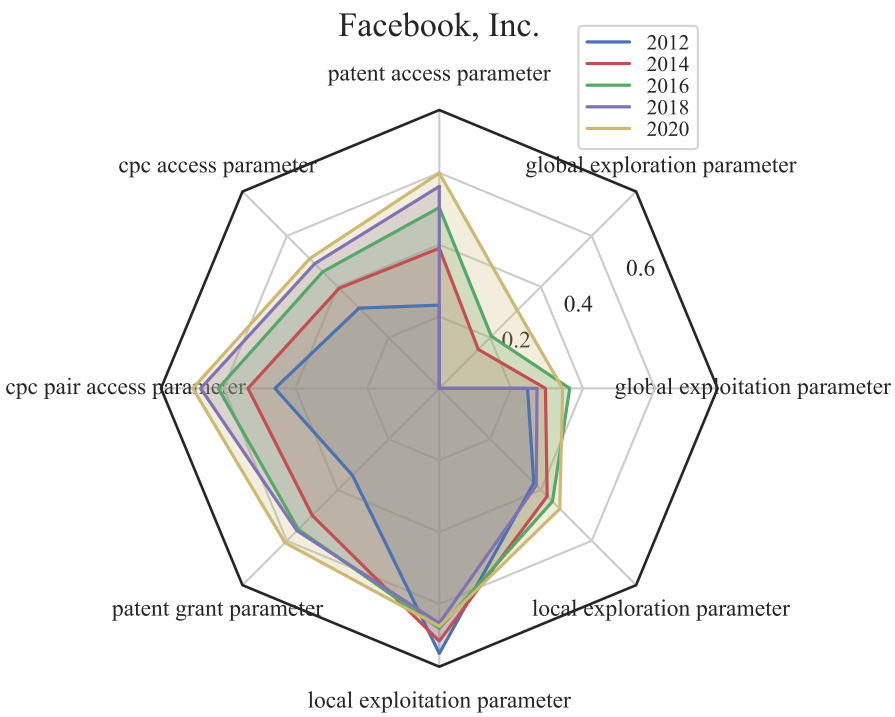


Figure 4.9: Biannual CEE parameters of Facebook, Inc.(currently known as Meta Platforms) from 2012 to 2020. Blue, red, green, purple, and yellow represent parameters in 2012, 2014, 2016, 2018, and 2020, respectively.

of new patent grants is slowing down. Instead, it simply implies that the proportion of Facebook's new AI patent grants in all new AI patents granted by the USPTO is growing at a declining rate.

Regarding knowledge discovery, it is observable from Figure 4.9 that the local exploitation parameter is declining. In contrast, the local exploration parameter and global exploitation parameter seem to be increasing, although not without fluctuations. Although no global exploration was observed in 2012 and 2018, the extent to which uncharted territory is explored has considerably expanded in the years that global exploration occurred (2014, 2016, and 2020). Therefore, it is not too hasty

to infer that as the local knowledge pool keeps growing, Facebook has been shifting its inventive preferences from merely exploiting its local knowledge to exploring new technologies, either exploring technologies mastered by its competitor or exploring technologies into which no one has ever delved.

Comparing Multiple Organizations' Change over Time

Line plots of time series would be informative to show how the CEE framework can be used to track the long-term change of inventive activities in organizations. Appendix Z shows five selected AI organizations' times series of the CEE parameters. I selected the five organizations as samples, each representing organizations that started AI patenting during different stages. For instance, IBM is selected because it has represented AI companies in the industry since the 1970s. Similarly, Google and Facebook are selected to represent respectively companies that were born in the aftermath of the dot-com bubble in the 2000s and the ones that gave rise to the era of social media in the 2010s.

Although these companies ventured into AI in different stages, some converge in some aspects. For instance, IBM, Microsoft, and Google have similar values in CPC code access parameters (orange, around 0.6) and CPC code-pair access parameters (green, around 0.8) in 2020, indicating that they have accumulated similar amounts of technological components and combinations of components. Nevertheless, Google's patent access parameter (blue, around 0.7) in 2020 is lower than that of IBM or Microsoft (around 0.8), indicating that Google has accumulated fewer AI patents than IBM or Microsoft, in turn implying that Google's AI patents on average cover a more comprehensive range of technical components than IBM or Microsoft. Regarding local exploitation (purple), Facebook has a different trend from the other four, of which the local exploitation parameters tend to grow gradually. On the contrary, the local

exploitation parameter of Facebook started with a very high value (around 0.6). It is worth noting that IBM took four decades to reach this level. IBM's value remains relatively stable at around 0.7, the highest among the five organizations.

As Panel A of Appendix Z shows, IBM's knowledge access and knowledge discovery have been relatively stable in the recent three decades, if not growing. This suggests that a certain balance between exploitation and exploration has been maintained. Among those parameters, IBM's technology code access parameter (orange) has been highly stable to the degree that it has been almost flat since the 1990s despite the slightly growing trend of patent access and code pair access.

Inventive activities in Motorola[9] as shown in Panel B, present a different tendency. Motorola's new AI patents (red), compared to the whole industry, peaked in the late-1990s and have continuously declined until the late 2010s. Similar trends can be observed for the four types of knowledge discovery. Those trends indicate that Motorola's inventive focus on AI-related technologies has gradually shifted away since the late 1990s, but in the late 2010s, it gained fresh momentum.

4.6 Conclusion

This study develops and applies a novel approach to systematically quantify organizations' knowledge access and knowledge discovery by recognizing combinatorial exploitation and exploration. I constructed a multivariate framework — the Combinatorial Exploitation and Exploration (CEE) framework — and described how to compute the eight parameters defined in this framework. This framework allows visualizing and exploring multiple dimensions of organizations' inventive activities based on the extent to which an organization has exhausted its space of possibilities.

The CEE framework introduced in this study represents one of the first attempts to

[9]Motorola here refers to Motorola Solutions, Inc.

conceptualize exploitation and exploration in organizations' combinatorial inventions using patent data. The empirical case study of AI patenting provided in this study not only demonstrates the CEE framework's potential usage and explanatory power that may enrich researchers' analytical toolkit but also opens up ample opportunities for future research agendas. For example, it could inspire studies focused on the organizational inventive activities within specific regions or industries, such as those involved in climate change mitigation, adaptation, and nanotechnology. In addition to the potential usage exemplified in this study, geographic analyses can also help study regional or national inventive activities. Appendices AA and AB offer two preliminary investigations into the geography of the combinatorial invention using the CEE framework, which respectively visualize the mean local exploration parameter of each state in the U.S. and the mean global exploitation parameters of each country that has AI patents filed at the USPTO. Future research in this direction can improve our understanding of the geographic characteristics of organizations' combinatorial inventions.

Many researchers have reported that technological diversification in organizations can impact innovation competence (Cristina and Benavides-Velasco, 2008). Providing an additional dimension of technological diversity to the proposed CEE framework can significantly improve its analytical capability, reflecting the depth and breadth of organizational inventions. Possible candidates of measurements include Shannon entropy and related variety that measures the diversity of different hierarchical levels of technology codes (Castaldi et al., 2015; Zabala-Iturriagagoitia et al., 2020).

Chapter 5

DISCUSSION AND CONCLUSION

Scientific and technological progress has been crucial in advancing societies and economies throughout history. Such progress relies heavily on the invention of new ideas, devices, and procedures, which are often considered combinations of existing knowledge and occasionally novel ones. However, as the knowledge stock of humanity grows exponentially, it becomes increasingly challenging for individuals to search for previous knowledge and discover valuable new combinations. As a result, the size of research teams has been growing in recent decades. To understand how the combinatorial process has occurred in previous records of inventive activity, it is essential to study specific fields or disciplines that strongly represent innovative advancement in science and technology during a given period.

Artificial intelligence (AI), a field that has mobilized enormous resources, shown considerable promise, and created unprecedented applications, is appropriate for investigating combinatorial inventions. This dissertation investigated combinatorial inventions related to AI from three perspectives: the creation of scientific knowledge using AI academic publications, the creation of technical knowledge using AI patents, and how organizations create AI inventions (i.e., patents) with the constraints of organizational and industrial boundaries.

This research found that while scientific publications and patents related to AI exhibit a markedly faster growth rate of inventive outcomes, they do not display significantly different patterns from previous episodes of scientific and technological advancement regarding knowledge combinations. Like science and technology in general, AI publications and patents heavily rely on combining existing knowledge

in highly conventional ways and are primarily driven by incremental improvements rather than revolutionary "paradigm shifts." In addition, AI publications that present conventional knowledge combinations tend to be more scientifically impactful when assessed by the number of citations they receive.

Furthermore, AI patents can be considered less novel than overall inventions, given their higher refinement rate and higher narrow invention rate compared to utility patents in general. Technical components involved in AI patenting are increasingly concentrated in a few fields, such as electric digital data processing. The number of organizations participating in AI patenting is growing, and the percentage of new organizations in AI patenting is significantly higher than in utility patents in general and other transformative areas, such as nanotechnology and climate change mitigation technologies. Like other organizations, AI organizations prioritize exploiting existing knowledge in their previous patent pools over exploring knowledge outside their organizational boundaries, especially in the early stages of their involvement in AI.

This dissertation provides insights into how combinatorial inventions related to AI have occurred and how they compare to previous records of inventive activity. By examining the creation of scientific and technical knowledge and the patenting activities of AI organizations, this dissertation offers empirical evidence and tools to assess the quality and outcomes of inventive activities in AI. It has implications for researchers and policymakers producing highly impactful inventions. It also sheds light on the future of inventive activities and how the combinatorial process may evolve.

5.1 Implications

5.1.1 Revolutionary Technology Can be Produced in a Normal Way

The historical progress of science and technology is often considered to repeat the Kuhn cycle of alternating "normal" and "revolutionary" phases (Kuhn, 1962; Bird, 2022). After a paradigm or a set of rules is established and well-accepted, the phase of "normal science" begins. Normal science describes the scientific and technological actives guided by the established paradigm and are mundane, routine, and intended to solve puzzles that are believed to have solutions. Such activities constitute a large proportion of work by most scientists and R&D personnel, who follow established pipelines and procedures to make cumulative improvements to existing knowledge or products. Their contributions are often considered incremental other than radical. If a puzzle is too hard to solve and starts to shatter people's confidence, a crisis arises, welcoming more radically different ideas and approaches, which may ultimately help solve the crisis, leading to the wide adoption of new ideas and giving rise to a paradigm shift and scientific revolution. Examples include Darwin's theory of evolution and Einstein's theory of relativity. Such new ideas are celebrated as "revolutionary" rather than "normal" because they appear to resolve the crisis in ways that normal science cannot. After a paradigm shift, new rules are established, science and R&D resume "normal," and another cycle has begun. Revolutionary ideas are believed to be neither incremental nor cumulative, and they often happen accidentally, contingently, unpredictably, and not by planning (Nickles, 2017).

AI has been heralded as profoundly revolutionary and transformative. In a way, AI is believed to be able to help solve many challenging puzzles and potentially change how the puzzles are solved. Those perplexing puzzles awaiting AI solutions include not only scientific or technological problems, but also economic, societal, political,

and cultural issues.

The findings of this research reveal that AI's scientific and technological development so far was not markedly different from the other "normal science" that relies heavily on incremental advancement. The evidence presented in this dissertation implies that scientific research and technological development related to AI primarily provides small and cumulative improvements to a well-established knowledge base, for example, tweaking the parameters or the arrangement of a small part in a widely-accepted neural network structure to improve the learning efficiency, or combining two existing neural networks with different objective functions to improve the quality of generated data (such as Generative Adversarial Networks) (Goodfellow et al., 2014). Inventions that are novel, original, and critical and pave the way to subsequent practices, such as back-propagation, are exceedingly rare. Even those anecdotally romanticized as "accidental" discoveries later proved to be profound, such as Andrew Ng pioneering the use of GPUs for training deep neural networks, are not accidents. Instead, they are built on careful experiments and trial-and-error and a deep understanding of previous knowledge in the field (Strickland, 2022).

This research certainly does not directly counter the argument of AI as a revolutionary technology because what was measured in this research was how AI knowledge has advanced rather than how AI affects science and technology and other aspects of society. On the contrary, this research provides an empirical portrayal of a potentially revolutionary technology, which has been advancing in an extremely "normal" and incremental way. It implies that, contrary to previous beliefs, revolutionary science and technology may occur in a way that is not accidental or contingent. Its own puzzles can be and must be solved using everyday, ordinary, bread-and-butter scientific practices. Revolutionary science and technologies can be and will likely be born from normal activities in addition to individuals' "eureka" moments.

Furthermore, this research implies that the framing of normal versus revolutionary science may be insufficient to describe AI, which, with the observations and evidence presented in this study, may be an example of where the distinction between normalcy and revolution breaks down. In Kuhn's theory, a revolution breaks out when unorthodox paradigms are pursued as responses to a crisis that cannot be solved in normal ways. Nevertheless, it is unclear where those new paradigms come from in the first place. It is likely that, before the crisis, they have already emerged, developed, and stabilized to some degree, although under the shadow of or even as a part of more mainstream, normal sciences. These new paradigms are under-recognized, under-appreciated, and underfunded to demonstrate their potential, although they likely are more advanced, progressive, or have enormous potential. Not until a crisis arises do they find themselves opportunities to shine, like mammals' ancestors that were living with dinosaurs and that did not prevail until the catastrophic mass extinction. Therefore, a new paradigm may co-exist with or even be born from an old one. As demonstrated in this research, AI is born and has been developing as a sub-field of computer science, an existing and well-established discipline. The way AI has been advancing, despite its unique characteristics, such as the requirement for large training data, has largely resembled traditional computer science and engineering. *There are arguably no fundamental revisions or conceptual discontinuities AI has brought to normal computer science.* Therefore, AI may necessitate revisiting and adapting the dichotomy of normalcy and revolution.

Accordingly, this research suggests that, rather than distinguishing normal and revolutionary science, it would be more constructive to conceptualize normalcy and revolutionary-ness as two relatively independent dimensions of science and technologies. The normalcy dimension would describe how much a technology relies on existing paradigms, while the revolution dimension measures how much it can transform

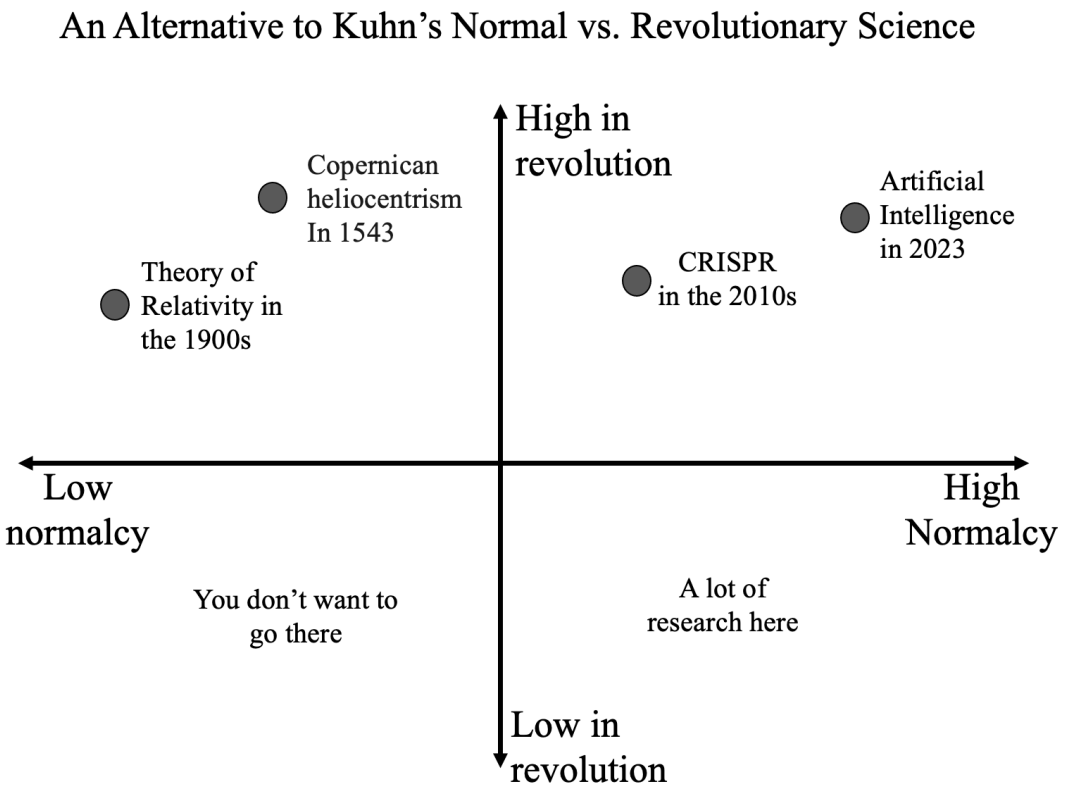


Figure 5.1: An alternative to Kuhn's normal vs. revolutionary science.

existing paradigms and open new opportunities. Accordingly, a four-quadrant coordinate system can be constructed (see Figure 5.1). While much research in the history of science and technology has been focused on the upper left quadrant, AI can be considered to reside in the upper right quadrant, where highly transformative technologies are produced with well-established rules and norms. With the increased specialization and division of research labor, more technologies can be anticipated emerging in the upper right quadrant.

5.1.2 Publishing and Inventing for High Impact

This research has implications for researchers who aim to publish high-impact articles and inventors who intend to create high-quality inventions that can profoundly affect future technologies.

Implications for Researchers

Scientific journals often recommend explicit or implicit limits on the number of references for articles seeking publication. For example, *Nature* allows up to fifty references for a research article.[1] Such a limit can affect authors' decisions regarding what articles to review and reference and what articles to exclude from references when writing their papers. The research presented here implies that referencing and connecting to more articles published in journals that have published foundational articles in the field of interest and are widely accepted as high quality would likely lead to higher citations.

Nevertheless, the results presented in this dissertation do not necessarily imply that referencing unorthodox journals reduces future citations. On the contrary, papers combining well-established journals and not-so-popular journals in their reference lists often have significantly high citations in the long run, especially for highly original and novel research. *A good balance between established and novel journals in the reference list may indicate a high-quality paper.*

References represent previous knowledge, inspiring the authors to create new knowledge. Therefore, in addition to insights into referencing strategies in paper writing, this research also implies that focusing on reviewing articles from well-established journals in the field may help researchers acquire adequate knowledge about a field and its state of the art, identify gaps that can be turned into research opportunities, craft research skills, formulate feasible research plans, and ultimately perform high-quality research. It may sound cliche'd because reviewing the literature published in the leading journals in the field is often the first step to starting a new research project. The findings of this research can be understood as accentuating the

[1] See `https://www.nature.com/nature/for-authors/formatting-guide`

significance of researchers building new research on a well-established foundation of conventional knowledge in the field. Regardless of how novel the new research turns out to be, the conventional combinations of previous knowledge referenced in the article would likely lead to higher citations. A possible reason such publications tend to be cited more is that they engage more researchers in the same field who speak the same disciplinary language and welcome incremental improvements requiring limited intellectual leaps. This implication is significant, especially in an age of knowledge explosion, where it is almost impossible for a team with limited resources to digest all previous articles in a field. Therefore, it can help researchers avoid too much time scrambling through tons of literature.

Nevertheless, the findings of this research should not be considered to undermine the value of highly novel ideas in scientific research, especially those in fields that do not have a well-established foundation or where transitions are to come. Novel combinations often occur more frequently in the early stage of a field or interdisciplinary realm and may precede radical changes. They may remain dormant in contesting citations. However, their value may be recognized in the future, a phenomenon called "*sleeping beauties in science*" (Ke et al., 2015). Moreover, novel combinations initiated in the early stage of a discipline may become conventional as they become increasingly more familiar to the researchers in the field and ultimately bonded into the field's knowledge base.[2] In this regard, it would only be mindful for researchers to be aware that highly novel research may take longer to exert impact, and it will benefit from a demonstration of stronger connections to a well-established knowledge

[2]The author defines such a process of novel combinations becoming conventional as "*conventionalization*." A separate analysis conducted by the author illustrates that, in the recent two decades, such a "conventionalization" process has sped up significantly in AI, indicating that new knowledge takes an increasingly shorter time to be accepted and diffused.

base.

Furthermore, this dissertation provides additional tools for researchers to identify well-established journals and research gaps. Undeniably, what journals are considered foundational and high quality is likely to be unspoken common knowledge for a research community, particularly those with established disciplinary boundaries. Nonetheless, such knowledge is often passed on through word-of-mouth messages and may appear vague, implicit, challenging, and confusing, especially in interdisciplinary research communities and for early career researchers. This research offers tools that can help in this regard. Specifically, well-established journals can be identified as those that are frequently cited. Research gaps and opportunities may exist in journal pairs less or never co-cited, indicating combinations of knowledge not extensively explored. Some of those less-cited combinations were introduced decades ago. However, they either remain dormant from their debut or become conventional and then novel again, maybe because they are forgotten, not promising, obsolete, or outshined by other new ideas. In the case of AI, it is also possible that in the early days of AI, many novel ideas were not implemented sufficiently due to technical limitations such as hardware or programming languages. However, those ideas may have become readily implementable with today's advanced supporting technologies.

A great example is the revival of perceptrons. The concept of perceptrons, a kind of algorithm for supervised learning of binary classifiers, was invented in 1958 by Frank Rosenblatt. It received bitter criticism on its applicability and died out (Minsky and Papert, 1988). However, it has regained vitality in recent decades due to growing computing power and helps promote many new AI advances, including neural networks and deep learning.

Here are some examples of less-co-cited journal pairs related to AI that may indicate research gaps:

- *Artificial Intelligence* and *Linguistics and Philosophy*, which represent an interdisciplinary space that involves AI and the theory of language,
- *Mind* and *Pattern Recognition*, covering a less explored agenda of philosophy, epistemology, metaphysics, and philosophy of mind intertwining with general pattern recognition techniques,
- *IEEE Transactions on Systems Man and Cybernetics* and *Journal of Documentation*, covering the overlapping arena of system engineering and information science,
- *Behavioral Sciences* and *The Lancet*, which link psychology, neuroscience, cognitive science, and behavioral biology with medical research in the realm of AI, and
- *Social Studies of Science* and *Technology and Culture*, which combine the history and philosophy of science with the history of technology related to AI.

These journal pairs may be valuable if reevaluated or exploited with appropriate, modern thoughts and techniques. Therefore, this research also implies that literature from the early age of AI may have great value for the future innovation and development of AI.

Implications for Inventors

This research provides inventors insights into how to create high-quality inventions. Again, *a good balance between common and novel technical components may be essential to a patent's success.*

Highly original patents have significantly high disruption. They lead to new paths that subsequent patents in the field are likely to follow. Nevertheless, they are less

likely to be cited because future inventors may find them more challenging to build upon, extend, and refine. On the other hand, refinement patents that only provide minor improvements to existing technologies are likely to be cited more because they provide ample familiarity that invites incremental improvements. Generally speaking, patents that are more balanced along the origination-refinement spectrum, particularly patents of novel combinations, which combine existing components with new ones, often have the most balanced performance. They are neither too radical nor cut and dried, and they often receive proper citations and cause significant disruption.

Such observations suggest that calculatedly creating patents of "novel combinations" may benefit inventors in achieving greater success. Undeniably, inventing is not simply playing legos with technical components. Arbitrarily combining components without assessing their patentability and usability would likely lead to rejection. Nevertheless, this research suggests that extra objectives can be added to inventors' searching and drafting strategies. Searching previous patents is critical for inventors to understand the state of the prior art so that what is novel and different in the new invention compared to previous technologies can be better described when drafting patent applications. This research suggests patent searches can be improved by providing not only what components and functionalities are standard in the field but also what is not, in other words, what components are rarely or never used in the field and potentially lead to patents of novel combinations. Assessing the feasibility of combining those components with common ones may provide inventors extra inspiration or revelation.

5.1.3 Assessing Inventive Performance

This research provides insights for funding agencies and organizations that wish to explore the combinatorial characteristics of publications. The combinatorial char-

acteristics discussed in this dissertation can be tested, embedded, and incorporated into the metrics that assess research projects' novelty.

In particular, the CEE framework developed in Chapter 4 provides a valuable tool for organizations to assess previous inventive performance, evaluate current inventive competence and resources, identify opportunities and challenges, and make reasonable predictions regardless of domains. The CEE framework measures an organization's inventive performance and competence with the consideration of its previous knowledge stock and its access to industry knowledge, which goes beyond the conventional measurements that simply count the number of certain types of patents. It prevents the common error of comparing ants with elephants. For example, the demonstration presented in Chapter 4 shows that although Alibaba has more AI patents than KAIST, the latter engaged in substantially more exploratory research than the former. It is unsurprising because KAIST is a research institute aiming to expand scientific boundaries, while Alibaba is a private company that produces commercial products and services to generate financial profit. Without the CEE framework, Alibaba would appear to perform better than KAIST in AI patenting. Nevertheless, the CEE framework reveals the different inventive activities occurring in the two organizations — KAIST, although with fewer patent grants, expands the knowledge boundary of the field of AI to a much greater extent, while Alibaba engages in no inventions that push AI's knowledge boundary. Therefore, policymakers in organizations interested in inventions of different natures rather than the number of patent grants can benefit from the CEE framework, which provides more meaningful metrics. Another example the CEE framework presents illustrates the evolution of the inventive behavior of Meta (former Facebook) from 2012 to 2020 — less exploitation and increased exploration.

In addition, national and regional governments can utilize the CEE framework to evaluate the inventive behaviors of different countries and regions.

5.1.4 Team sizes matter

The results of this research regarding how team size matters to inventions' nature and quality carry meaningful implications for allocating human resources to R&D projects. One general observation is that more people bring more knowledge — for instance, more than half of lone inventors in machine learning engage only two components in their patents, and such a percentage is halved to less than 25% for teams of ten. More knowledge means a more extensive foundation to build upon.

However, larger teams do not necessarily perform better. When the size of a team grows, the amount of prior knowledge brought to the conversation naturally increases. However, it may also introduce more constraints and conflicts of opinion, dilute responsibilities, and impair efficiency. Larger teams may face more challenges in reaching consensus, particularly on radical or controversial ideas. This could potentially result in minor modifications and less novel advancements

This research implies that larger teams may perform well for projects that engage in highly integrated systems with more components and aim to improve existing practices incrementally. On the other hand, highly original and novel patents that intend to disrupt future practices in the field are likely to be created by smaller but efficient teams, such as teams of two or three people or even lone inventors.

Those implications are meaningful for inventors and policymakers in organizations responsible for allocating research budgets and human resources. By recognizing the advantages and disadvantages of large and small teams, more realistic expectations of inventive outcomes can be set up, and more detailed guidance on team size can be offered.

5.1.5 Public Supported AI

A strong reliance on public funding often indicates that the field is still in its early stage and requires a longer time for commercialization. This research found that AI patents have relied heavily on scientific knowledge funded by the government, more heavily than utility patents in general, with patents related to computer vision, AI hardware, NLP, and machine learning having the strongest reliance on public support. Machine learning and computer vision also have the highest percentage of academic patents, which universities own.

Publicly-supported AI patents are of high quality — they tend to receive more citations and have a significantly higher chance of becoming the most disruptive. Therefore, science and technology would benefit from continuous public funding for AI, especially machine learning, computer vision, AI hardware, and NLP. The findings of this research can also be interpreted as an indication that there is a treasure of AI knowledge (and human resources) in universities and government, especially in machine learning, computer vision, and knowledge processing. Collaborations between academic institutes and the private sector will likely stimulate technology transfer and generate economic benefits for society.

5.2 Limitations

This dissertation has several limitations.

First, the datasets used in this research are limited by their scopes and accuracy. As detailed in Section 2.3.1, the AI publications dataset was collected in 2020 from the Web of Science (WOS) using keyword-search in a snowball fashion. According to their agreements, the WOS provides different services and databases to institutional subscribers. Limited by the institutional subscription available to the author at the

time such dataset was collected (2020), only four databases are searched (see footnote 2 in page 29) and the Conference Proceedings Citation Index was not among them. Therefore, the AI publication dataset does not systematically include conference proceedings. Nevertheless, AI researchers increasingly choose to present their research at conferences, like the conference of Neural Information Processing Systems (NIPS) and International Conference on Machine Learning (ICML). Accordingly, many of those papers are published in conference proceedings rather than journals. Therefore, the AI publication dataset collected for this dissertation can only be considered a bounded sample of AI research that does not encompass conference proceedings. Accordingly, the conclusions and implications drawn on the analysis of this dataset are limited by such a scope. Nevertheless, the author believes the AI publication dataset used in this research remains highly relevant, appropriate, and significant in representing AI research. Conference papers are often more concise than journal articles and intend to communicate the latest progress of the authors' research to a significantly specialized audience rather than thorough investigations. The peer-review processes for conference papers and journal articles are also different. While the former is completed in more fixed windows of time, the latter is more comprehensive, extensive, and extended and allows more iterations of discussion and revision. Therefore, AI publications in journals can be considered a representative sample of high-quality AI research.

The AI publication dataset is also limited by the search terms used to construct the dataset. The initial keywords were selected heuristically, and even though supplemented by the additional search terms recruited in the snowball process, they may be insufficient to encompass the topics of AI because many new terms related to AI keep emerging. Furthermore, during each iteration of the snowball process, only the twenty most frequent keywords are selected as candidates for the next iteration. This step

is necessary given the limited time and resources available to this dissertation, but it risks omitting some important keywords, especially emerging ones. The synonym aggregation conducted by the author is heuristic in nature. Limited by the author's knowledge about AI, mistakes and omissions may exist. In addition, the analysis of AI publications is limited by the language of choice. As mentioned in Chapter 2, WOS only includes publications in English and publications with metadata translated into English. Therefore, WOS contains primarily English publications. So is the AI publication dataset used in this dissertation. 98.4% of the papers in the dataset is written in English, 0.5% in Chinese, 0.3% in Spanish, and 0.8% in others. With a significant proportion of AI research happening in China and conducted by Chinese researchers,[3] such a language limitation may lead to biases that underestimate the knowledge combinations recorded in publications in other languages.

The AI patent dataset used in Chapters 3 and 4 also has limitations. It is limited by the machine learning models developed by the USPTO to predict the "AI probability" of each patent (Giczy et al., 2021). Such models generate prediction scores primarily based on the text of patents' abstracts and claims, and they are accurate enough but not 100% accurate. During data exploration, the author has semantically identified several patents in the dataset with high AI scores but in fact unrelated to AI. They are misclassified as AI patents because they contain AI-like terms that mean different things in different contexts, such as "learning devices," which may refer to a device that helps humans learn knowledge or skills rather than machine learning.

[3]To iterate from Chapter 2: "Although many Chinese-speaking authors chose to publish their works in English journals (the dataset used in this research shows that authors based in China have published more AI papers than in other countries, and among the top 10 research institutions that have the most AI publications, five are located in China), publications in the Chinese language are likely to be significant in number."

The assignee dataset used in Chapter 4 is limited by the accuracy of the disambiguation algorithm used by the *PatentsView* team to identify organizations (Monath et al., 2021). Because, unlike some other patent offices in the world (such as the patent office of China), the USPTO does not mandate patent applicants to provide personally or organizationally identifiable information in patent filing, it becomes challenging to identify individual inventors and assignees. The assignee dataset collected for this dissertation provides a unique identifier for each distinct inventor and assignee generated by inference. In the real world, inventors and companies may change names, and typos exist. For example, the author has identified dozens of misspelled names for IBM in the AI patent dataset, and because of this, they are associated with different identifiers from IBM. These factors may undermine, to some degree, the accuracy of the analysis conducted in this dissertation.

In addition, only U.S. patents are included in this research, limiting the geographic scope of this dissertation. Undeniably, the USPTO routinely receives a large number of foreign-invented patent applications, more than domestic-invented applications (WIPO, 2019), and these patents' rights "extend only throughout the territory of the United States and have no effect in a foreign country."[4] Patents granted by the USPTO cannot be considered to match the record of world patents, even though those patents may belong to patent families that have equivalent applications in other countries. Therefore, this dissertation can only be interpreted within the patents whose property rights are only valid in the U.S.

Secondly, this dissertation is limited by its methodologies. The novelty analysis is limited within the domain of AI and cannot be generalized without caution into other domains and scientific publications and patenting in general. Specifically, Chapter 2 investigated how novel AI publications are and how much combinatorial novelty re-

[4] `https://www.uspto.gov/ip-policy/ipr-toolkits`

lates to AI publications' scientific impact. Those results are not tested in publications related to other domains, such as nanotechnology and civil engineering, let alone all scientific publications in general. Nevertheless, novelty is relative to the background. In other words, whether something is novel depends on to what background it is compared. A novel idea in AI may be extensively investigated in another discipline — vice versa. Without comparing to other backgrounds, this research is restricted in its explanatory power regarding the trajectory of science and technology in general. The author attempted to overcome such a limitation in Chapter 3 by providing analysis on characteristics like novelty within utility patents in general in addition to AI, for instance, see Figure 3.4 on page 85. Nevertheless, such analyses on other disciplines and for scientific publication are still insufficient.

Some measurements and variables in this dissertation were selected because of resource limitations and can be improved. For example, annual citations are used to measure the scientific impact of AI publications. Undeniably, the annual citation of an article captures to some degree its average impact during the years after its publication. Nevertheless, more researchers are inclined to consider citation ages and use citation counts during the first several years (usually five years) as a more appropriate indicator for scientific impact because the citation of a paper often reaches its peak during the first two years and then gradually declines. Therefore, using citation counts may risk an overestimation of aged articles and an underestimation of recent articles. On the other hand, using annual citations may lead to underestimating the impact of aged articles. Nevertheless, limited by the author's institutional subscription to WOS, it is challenging to identify publications' citation ages. Therefore, additional variables, such as citation percentile of the year, are computed to allow comparison among publications and patents in the same year.

Another methodological limitation lies in the representations of knowledge. Chap-

ter 2 identifies referenced journals as representing previous knowledge combined in publications. This simplified approximation may be insufficient to capture the nuance of knowledge. A journal is more of a collection than a piece of knowledge, evolving over time. It is even more questionable in the case of journals targeting broader disciplines, such as *Nature* and *Science*. In analyzing patents (Chapters 3 and 4), technology codes are considered to represent knowledge or technical components combined in the patents. Again, it is simplified to consider the set of technology codes to represent patents' technical content. A patent describes or discloses its technical content in the form of claims. A claim rather than a patent itself represents a scope of protection or an invention. A patent may contain multiple claims, and a claim could be dependent or independent. In other words, depending on the structure of claims, a patent could include multiple inventions, each representing a separate combination of components. However, this dissertation's method does not reflect such nuances and could be drastically improved.

Chapter 4, which discussed how organizations combine existing technologies to create new AI inventions, is limited by its assumptions that a patent is owned by its assignees. Limited by the data source, such assignees only reflect the assignees "at issue"[5]. In other words, the assignee dataset collected for Chapter 4 only reflects the organizations or individuals who receive the patent rights the moment the patents are granted and "does not collect or carry reassignment data for if additional changes to the assignee occur after the patent is granted." However, corporate structural changes, such as acquisitions, mergers, bankruptcies, or even name changes, are not uncommon in the business world. Patents and any other intellectual properties owned by a company acquired by another are passed to the latter. Such ownership shifts resize the pools of patents and components of organizations involved but are not

[5] `https://patentsview.org/forum/7/topic/536`

addressed in this dissertation.

5.3 Next Steps

Plans to extend the research in this dissertation will be the following — diversifying invention datasets, generalizing to other inventive areas besides AI, testing other representations of "knowledge," and improving methodologies.

Chapter 2 is based only on publications that went through standard academic publishing practice and, in particular, the peer-review process. That is to say, reports, non-reviewed publications, and non-technical discussions (such as arXiv,[6] GitHub repositories,[7] and social media discussion) are not included in the analysis. AI researchers increasingly choose to publish their papers on non-traditional platforms like arXiv. Organizations, individual practitioners, and hobbyists tend to publish AI codes on Github without writing papers. Those efforts are growing in scale and should not be overlooked in future assessments of AI invention. Therefore, incorporating records from arXiv as a complementary for peer-reviewed publications would be an appropriate next step.

Furthermore, Chapters 3 and 4 analyzed U.S. patent grants related to AI. In other words, patents granted by other countries are excluded. Although the USPTO is one of the most prominent patent offices in the world and the U.S. patents can be considered a representative sample of global innovation, there is a risk that some vital information may be omitted because of this selection. To complete the assessment, the next step would be to collect a global patent dataset from WIPO and use patent families to identify patents with the same technical content but filed in different jurisdictions.

[6] `https://arxiv.org`

[7] `https://github.com`

Moreover, when assessing the impact of publications and patents, in addition to the frequently-used metrics such as citation counts and disruptive index, social media discussion becomes increasingly crucial to scholarly communication in disseminating, promoting, and democratizing scientific knowledge and technical accomplishment (Sugimoto et al., 2017). It would only be meaningful to incorporate social media discussion into the assessment of scholarly impact. Therefore, the future research following this dissertation is to integrate social media metrics, particularly Twitter content, into evaluating the impact of scientific publications as complementary to citation metrics.

To understand the general laws of inventions, analyzing one discipline is far from completion. Therefore, it would be helpful to collect publication and patent datasets of other technologies, such as nanotechnology, biotechnology, and climate change mitigation technology, and conduct similar analyses to compare the similarities and differences to draw more generalizable conclusions.

Another meaningful question that deserves exploration is to look for better ways to represent the quanta of knowledge. This dissertation adopted previous research in which referenced journals in publications and technology codes in patents are considered previous knowledge that is combined. Nevertheless, criticism has pointed out that those representations are only approximations that may produce misleading results. A scientific journal is a collection of intellectual fruits, and it would be arguably overly simplistic to reduce its rich and cumulative content to represent a piece of previous knowledge. Moreover, it is debatable what specific previous knowledge it represents when it comes to multi-disciplinary journals or journals targeting non-specific disciplinary audiences, such as *Science* and *Nature*. In the case of patents, although it is more widely accepted to consider a technology code (especially a full-digit code) as a technical component, we should be cautious when using it because technology codes

are created for classification and search purposes. A technology code, by definition, does not necessarily refer to a specific piece of technology. Moreover, as time goes by and technologies change, re-classification occurs frequently because new codes may emerge, and existing codes may disappear or merge with others. In this regard, a technology code is more of a cluster or collection of technical information well-defined and understood by a skilled technician in the field where more granular details become irrelevant. What can be considered well-defined differs across fields. Therefore, many new approaches, especially computational ones, such as NLP and machine learning, have started to be experimented with to identify "knowledge" in inventive records. Some approaches intend to summarize a field's state of the art rather than identify or quantify specific knowledge. It would also be a meaningful attempt to extract information from patents' claims using NLP approaches.

In addition, some patents are assigned to "c-sets." A c-set is a combination of several codes considered as one classification. Only a few CPC subclasses are authorized to form c-sets (EPO and USPTO, 2022). By definition, a c-set is considered one technology, and codes contained in a c-set cannot be considered as combined with other codes that are not in the same c-set. The next step would be considering c-sets and formulating more accurate representations of combinations of technical components in patents.

REFERENCES

Agrawal, A., McHale, J., and Oettl, A. (2018). Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth. Working paper w24541, National Bureau of Economic Research, Cambridge, MA.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., and Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, 60:102387.

Albarrán, P., Crespo, J. A., Ortuño, I., and Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, 88(2):385–397. Place: Dordrecht Publisher: Springer Netherlands.

Alcácer, J., Gittelman, M., and Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research policy*, 38(2):415–427. Place: Amsterdam Publisher: Elsevier B.V.

Andriopoulos, C. and Lewis, M. W. (2009). Exploitation-Exploration Tensions and Organizational Ambidexterity: Managing Paradoxes of Innovation. *Organization Science*, 20(4):696–717. Publisher: INFORMS.

Arthur, W. B. (2009). *The Nature of Technology: What It Is and How It Evolves*. Free Press, New York.

Austen, J. and Grahame-Smith, S. (2009). *Pride and Prejudice and Zombies*. Quirk Books, 1st edition edition.

Bali, J., Garg, R., and Bali, R. (2019). Artificial Intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian Journal of Ophthalmology*, 67(1):3–6.

Barham, L. (2013). *From Hand to Handle: The First Industrial Revolution*. Oxford University Press, New York, NY, illustrated edition edition.

Basalla, G. (1988). *The evolution of technology*. Cambridge history of science. Cambridge University Press, Cambridge [England].

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):361–362. Number: 1.

Becker, H. S. (1982). *Art worlds*. University of California Press, Berkeley.

Benner, M. J. and Tushman, M. L. (2003). Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited. *The Academy of Management Review*, 28(2):238–256. Publisher: Academy of Management.

Bezerra, C. O., Carneiro, L. L., Carvalho, E. A., das Chagas, T. P., de Carvalho, L. R., Uetanabaro, A. P. T., da Silva, G. P., da Silva, E. G. P., and da Costa, A. M. (2021). Artificial Intelligence as a Combinatorial Optimization Strategy for Cellulase Production by Trichoderma stromaticum AM7 Using Peach-Palm Waste Under Solid-State Fermentation. *BioEnergy Research*.

Biernacki, P. and Waldorf, D. (1981). Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research*, 10(2):141–163. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

Bird, A. (2022). Thomas Kuhn. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2022 edition.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Biswas, R., Blackburn, L., Cao, J., Essick, R., Hodge, K. A., Katsavounidis, E., Kim, K., Kim, Y.-M., Bigot, E.-O. L., Lee, C.-H., Oh, J. J., Oh, S. H., Son, E. J., Vaulin, R., Wang, X., and Ye, T. (2013). Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data. *Physical Review D*, 88(6):062003. arXiv: 1303.6984.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics*, 2008(10):P10008–12. Place: Ithaca Publisher: IOP Publishing.

Blum, C., Puchinger, J., Raidl, G. R., and Roli, A. (2011). Hybrid metaheuristics in combinatorial optimization: A survey. *Applied soft computing*, 11(6):4135–4151. Publisher: Elsevier BV, Elsevier.

Bornmann, L. (2020). How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles. *Quantitative Science Studies*, 1(4):1553–1569.

Borschke, M. (2017). *This is Not a Remix: Piracy, Authenticity and Popular Music*. Bloomsbury Publishing Inc, Bloomsbury Academic & Professional, Bloomsbury Academic, New York, 1 edition. Pages: 11, 33, 1, 167, 113, 159, 71–166, 112, 168, 158, 70, 186, 10, 32.

Brundage, M. (2019). *Responsible Governance of Artificial Intelligence: An Assessment, Theoretical Framework, and Exploration*. PhD thesis, Arizona State University, Tempe, Arizona.

Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370):1530–1534. Publisher: American Association for the Advancement of Science Section: Policy Forum.

Castaldi, C., Frenken, K., and Los, B. (2015). Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting. *Regional Studies*, 49(5):767–781.

Chalmers, D., MacKenzie, N. G., and Carter, S. (2021). Artificial Intelligence and Entrepreneurship: Implications for Venture Creation in the Fourth Industrial Revolution. *Entrepreneurship Theory and Practice*, 45(5):1028–1053. Publisher: SAGE Publications Inc.

Chang, A. C. (2020). *Intelligence-based medicine: artificial intelligence and human cognition in clinical medicine and healthcare*. Academic Press, London, England. Book Title: Intelligence-based medicine : artificial intelligence and human cognition in clinical medicine and healthcare.

Chen, L., Xuan, J., Riggins, R. B., Clarke, R., and Wang, Y. (2011). Identifying cancer biomarkers by network-constrained support vector machines. *BMC systems biology*, 5:161.

Choi, J., Jang, D., Jun, S., and Park, S. (2015). A Predictive Model of Technology Transfer Using Patent Analysis. *Sustainability*, 7(12):16175–16195. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Christensen, C. M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. The management of innovation and change series. Harvard Business School Press, Boston, Mass.

Church, K. (2011). A Pendulum Swung too Far. *Linguistic Issues in Language Technology*, 6(5).

Clancy, M. (2015). *Combinatorial innovation, evidence from patent data, and mandated innovation*. PhD Thesis, Iowa State University. ISBN: 9781339143118.

Clarivate Analytics (2018). KeyWords Plus generation, creation, and changes.

Cockburn, I. M., Henderson, R., and Stern, S. (2019). The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Coles, P. (2019). Einstein, Eddington and the 1919 eclipse. *Nature*, 568(7752):306–307. Place: England Publisher: Nature Publishing Group.

Copeland, S. (2019). On serendipity in science: discovery at the intersection of chance and wisdom. *Synthese*, 196(6):2385–2406.

Correia, A. and Reyes, I. (2020). AI research and innovation: Europe paving its own way. Technical report, European Commission, Directorate-General for Research and Innovation, Publications Office.

Crafts, N. (2021). Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford review of economic policy*, 37(3):521–536. Place: UK Publisher: Oxford University Press.

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, New York, NY. Book Title: AI : the tumultuous history of the search for artificial intelligence.

Cristina, Q.-G. and Benavides-Velasco, C. A. (2008). Innovative competence, exploration and exploitation: The influence of technological diversification. *Research policy*, 37(3):492–507. Place: Amsterdam Publisher: Elsevier B.V.

Crouch, S. R., Cullen, T. F., Scheeline, A., and Kirkor, E. S. (1998). Kinetic Determinations and Some Kinetic Aspects of Analytical Chemistry. *Analytical chemistry (Washington)*, 70(12):53–106. Place: Washington Publisher: American Chemical Society.

Das, J. P., Aherne, E., and Kavanagh, E. (2019). Imaging of the Spine: A Bibliometric Analysis of the 100 Most-Cited Articles. *Spine*, 44(22):1593–1598. Place: United States Publisher: Copyright Wolters Kluwer Health, IncAll rights reserved.

Davis, D. S., Sanger, M. C., and Lipo, C. P. (2019). Automated mound detection using lidar and object-based image analysis in Beaufort County, South Carolina. *Southeastern Archaeology*, 38(1):23–37. Publisher: Routledge _eprint: https://doi.org/10.1080/0734578X.2018.1482186.

de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L.-Q., Moore, J. E., Morgan, D., Jacobs, R., Pollock, T., Schlom, D. G., Toberer, E. S., Analytis, J., Dabo, I., DeLongchamp, D. M., Fiete, G. A., Grason, G. M., Hautier, G., Mo, Y., Rajan, K., Reed, E. J., Rodriguez, E., Stevanovic, V., Suntivich, J., Thornton, K., and Zhao, J.-C. (2019). New frontiers for the materials genome initiative. *npj Computational Materials*, 5(1):1–23. Number: 1 Publisher: Nature Publishing Group.

Deacon, T. W. (1997). *The symbolic species: the co-evolution of language and the brain*. W.W. Norton, New York, 1st ed edition.

Derry, T. K. and Williams, T. I. (1993). *A short history of technology: from the earliest times to A.D. 1900*. Dover Publications, New York. OCLC: 868510009.

Dixon, S. E., Meyer, K. E., and Day, M. (2007). Exploitation and exploration learning and the development of organizational capabilities: A cross-case analysis of the Russian oil industry. *Human Relations*, 60(10):1493–1523. Publisher: SAGE Publications Ltd.

Donnellan, P. R. (2018). The Future of Mobility-Electric, Autonomous, and Shared Vehicles. *IEEE Engineering Management Review*, 46(4):16–18.

Edquist, C. (2005). Systems of Innovation: Perspectives and Challenges. In Fagerberg, J., Mowery, D. C., and Nelson, R., editors, *The Oxford handbook of innovation*, pages 181–208. Oxford University Press, Oxford ;.

Edquist, C. and Johnson, B. (1997). Institutions and Organizations in Systems of Innovation. In Edquist, C. and de la Mothe, J., editors, *Systems of Innovation: Technologies, Institutions, and Organizations*, Science, Technology, and the International Political Economy, pages 41–63. Pinter. Google-Books-ID: Sf0POR0ffWEC.

EPO and USPTO (2022). Guide to the CPC (Cooperative Patent Classification).

Eric, S. and Work, B. (2021). 2021 Final Report: National Security Commission on Artificial Intelligence. Technical report, National Security Commission on Artificial Intelligence, Washington, D.C.

Fang, C., Lee, J., and Schilling, M. A. (2010). Balancing Exploration and Exploitation Through Structural Design: The Isolation of Subgroups and Organizational Learning. *Organization Science*, 21(3):625–642. Publisher: INFORMS.

Fang, W. and Lou, L. (2019). Constructing the Evaluation System of Urban Land Ecosystem Based on Entropy Method and Analytic Hierarchy Process-Taking Shaanxi Province as an Example. *IOP conference series. Earth and environmental science*, 295(2):12072–. Place: Bristol Publisher: IOP Publishing.

Fernandez, M., Fernandez, M., Caballero, J., Caballero, J., Fernandez, L., Fernandez, L., Sarai, A., and Sarai, A. (2011). Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Molecular diversity*, 15(1):269–289. Place: Dordrecht Publisher: Springer Netherlands, Springer, Springer Nature BV.

Fernandez-Luque, L. and Imran, M. (2018). Humanitarian Health Computing Using Artificial Intelligence and Social Media: A Narrative Literature Review. *International journal of medical informatics (Shannon, Ireland)*, 114:136–142. Place: Ireland Publisher: Elsevier BV.

Fleming, L., Green, H., Li, G.-C., Marx, M., and Yao, D. (2019a). Replication Data for: Government-funded research increasingly fuels innovation.

Fleming, L., Greene, H., Li, G., Marx, M., and Yao, D. (2019b). Government-funded research increasingly fuels innovation. *Science (American Association for the Advancement of Science)*, 364(6446):1139–1141. Place: United States Publisher: The American Association for the Advancement of Science.

Fleming, L. and Marx, M. (2006). Managing Creativity in Small Worlds. *California Management Review*, 48(4):6–27.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. (2019a). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539. Publisher: National Academy of Sciences Section: Perspective.

Frank, M. R., Wang, D., Cebrian, M., and Rahwan, I. (2019b). The Evolution of Citation Graphs in Artificial Intelligence Research. *Nature Machine Intelligence*, 1(2):79–85. Number: 2 Publisher: Nature Publishing Group.

Fujii, H. and Managi, S. (2018). Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. *Economic analysis and policy*, 58:60–69. Publisher: Elsevier BV, Elsevier BV.

Funk, R. J. and Owen-Smith, J. (2017). A Dynamic Network Measure of Technological Change. *Management Science*, 63(3):791–817. Publisher: INFORMS.

Funk, R. J., Park, M., and Leahey, E. (2022). Papers and patents are becoming less disruptive over time (1.0).

Gal, O. (2021). *The Origins of Modern Science: From Antiquity to the Scientific Revolution*. Cambridge University Press, Cambridge, United Kingdom; New York, NY.

Geiger, S. W. and Makri, M. (2006). Exploration and exploitation innovation processes: The role of organizational slack in R & D intensive firms. *The Journal of High Technology Management Research*, 17(1):97–108.

Gemba, K. and Kodama, F. (2001). Diversification dynamics of the Japanese industry. *Research policy*, 30(8):1165–1184. Place: Amsterdam Publisher: Elsevier B.V.

Gertner, J. (2012). *The idea factory: Bell Labs and the great age of American innovation*. Penguin Press, New York. OCLC: 733230713.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature (London)*, 521(7553):452–459. Place: England Publisher: Springer Science and Business Media LLC, Nature Publishing Group.

Giczy, A. V., Pairolero, N. A., and Toole, A. A. (2021). Identifying artificial intelligence (AI) invention: a novel AI patent dataset. *The Journal of technology transfer*, 47(2):476–505. Place: New York Publisher: Springer US.

Gong, Y., Le, Y., Zhang, X., and Chen, X. (2021). Impacts of Practice Combinations on Organizational Knowledge: Based on March's Exploration-Exploitation Model. *Complexity*, 2021:e5618287. Publisher: Hindawi.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Greve, H. R. (2007). Exploration and exploitation in product innovation. *Industrial and corporate change*, 16(5):945–975. Place: Oxford Publisher: University Press.

Gruber, M., Harhoff, D., and Hoisl, K. (2013). Knowledge Recombination Across Technological Boundaries: Scientists vs. Engineers. *Management science*, 59(4):837–851. Publisher: Institute for Operations Research and the Management Sciences INFORMS, INFORMS, Institute for Operations Research and the Management Sciences.

Gunkel, D. J. (2016). *Of remixology: ethics and aesthetics after remix*. The MIT Press, Cambridge, Massachusetts ;.

Gupta, A. K., Smith, K. G., and Shalley, C. E. (2006). The interplay between exploration and exploitation. *Academy of Management Journal*, 49:693–706. Place: US Publisher: Academy of Management.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M., Adams, R. P., and Aspuru-Guzik, A. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127. Place: England Publisher: Springer Nature - Nature Publishing Group.

Haas, B. (2018). 'Killer robots': AI experts call for boycott over lab at South Korea university. *The Guardian*.

Hathaway, H. (1953). Niagara. IMDb ID: tt0046126 event-location: USA.

Haugeland, J. (1985). *Artificial intelligence: the very idea*. MIT Press, Cambridge, Mass. OCLC: 11840529.

He, L., Liu, Y., Yang, K., Zou, Z., Fan, C., Yao, Z., Dai, Y., Li, K., Chen, J., and Yao, X. (2021). The discovery of Q-markers of Qiliqiangxin Capsule, a traditional Chinese medicine prescription in the treatment of chronic heart failure, based on a novel strategy of multi-dimensional "radar chart" mode evaluation. *Phytomedicine (Stuttgart)*, 82:153443–153443. Place: Germany Publisher: Elsevier GmbH.

Henderson, R. and Cockburn, I. (1994). Measuring Competence? Exploring Firm Effects in Pharmaceutical Research. *Strategic management journal*, 15(S1):63–84. Place: Chichester Publisher: John Wiley & Sons, Ltd.

Hendler, J. (2008). Avoiding Another AI Winter. *IEEE intelligent systems*, 23(2):2–4. Publisher: IEEE, IEEE Computer Society.

Hinton, G. E., Krizhevsky, A., Sutskever, I., and Srivastva, N. (2016). System and method for addressing overfitting in a neural network.

Hippel, E. v. (1988). *The sources of innovation*. Oxford University Press, New York.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., and McFarland, D. A. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences - PNAS*, 117(17):9284–9291. Place: United States Publisher: National Academy of Sciences.

Howe, J. (2007). History of Artificial Intelligence at Edinburgh. Publisher: School of Informatics, The University of Edinburgh.

Hutson, M. (2020). Core progress in AI has stalled in some fields. *Science*, 368(6494):927–927. Publisher: American Association for the Advancement of Science Section: In Depth.

Introna, L. and Wood, D. (2002). Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems. *Surveillance & Society*, 2(2/3):177–198.

Jacquemin, A. P. and Berry, C. H. (1979). Entropy Measure of Diversification and Corporate Growth. *The Journal of Industrial Economics*, 27(4):359–369.

Jones, C. I. (1995). R&D-based models of economic growth. *The Journal of political economy*, 103(4):759–. Place: Chicago Publisher: University of Chicago Press.

Kaempffert, W. (1930). *Invention and society*. American library association, Illinois. Book Title: Invention and society, by Walkemar Kaempffert.

Kaplan, J. (2016). *Artificial intelligence: what everyone needs to know*. What everyone needs to know. Oxford University Press, New York, NY. OCLC: 945693664.

Kauffman, S., Lobo, J., and Macready, W. G. (2000). Optimal search on a technology landscape. *Journal of economic behavior & organization*, 43(2):141–166. Place: Amsterdam Publisher: Elsevier BV, Elsevier, North-Holland PubCo.

Kauffman, S. A. (1993). *The origins of order*. Oxford University Press.

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431. Publisher: Proceedings of the National Academy of Sciences.

Kim, C., Batra, R., Chen, L., Tran, H., and Ramprasad, R. (2021). Polymer design using genetic algorithm and machine learning. *Computational materials science*, 186. Publisher: Elsevier BV.

Kirby, R. S., Withington, S., Darling, A. B., and Kilgour, F. G. (1956). *Engineering in history*. McGraw-Hill, New York. OCLC: 930464723.

Kogut, B. and Zander, U. (1992). Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. *Organization science (Providence, R.I.)*, 3(3):383–397. Place: Providence, RI Publisher: INFORMS.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press, Chicago. Book Title: The structure of scientific revolutions.

Kulkarni, A. J., Baki, M. F., and Chaouch, B. A. (2016). Application of the cohort-intelligence optimization method to three selected combinatorial optimization problems. *European journal of operational research*, 250(2):427–447. Publisher: Elsevier BV.

Lambiotte, R., J -C Delvenne, and Barahona, M. (2009). Laplacian Dynamics and Multiscale Modular Structure in Networks. *arXiv.org*. Place: Ithaca Publisher: Cornell University Library, arXiv.org.

Lazer, D. and Friedman, A. (2007). The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, 52(4):667–694. Publisher: SAGE Publications Inc.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

Leonard-Barton, D. and Sviokla, J. (1988). Putting Expert Systems to Work. *Harvard Business Review*, 0(March 1988). Section: IT.

Li, J., Zhao, Y., Sun, C., Bao, X., Zhao, Q., and Zhou, H. (2018). A Survey of Development and Application of Artificial Intelligence in Smart Grid. In *IOP Conference Series: Earth and Environmental Science*, volume 186, pages 12066–. IOP Publishing. ISSN: 1755-1307 Issue: 4.

Liu, T. (2020). Empirical Research on the Application of Computer Artificial Intelligence in Law. In *Journals of Physics: Conference Series*, volume 1648. IOP Publishing. ISSN: 1742-6588 Issue: 3.

Lobo, J. and Strumsky, D. (2019). Sources of inventive novelty: two patent classification schemas, same story. *Scientometrics*, 120(1):19–37. Place: Cham Publisher: unav, Springer International Publishing, Springer.

Lu, C. and Tang, X. (2014). Surpassing Human-Level Face Verification Performance on LFW with GaussianFace. *arXiv:1404.3840 [cs, stat]*. arXiv: 1404.3840.

Lund, B. D. and Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*, ahead-of-print(ahead-of-print).

Maclaurin, W. R. (1953). The Sequence from Invention to Innovation and Its Relation to Economic Growth. *The Quarterly journal of economics*, 67(1):97–111. Place: Cambridge, Mass. [etc.] Publisher: MIT Press.

Maclure, J. (2020). The new AI spring: a deflationary view. *AI & society*, 35(3):747–750.

Macready, W. G., Siapas, A. G., and Kauffman, S. A. (1996). Criticality and Parallelism in Combinatorial Optimization. *Science*, 271(5245):56–59. Publisher: American Association for the Advancement of Science Section: Reports.

Madonna (1998). Ray of Light.

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1):71–87. Publisher: INFORMS.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv:1801.00631 [cs, stat]*. arXiv: 1801.00631.

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *ArXiv*.

Matheny, M., Israni, S. T., Ahmed, M., and Whicher, D. (2019a). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Technical report, National Academy of Medicine, Washington, D.C.

Matheny, M., Israni, S. T., Ahmed, M., and Whicher, D. (2019b). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Technical report, National Academy of Medicine, Washington, D.C.

Maynard, A. (2015). Navigating the fourth industrial revolution. *Nature Nanotechnology*, 10(12):1005–1006.

Maynard, A. (2020). *Future Rising: A Journey from the Past to the Edge of Tomorrow*. Mango, Coral Gables.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McGrath, R. G. (2001). Exploratory Learning, Innovative Capacity, and Managerial Oversight. *Academy of Management Journal*, 44(1):118–131.

McLevey, J. and McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics*, 11(1):176–197.

Mewes, L. and Broekel, T. (2022). Technological complexity and economic growth of regions. *Research policy*, 51(8):104156–. Publisher: Elsevier B.V.

Miller, K. D., Zhao, M., and Calantone, R. J. (2006). Adding Interpersonal Learning and Tacit Knowledge to March's Exploration-Exploitation Model. *The Academy of Management Journal*, 49(4):709–722. Publisher: Academy of Management.

Minsky, M. L. and Papert, S. (1988). *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, Mass, expanded ed edition.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.

Monath, N., Jones, C., and Madhavan, S. (2021). PatentsView: Disambiguating Inventors, Assignees, and Locations. Technical report, American Institutes for Research.

Mozer, M. C., Wiseheart, M., and Novikoff, T. P. (2019). Artificial intelligence to support human instruction. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):3953–3955. Publisher: National Academy of Sciences Section: Commentary.

Mukherjee, S., Uzzi, B., Jones, B., and Stringer, M. (2016). A New Method for Identifying Recombinations of Existing Knowledge Associated with High-Impact Innovation. *Journal of Product Innovation Management*, 33(2):224–236.

National Academies of Sciences, Engineering, and Medicine (2021). *Accelerating Decarbonization of the U.S. Energy System*. The National Academies Press, Washington, DC.

National Research Council (1997). *Computer Science and Artificial Intelligence*. The National Academies Press, Washington, DC.

Navas, E. (2012). *Remix Theory: The Aesthetics of Sampling*. Springer Verlag, Wien, Springer, Ambra, Dordrecht, 1. aufl. edition.

Nelson, R. R. and Winter, S. G. (1982). *An evolutionary theory of economic change*. Belknap Press of Harvard University Press, Cambridge, Mass.

Newell, A. and Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3):113–126.

Newton, I. (1675). Letter from Sir Isaac Newton to Robert Hooke.

Nickles, T. (2017). Scientific Revolutions. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Niu, J., Tang, W., Xu, F., Zhou, X., and Song, Y. (2016). Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis. *ISPRS international journal of geo-information*, 5(5):66–. Publisher: MDPI AG.

NYT (1958). NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser. *New York Times*, page 25.

Oppenlaender, J. (2022). The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, Academic Mindtrek '22, pages 192–202, New York, NY, USA. Association for Computing Machinery.

Pacey, A. (1990). *Technology in world civilization: a thousand-year history*. Basil Blackwell, Oxford, [England. Book Title: Technology in world civilization : a thousand-year history.

Pareto, V., Bousquet, G.-H., and Busino, G. (1964). *Cours d'économie politique*. Droz, Genève. OCLC: 29127729.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144. Number: 7942 Publisher: Nature Publishing Group.

Peirce, C. S. (1960). *Collected papers of Charles Sanders Peirce*. Belknap Press of Harvard University Press, Cambridge ;.

Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45(1):117–136. Place: Dordrecht Publisher: Springer Science and Business Media LLC, Kluwer Academic Publishers.

Ponomarev, I. V., Williams, D. E., Lawton, B. K., Cross, D. H., Seger, Y., Schnell, J., and Haak, L. (2012). Breakthrough Paper Indicator: early detection and measurement of ground-breaking research. In *Proceedings of the 11th international conference on current research information systems*, pages 295–304, Prague, Czech Republic.

Porter, M. M. and Niksiar, P. (2018). Multidimensional mechanics: Performance mapping of natural biological systems using permutated radar charts. *PLoS ONE*, 13(9):e0204309.

Raghupathi, W. and Nerur, S. (1999). Research themes and trends in artificial intelligence: an author co-citation analysis. *Intelligence (New York, N.Y. : 1999)*, 10(2):18–23. Publisher: ACM.

Richerson, P. J., Boyd, R., and Henrich, J. (2013). The Cultural Evolution of Technology. In *Cultural Evolution: Society, Technology, Language, and Religion*. The MIT Press.

Ridley, M. (2020). *How Innovation Works: And Why It Flourishes in Freedom*. Harper, New York, 1st edition edition.

Rocchetta, S. and Mina, A. (2019). Technological coherence and the adaptive resilience of regional economies. *Regional studies*, 53(10):1421–1434. Place: Cambridge Publisher: Routledge.

Rocchetta, S., Ortega-Argilés, R., and Kogler, D. F. (2022). The non-linear effect of technological diversification on regional productivity: implications for growth and Smart Specialisation Strategies. *Regional Studies*, 56(9):1480–1495.

Roetzel, P. G. (2018). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business research (Göttingen)*, 12(2):479–522.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

Romer, P. M. (2010). What Parts of Globalization Matter for Catch-Up Growth? *The American economic review*, 100(2):94–98. Publisher: American Economic Association.

Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal*, 22(4):287–306. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.160.

Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Computational models of cognition and perception. MIT Press, Cambridge, Mass. OCLC: 12837549.

Russell, S. J. and Norvig, P. (2010). *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, N.J, 3rd ed edition. OCLC: 359890490.

Russell, S. J. and Norvig, P. (2020). *Artificial intelligence: a modern approach*. Pearson, Boston. OCLC: 1021874142.

Saltiel, J. (2019). Five years after Alice: five lessons learned from the treatment of software patents in litigation. *WIPO Magazine*, 2019(4).

Sammut, C. and Webb, G. I., editors (2011). *Encyclopedia of Machine Learning*. Springer, New York ; London, 2010 edition edition.

Schumpeter, J. A. (1934). *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle*. Harvard University Press, Cambridge, MA.

Schumpeter, J. A. (1950). *Capitalism, Socialism, and Democracy*. Harper, New York, 3d ed. edition. Book Title: Capitalism, socialism, and democracy.

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, 43(9):628–638.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences - PNAS*, 117(48):30033–30038. Place: United States Publisher: National Academy of Sciences.

Service, R. F. (2021). Protein structures for all. *Science*, 374(6574):1426–1427. Place: United States.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Shukla, A. K., Janmaijaya, M., Abraham, A., and Muhuri, P. K. (2019). Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). *Engineering applications of artificial intelligence*, 85:517–532. Publisher: Elsevier Ltd, Elsevier BV.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

State Council of the People's Republic of China (2017). New Generation of Artificial Intelligence Development Plan. Technical report, State Council of the People's Republic of China.

Strickland, E. (2022). Andrew Ng: Unbiggen AI.

Strumsky, D. and Lobo, J. (2015a). Identifying the sources of technological novelty in the process of invention. *Research policy*, 44(8):1445–1461. Place: Netherlands Publisher: Elsevier B.V.

Strumsky, D. and Lobo, J. (2015b). Identifying the sources of technological novelty in the process of invention. *Research policy*, 44:1445–1461.

Stuart, T. E. and Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17(S1):21–38. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.4250171004.

Sugimoto, C. R., Work, S., Larivière, V., and Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9):2037–2062. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23833.

Surden, H. (2019). Artificial Intelligence and Law: An Overview. *Georgia State University law review*, 35(4):1305–1337. Publisher: Georgia State University.

Szabadföldi, I. (2021). Artificial Intelligence in Military Application – Opportunities and Challenges. *Land Forces Academy review*, 26(2):157–165. Publisher: Sciendo.

Tarde, G. d. (1903). *The laws of imitation*. Henry Holt and Company, New York. Book Title: The laws of imitation.

Toole, A., Jones, C., and Madhavan, S. (2021). PatentsView: An Open Data Platform to Advance Science and Technology Policy. SSRN Scholarly Paper ID 3874213, Social Science Research Network, Rochester, NY.

Toole, A. A. and Pairolero, N. A. (2020). Adjusting to Alice: USPTO patent examination outcomes after Alice Corp. v. CLS Bank International. Technical report, USPTO.

Toole, A. A., Pairolero, N. A., Giczy, A., Forman, J. Q., Pulliam, C., Such, M., Chaki, K., Orange, D. B., Homescu, A. T., Frumkin, J., Chen, Y. Y., Gonzales, V. M., Hannon, C., Melnick, S., Nilsson, E., and Rifkin, B. M. (2020). Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents. Technical report, USPTO, Alexandria, VA.

Tran, J. L. (2016). Two years after Alice v. CLS Bank. *Journal of the Patent and Trademark Office Society*, 98(3):354–356. Publisher: Patent and Trademark Office Society.

Triulzi, G., Alstott, J., and Magee, C. L. (2020). Estimating technology performance improvement rates by mining patent data. *Technological Forecasting and Social Change*, 158:120100.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433–460. Publisher: [Oxford University Press, Mind Association].

Unger, J. M., Barlow, W. E., Tangen, C. M., Ramsey, S. D., Thompson, I. M., Klein, E. A., LeBlanc, M., Blanke, C. D., Goodman, P. J., Minasian, L. M., Nghiem, V. T., and Hershman, D. L. (2018). The scientific impact and value of large, NCI-sponsored randomized phase III cancer chemoprevention trials. *Cancer epidemiology*, 55:117–122. Place: Netherlands Publisher: Elsevier Ltd, Elsevier BV.

U.S.Code (2011). 35 U.S. Code § 103 - conditions for patentability; non-obvious subject matter.

Usher, A. P. (1954). *A history of mechanical inventions*. Harvard University Press, Cambridge. OCLC: 514178.

USPTO (2014). Preliminary Examination Instructions for Determining Subject Matter Eligibility in view of Alice Corp. v. CLS Bank. Last Modified: 2014-06-26T09:11:33-0400.

USPTO (2015a). 2202 Citation of Prior Art and Written Statements [R-08.2017]. In *Manual of Patent Examining Procedure*. USPTO.

USPTO (2015b). 905 Cooperative Patent Classification [R-07.2015]. In *Manual of Patent Examining Procedure*. USPTO.

USPTO (2019). Application of Cooperative Patent Classification (CPC) in biotechnology areas.

USPTO (2021a). General information concerning patents.

USPTO (2021b). Patent Classification.

USPTO (2022a). CPC Scheme - G03H HOLOGRAPHIC PROCESSES OR APPARATUS.

USPTO (2022b). PatentsView. Last Modified: 2022-07-01T13:42:59-0400.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Vesnic-Alujevic, L., Nascimento, S., and Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, 44(6):101961.

Wagner, A. and Rosen, W. (2014). Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. *Journal of the Royal Society interface*, 11(97):20131190–20131190. Place: England Publisher: The Royal Society.

Wang, C.-H. and Hsu, L.-C. (2014). Building exploration and exploitation in the high-tech industry: The role of relationship learning. *Technological Forecasting and Social Change*, 81:331–340.

Wang, J., Lobo, J., and Strumsky, D. (2022a). How Novel is Invention in Artificial Intelligence? Evidence from the Scientific Literature and Patenting. in revision.

Wang, J., Maynard, A., Lobo, J., Michael, K., Motsch, S., and Strumsky, D. (2022b). Knowledge Combination Analysis Reveals That Artificial Intelligence Research Is More Like "Normal Science" Than "Revolutionary Science". Manuscript submitted for publication.

Ward, J., Sypniewski, A., and Stephenson, S. (2020). Deep learning internal state index-based search and classification.

Warhol, A. (1962). Marilyn Diptych.

Web of Science (2020). Web of Science Core Collection.

Weinberg, S. (2015). *To explain the world: the discovery of modern science*. Harper, New York, first edition. edition. Book Title: To explain the world : the discovery of modern science.

Welsh, S., Barela, S., Galliott, J. C., and Michael, K. (2018). *Ethics and Security Automata: Policy and Technical Challenges of the Robotic Use of Force*. Routledge, London.

WIPO (2019). Technology Trends 2019: Artificial Intelligence. Technical report, World Intellectual Property Organization, Geneva Switzerland.

Wootton, D. (2015). *The Invention of Science: A New History of the Scientific Revolution*. Harper Collins. Google-Books-ID: 7exeBwAAQBAJ.

World Economic Forum (2018). Harnessing Artificial Intelligence for the Earth. Technical report, World Economic Forum, Geneva, Switzerland.

Wu, L., Wang, D., and Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382. Number: 7744 Publisher: Nature Publishing Group.

Youn, H., Strumsky, D., Bettencourt, L. M. A., and Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society, Interface*, 12(106):20150272–20150272.

Zabala-Iturriagagoitia, J. M., Porto Gómez, I., and Aguirre Larracoechea, U. (2020). Technological diversification: a matter of related or unrelated varieties? *Technological forecasting & social change*, 155:119997–.

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27. Place: Cham.

Ziman, J. M. (1978). *Reliable knowledge: an exploration of the grounds for belief in science*. Cambridge University Press, Cambridge. Book Title: Reliable knowledge : an exploration of the grounds for belief in science.

APPENDIX A

AI KEYWORDS SYNONYM AGGREGATION TO SEARCH FOR DATASET CONSTRUCTION

In every iteration of the snowball process for collecting the AI publication dataset (see Chapter 2), the keywords mentioned by the papers collected from previous iterations are ranked by their frequencies. Nevertheless, different keywords may have significantly similar meanings. Conducting synonym aggregation to aggregate their frequencies for a more accurate ranking would be meaningful.

This appendix provides the synonym aggregations conducted in the data collection. The synonyms are identified and aggregated heuristically and manually. After the aggregation, new frequencies are counted, and a new ranking is generated to select the candidate search terms for the next iteration.

1. 'artificial intelligence' replaces 'artificial intelligence (ai)', 'artificial intelligence techniques', 'computational and artificial intelligence', 'artificial intelligence methods', and 'ai'.

2. 'artificial neural networks' replaces 'artificial neural network', 'ann', 'anns', 'artificial neural network (ann)', 'artificial neural networks (anns)', 'artificial neural networks (ann)', 'ann model', 'ann modeling', 'artificial neural network model', 'functional link artificial neural network', 'artificial neural network modeling', 'artificial neural network modelling', 'artificial neural network analysis', 'artificial neural', 'artificial neural-network', 'ann (artificial neural network)', 'functional link artificial neural network (flann)', 'kohonen artificial neural network', 'artificial neural nets', 'functional link artificial neural networks', 'artificial neural network models', 'artificial neural networking', and 'artificial neural network(ann)'.

3. 'neural networks' replaces 'neural network', and 'nn'.

4. 'machine learning' replaces 'machine learning algorithms', 'machine learning (ml)', 'machine learning techniques', 'machine learning algorithm', 'learning (artificial intelligence)', 'machine learning methods', 'machine learning models', 'machine learning method', 'machine learning technique', 'machine learning approach', 'machine learning model', 'machine learning tools', and 'machine learning approaches'.

5. 'support vector machine' replaces 'support vector machines', 'support vector machine (svm)', 'support vector machines (svms)', 'svm', 'support vector machines (svm)', 'least squares support vector machine', 'least square support vector machine', 'least squares support vector machines', 'twin support vector machine', 'support vector machine regression', 'least-squares support vector machine', 'twin support vector machines', 'least square support vector machine (lssvm)', 'one-class support vector machine', 'lssvm', 'ls-svm', and 'svms'.

6. 'random forest' replaces 'random forests', 'random forest (rf)', and 'random forest regression'.

7. 'genetic algorithm' replaces 'genetic algorithms', 'genetic programming', 'genetic algorithm (ga)', 'genetic programming (gp)', and 'genetic algorithms (gas)'.

8. ‘prediction’ replaces ‘forecasting’, ‘predictive models’, ‘predictive model’, ‘predictive modeling’, ‘predictive analytics’, ‘predictive model’, ‘predictive modelling’, ‘prediction models’, ‘prediction methods’, and ‘prediction algorithms’.

9. ‘big data’ replaces ‘big data analytics’, ‘big data analysis’, ‘big data applications’, and ‘big data processing’.

10. ‘natural language processing’ replaces ‘nlp’, ‘natural language processing (nlp)’, ‘natural language’, ‘natural language understanding’, and ‘natural language processing (nlp)’.

11. ‘expert systems’ replaces ‘expert system’, ‘artificial intelligence, applications and expert systems’, ‘fuzzy expert system’, ‘applications and expert systems’, ‘medical expert systems’, and ‘expert knowledge’.

12. ‘recurrent neural networks’ replaces ‘rnn’, ‘rnns’, ‘recurrent neural network’, ‘recurrent neural network (rnn)’, ‘recurrent neural nets’, ‘recurrent neural networks (rnns)’, ‘recurrent network’, ‘recurrent networks’, ‘recurrent’, ‘recurrent neural networks (rnn)’, ‘recurrent artificial neural network’, ‘recurrent artificial neural networks’, ‘grnn’, and ‘lstm-rnn’.

13. ‘bayesian networks’ replaces ‘naive bayes’, ‘bayesian network’, ‘bayesian’, ‘bayes methods’, ‘naive bayes classifier’, ‘bayesian methods’, ‘bayesian learning’, ‘bayesian analysis’, ‘bayesian modeling’, ‘bayesian neural networks’, ‘bayesian belief network’, and ‘bayesian models’.

14. ‘ethics’ replaces ‘machine ethics’, ‘bioethics’, ‘ai ethics’, ‘roboethics’, ‘computer ethics’, ‘robot ethics’, ‘medical ethics’, ‘ethical issues’, ‘ethical design’, ‘digital ethics’, ‘data ethics’, ‘business ethics’, ‘ethics of ai’, ‘research ethics’, ‘technology ethics’, ‘ethical evaluation’, ‘professional ethics’, ‘morality’, ‘moral judgment’, and ‘moral psychology’.

15. ‘risk’ replaces ‘risk assessment’, ‘risk prediction’, ‘risk factors’, ‘risk management’, ‘risk analysis’, ‘risk stratification’, ‘credit risk’, ‘risk factor’, ‘empirical risk minimization’, ‘risk adjustment’, ‘risk model’, ‘credit risk assessment’, ‘risk score’, ‘existential risk’, ‘risks’, ‘structural risk minimization’, ‘value at risk’, ‘risk evaluation’, ‘risk and uncertainty’, ‘operational risk’, ‘risk perception’, ‘health risk’, and ‘financial risk’.

16. ‘internet of things’ replaces ‘iot’, ‘internet of things (iot)’, ‘iot devices’, ‘industrial internet of things’, and ‘industrial internet of things (iiot)’.

17. ‘convolutional neural network’ replaces ‘convolutional neural networks’, ‘convolutional neural network (cnn)’, ‘cnn’, ‘cnns’, ‘convolution’, ‘convolutional neural nets’, ‘convolution neural network’, ‘deep convolutional neural network’, ‘deep convolutional neural networks’, ‘convolutional neural networks (cnns)’, ‘convolutional neural networks (cnn)’, ‘convolutional networks’, ‘convolution neural networks’, ‘deconvolution’, ‘convolution neural network (cnn)’, ‘fully convolutional network’, ‘fully convolutional networks’, ‘graph convolutional networks’,

'convolutional network', 'deep convolutional neural network (dcnn)', 'convolutional', 'deep convolution neural network', 'convolutional layers', 'fully convolutional neural network', 'cnn-lstm', and 'faster r-cnn'.

18. 'clustering' replaces 'cluster analysis', 'k-means clustering', 'fuzzy clustering', 'data clustering', 'clustering algorithms', 'clustering analysis', 'pattern clustering', 'clustering methods', 'clustering algorithm', 'cluster', 'k-means clustering algorithm', 'cluster computing', 'clustering techniques', 'k-means', 'k-means algorithm', 'kernel k-means', 'k-mean', 'k-means method', and 'fuzzy k-means'.

19. 'computer vision' replaces 'machine vision', 'vision', 'robot vision', 'computer vision system'.

20. 'robotics' replaces 'robots', 'human-robot interaction', 'robot', 'mobile robots', 'evolutionary robotics', 'mobile robot', 'autonomous robots', 'cognitive robotics', 'robot control', 'rehabilitation robotics', 'mobile robotics', 'intelligent robots', 'humanoid robots', and 'robot learning'.

21. 'image processing' replaces 'image analysis', 'image segmentation', 'image recognition', 'medical image processing', 'image representation', and 'image retrieval'.

22. 'deep learning' replaces 'deep neural network', 'deep neural networks', 'deep belief network'.

23. 'reinforcement learning' replaces 'deep reinforcement learning', 'reinforcement', 'reinforcement learning (rl)', 'deep reinforcement learning (drl)', 'inverse reinforcement learning', and 'hierarchical reinforcement learning'.

24. 'multilayer perceptron' replaces 'multi-layer perceptron', 'multilayer perceptrons', 'multilayer perceptron (mlp)', 'perceptron', 'multi layer perceptron', 'multi-layer perceptron (mlp)', 'multilayer perceptron neural network', 'multilayer perceptron network', 'multi-layer perceptrons', 'multi-layer perceptron neural network', 'multi-layer perceptron network', 'multi layer perceptron (mlp)', 'multilayer perceptrons (mlps)', 'multilayered perceptron', 'perceptrons', 'multilayer perceptron neural networks', 'multilayer perceptron networks', 'multi-layer perceptron neural networks', 'multilayer perceptron model', 'multi-layered perceptron', 'multilayer perceptron (mlp) neural network', 'three-layer perceptron', 'single-layer perceptron', 'perceptron neural network', 'mlps', and 'mlpr'.

25. 'privacy' replaces 'differential privacy', 'data privacy', 'privacy preservation', 'privacy protection', 'privacy preserving'.

26. 'principal component analysis' ['pca', 'principal component analysis (pca)', 'principal components analysis', 'principal component regression', 'principal components', 'kernel principal component analysis', 'principal component regression (pcr)', 'principal component', 'principal components analysis (pca)', 'robust principal component analysis', and 'robust pca'.

27. 'logistic regression' replaces 'logistic regression (lr)', 'multinomial logistic regression', 'kernel logistic regression', 'logistic models', and 'logistic regression analysis'.

28. ‘gradient descent’ replaces ‘stochastic gradient descent’.

29. ‘backpropagation’ replaces ‘back propagation’, ‘back-propagation’, ‘back propagation neural network’, ‘back-propagation neural network’, ‘back-propagation algorithm’, ‘back propagation algorithm’, ‘backpropagation algorithm’, ‘back propagation artificial neural network’, ‘backpropagation neural network’, ‘back-propagation artificial neural network’, ‘back-propagation network’, ‘back-propagation neural networks’, ‘backpropagation neural networks’, ‘back propagation network’, ‘back propagation neural network (bpnn)’, ‘back propagation (bp)’.

30. ‘gradient boosting’ replaces ‘extreme gradient boosting’, ‘stochastic gradient boosting’, ‘gradient boosting decision tree’, ‘gradient boosting machine’, ‘gradient boosting machines’, ‘gradient boosted trees’, ‘gradient boosting trees’, and ‘extreme gradient boosting (xgboost)’.

31. ‘generative adversarial networks’ replaces ‘gan’, ‘generative adversarial network’, ‘generative models’, and ‘generative model’.

32. ‘supervised learning’ replaces ‘supervised machine learning’, and ‘unsupervised’.

33. ‘unsupervised learning’ replaces ‘unsupervised machine learning’, and ‘unsupervised’.

34. ‘semi-supervised learning’ replaces ‘semisupervised learning’, ‘semi-supervised’, and ‘semi-supervised machine learning’.

APPENDIX B

AI KEYWORDS USED IN THE SECOND ITERATION OF SNOWBALL SEARCH

1. support vector machine
2. deep learning
3. genetic algorithm
4. random forest
5. convolutional neural network
6. perceptron
7. backpropagation
8. natural language processing
9. supervised learning
10. unsupervised learning
11. semi-supervised learning
12. reinforcement learning
13. recurrent neural networks

APPENDIX C

THE TOP 20 KEYWORDS AFTER TWO ITERATIONS OF SNOWBALL SEARCHING

1. machine learning
2. artificial neural networks
3. artificial intelligence
4. neural networks
5. support vector machine
6. deep learning
7. prediction
8. genetic algorithm
9. classification
10. data mining
11. random forest
12. convolutional neural network
13. clustering
14. feature selection
15. image processing
16. optimization
17. feature extraction
18. big data
19. pattern recognition
20. principal component analysis

APPENDIX D

CONDITIONAL AND UNCONDITIONAL MEAN VALUES AND VARIANCES OF CITATION COUNTS OF THE FOUR CATEGORIES OF AI PUBLICATIONS

Category	Mean	Var
Accepted Wisdom	22.07	18547.32
Avant Garde	15.93	3696.38
Darwin's Tower	28.41	54161.42
Platypus	12.33	18391.57
Unconditional	19.95	17516.07

APPENDIX E

SUMMARY OF NEGATIVE BINOMIAL REGRESSION ON THE KNOWLEDGE RECOMBINATION TAXONOMY AND AI PUBLICATIONS' CITATION COUNTS

Dep. Variable:	Z9	**No. Observations:**	296378
Model:	GLM	**Df Residuals:**	296373
Model Family:	NegativeBinomial	**Df Model:**	4
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1.0917e+06
Date:	Tue, 20 Sep 2022	**Deviance:**	6.1970e+05
Time:	14:04:55	**Pearson chi2:**	9.70e+06
No. Iterations:	17	**Pseudo R-squ. (CS):**	0.4877
Covariance Type:	nonrobust		

	coef	**std err**	z	$P > \|z\|$	**[0.025**	**0.975]**
Intercept	317.7754	0.621	512.108	0.000	316.559	318.992
category_20_80_cumulative[T.Avant Garde]	-0.5230	0.004	-120.350	0.000	-0.532	-0.515
category_20_80_cumulative[T.Darwin's Tower]	-0.2939	0.006	-46.170	0.000	-0.306	-0.281
category_20_80_cumulative[T.Platypus]	-0.4640	0.008	-59.943	0.000	-0.479	-0.449
PY	-0.1564	0.000	-507.582	0.000	-0.157	-0.156

APPENDIX F

OLS REGRESSION RESULTS OF NORMALIZED LOGARITHM OF THE NUMBER OF AI PUBLICATIONS

Dep. Variable:	normalized_annual_count_ln	**R-squared:**	0.967
Model:	OLS	**Adj. R-squared:**	0.967
Method:	Least Squares	**F-statistic:**	1869.
Date:	Thu, 09 Feb 2023	**Prob (F-statistic):**	1.52e-48
Time:	14:59:40	**Log-Likelihood:**	-60.289
No. Observations:	65	**AIC:**	124.6
Df Residuals:	63	**BIC:**	128.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	t	$P > \|t\|$	**[0.025**	**0.975]**
Intercept	-339.8149	7.982	-42.575	0.000	-355.765	-323.865
year	0.1736	0.004	43.226	0.000	0.166	0.182

Omnibus:	3.411	**Durbin-Watson:**	0.554
Prob(Omnibus):	0.182	**Jarque-Bera (JB):**	0.361
Skew:	0.182	**Prob(JB):**	0.186
Kurtosis:	4.105	**Cond. No.**	2.06e+05

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 2.06e+05. This might indicate that there are strong multicollinearity or other numerical problems.

APPENDIX G

OLS REGRESSION RESULTS OF NORMALIZED LOGARITHM OF THE NUMBER OF AI PATENTS

Dep. Variable:	normalized_annual_count_ln	**R-squared:**	0.971
Model:	OLS	**Adj. R-squared:**	0.971
Method:	Least Squares	**F-statistic:**	1455.
Date:	Thu, 09 Feb 2023	**Prob (F-statistic):**	8.66e-35
Time:	15:20:37	**Log-Likelihood:**	-16.099
No. Observations:	45	**AIC:**	36.20
Df Residuals:	43	**BIC:**	39.81
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	t	$P > \|t\|$	**[0.025**	**0.975]**
Intercept	-306.1568	8.119	-37.710	0.000	-322.530	-289.784
year	0.1550	0.004	38.138	0.000	0.147	0.163

Omnibus:	0.549	**Durbin-Watson:**	0.279
Prob(Omnibus):	0.760	**Jarque-Bera (JB):**	0.652
Skew:	-0.076	**Prob(JB):**	0.722
Kurtosis:	2.430	**Cond. No.**	3.07e+05

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 3.07e+05. This might indicate that there are strong multicollinearity or other numerical problems.

APPENDIX H

OLS REGRESSION RESULTS OF NORMALIZED LOGARITHM OF THE NUMBER OF SCIENTIFIC PUBLICATIONS

Dep. Variable:	normalized_annual_count_ln	**R-squared:**	0.949
Model:	OLS	**Adj. R-squared:**	0.949
Method:	Least Squares	**F-statistic:**	1372.
Date:	Thu, 09 Feb 2023	**Prob (F-statistic):**	4.50e-49
Time:	15:31:04	**Log-Likelihood:**	0.061771
No. Observations:	75	**AIC:**	3.876
Df Residuals:	73	**BIC:**	8.511
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	t	$P > \|t\|$	**[0.025**	**0.975]**
Intercept	-93.7698	2.592	-36.175	0.000	-98.936	-88.604
year	0.0484	0.001	37.044	0.000	0.046	0.051

Omnibus:	19.057	**Durbin-Watson:**	0.047
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5.937
Skew:	0.386	**Prob(JB):**	0.0514
Kurtosis:	1.859	**Cond. No.**	1.82e+05

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 1.82e+05. This might indicate that there are strong multicollinearity or other numerical problems.

APPENDIX I

OLS REGRESSION RESULTS OF NORMALIZED LOGARITHM OF THE NUMBER OF UTILITY PATENTS

Dep. Variable:	normalized_annual_count_ln	**R-squared:**	0.888
Model:	OLS	**Adj. R-squared:**	0.886
Method:	Least Squares	**F-statistic:**	442.2
Date:	Thu, 09 Feb 2023	**Prob (F-statistic):**	2.98e-28
Time:	15:35:59	**Log-Likelihood:**	12.091
No. Observations:	58	**AIC:**	-20.18
Df Residuals:	56	**BIC:**	-16.06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	t	$P > \|t\|$	**[0.025**	**0.975]**
Intercept	-74.7722	3.123	-20.741	0.000	-71.028	-58.516
year	0.0330	0.002	21.028	0.000	0.030	0.036

Omnibus:	8.687	**Durbin-Watson:**	0.295
Prob(Omnibus):	0.013	**Jarque-Bera (JB):**	2.888
Skew:	-0.127	**Prob(JB):**	0.236
Kurtosis:	1.937	**Cond. No.**	2.37e+05

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 2.37e+05. This might indicate that there are strong multicollinearity or other numerical problems.

APPENDIX J

TEAM SIZE DISTRIBUTION IN AI PATENTS

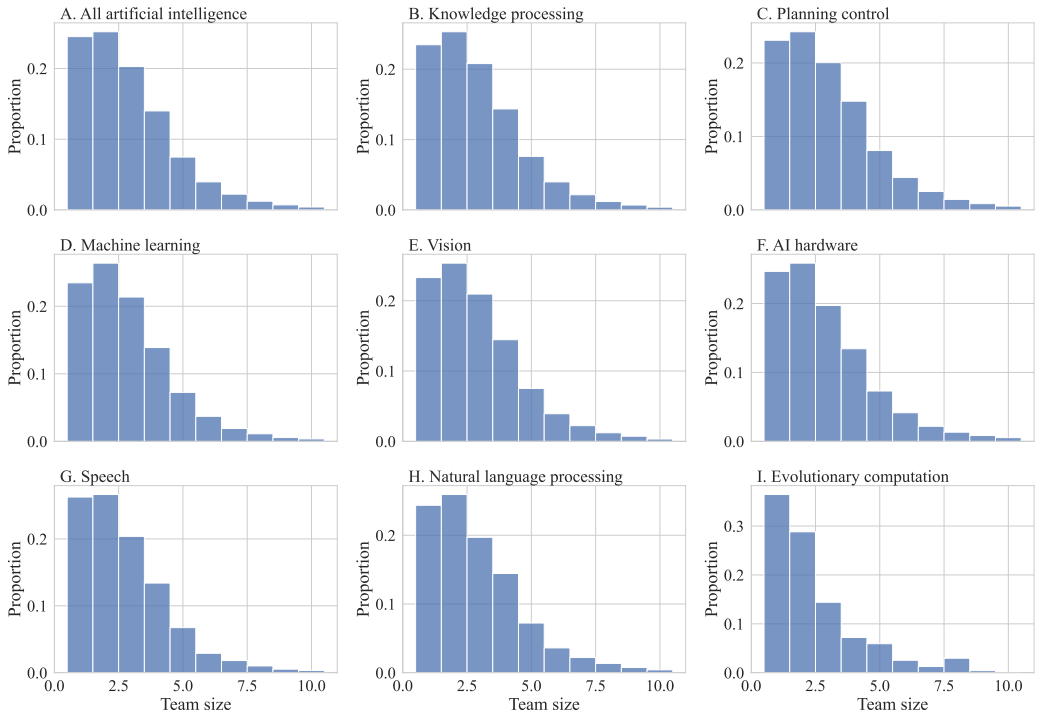


Figure J.1: Team size distribution in patents in AI and its subdomains.

AI subdomain	count	mean	std	min	25%	median	75%	max
any_ai	257208.0	2.970378	2.024852	1.0	2.0	3.0	4.0	65.0
evo	237.0	2.514768	1.867653	1.0	1.0	2.0	3.0	12.0
hardware	29105.0	2.985501	2.085200	1.0	2.0	2.0	4.0	53.0
kr	156614.0	2.982186	1.975265	1.0	2.0	3.0	4.0	43.0
ml	41078.0	2.921345	1.921791	1.0	2.0	3.0	4.0	61.0
nlp	13822.0	2.946317	1.967101	1.0	2.0	3.0	4.0	26.0
planning	126753.0	3.101639	2.150067	1.0	2.0	3.0	4.0	65.0
speech	18153.0	2.789732	1.838008	1.0	1.0	2.0	4.0	26.0
vision	41269.0	2.986624	1.962776	1.0	2.0	3.0	4.0	26.0

Table J.1: Statistical summary of team sizes in AI subdomains

APPENDIX K

PERCENTILE COMPOSITIONS OF CITATION AND DISRUPTION BY TEAM SIZES

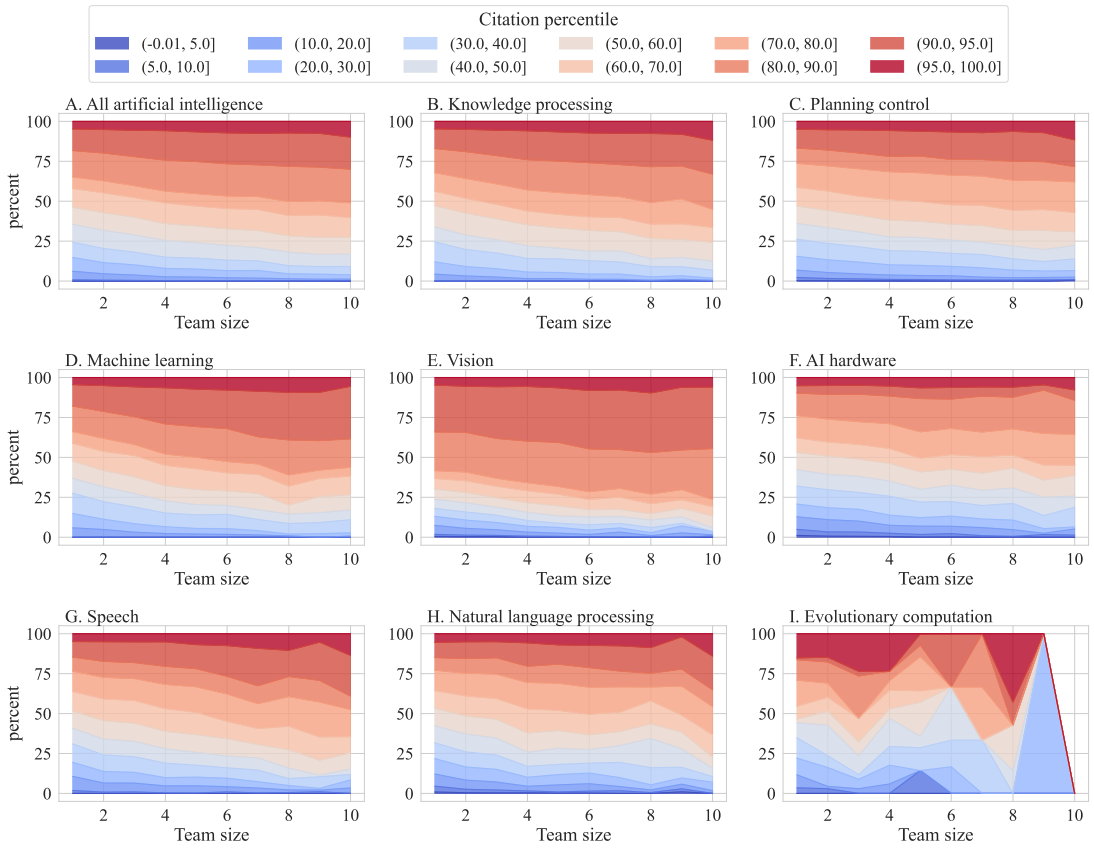


Figure K.1: Citation percentile composition by team size

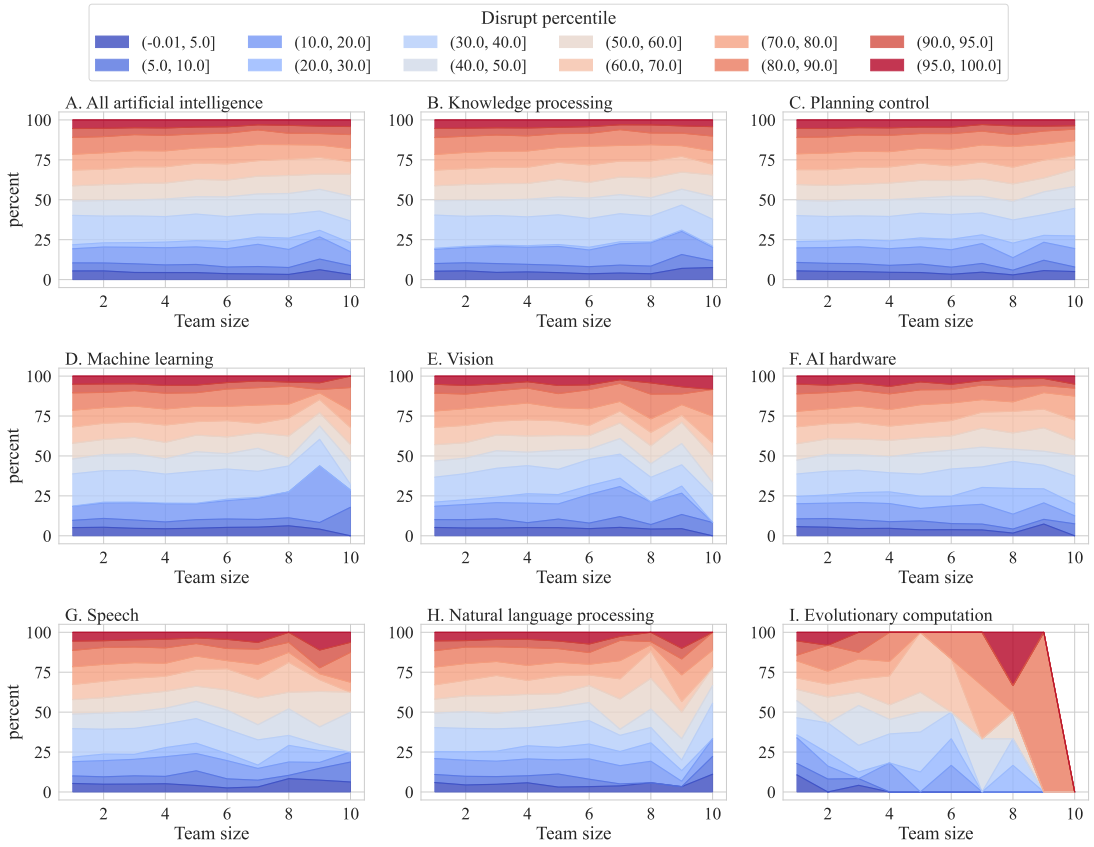


Figure K.2: Disruption percentile composition by team size

APPENDIX L

SUMMARY STATISTICS OF TEAM SIZES IN AI SUBDOMAINS BY SOURCES OF TECHNOLOGICAL NOVELTY

AI subdomain	Source of Novelty	Count	Mean	std	min	25%	median	75%	max
any_ai	combination	103433.00	3.01	2.03	1.00	2.00	3.00	4.00	53.00
	novel combination	22061.00	2.77	1.94	1.00	1.00	2.00	4.00	61.00
	origination	2535.00	2.21	1.57	1.00	1.00	2.00	3.00	20.00
	refinement	129179.00	2.98	2.04	1.00	2.00	3.00	4.00	65.00
evo	combination	15.00	1.93	0.96	1.00	1.00	2.00	2.50	4.00
	novel combination	82.00	2.51	2.12	1.00	1.00	2.00	3.00	12.00
	origination	41.00	3.00	2.04	1.00	1.00	2.00	4.00	8.00
	refinement	99.00	2.40	1.65	1.00	1.00	2.00	3.00	9.00
hardware	combination	11182.00	3.09	2.16	1.00	2.00	3.00	4.00	53.00
	novel combination	5146.00	2.87	2.03	1.00	1.00	2.00	4.00	24.00
	origination	814.00	2.29	1.68	1.00	1.00	2.00	3.00	16.00
	refinement	11963.00	2.99	2.06	1.00	2.00	3.00	4.00	26.00
kr	combination	60815.00	3.05	2.00	1.00	2.00	3.00	4.00	36.00
	novel combination	15923.00	2.84	1.90	1.00	1.00	2.00	4.00	34.00
	origination	2037.00	2.27	1.60	1.00	1.00	2.00	3.00	23.00
	refinement	77839.00	2.98	1.98	1.00	2.00	3.00	4.00	43.00
ml	combination	15021.00	3.08	1.95	1.00	2.00	3.00	4.00	18.00
	novel combination	8702.00	2.86	2.02	1.00	1.00	2.00	4.00	61.00
	origination	1357.00	2.34	1.68	1.00	1.00	2.00	3.00	20.00
	refinement	15998.00	2.86	1.85	1.00	2.00	2.00	4.00	22.00
nlp	combination	4686.00	3.02	1.98	1.00	2.00	3.00	4.00	26.00
	novel combination	1716.00	2.94	2.08	1.00	1.00	2.00	4.00	15.00
	origination	270.00	2.31	1.64	1.00	1.00	2.00	3.00	13.00
	refinement	7150.00	2.92	1.94	1.00	2.00	2.00	4.00	26.00
planning	combination	47473.00	3.14	2.12	1.00	2.00	3.00	4.00	41.00
	novel combination	11295.00	2.93	2.14	1.00	1.00	2.00	4.00	53.00
	origination	1249.00	2.31	1.54	1.00	1.00	2.00	3.00	16.00
	refinement	66736.00	3.12	2.18	1.00	2.00	3.00	4.00	65.00
speech	combination	6025.00	2.88	1.92	1.00	1.00	2.00	4.00	26.00
	novel combination	2557.00	2.71	1.80	1.00	1.00	2.00	4.00	21.00
	origination	346.00	2.15	1.47	1.00	1.00	2.00	3.00	13.00
	refinement	9225.00	2.77	1.80	1.00	1.00	2.00	4.00	26.00
vision	combination	21484.00	3.04	2.01	1.00	2.00	3.00	4.00	25.00
	novel combination	8377.00	2.97	1.96	1.00	2.00	3.00	4.00	21.00
	origination	703.00	2.36	1.50	1.00	1.00	2.00	3.00	11.00
	refinement	10705.00	2.92	1.88	1.00	2.00	3.00	4.00	26.00

Table L.1: Summary statistics of team sizes in AI subdomains by sources of technological novelty

APPENDIX M

NUMBER AND PERCENT OF AI PATENTS BY DIFFERENT TEAM SIZES AND SOURCES OF NOVELTY

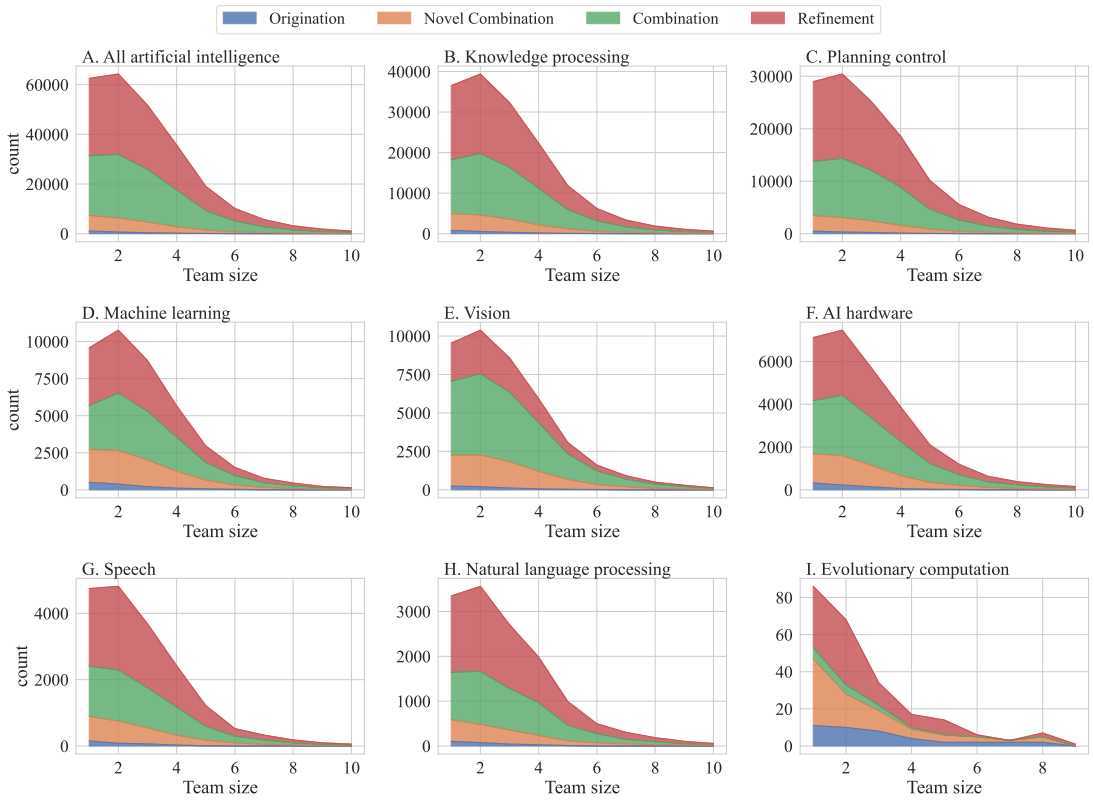


Figure M.1: Number of AI patents by different team sizes categorized by sources of technological novelty

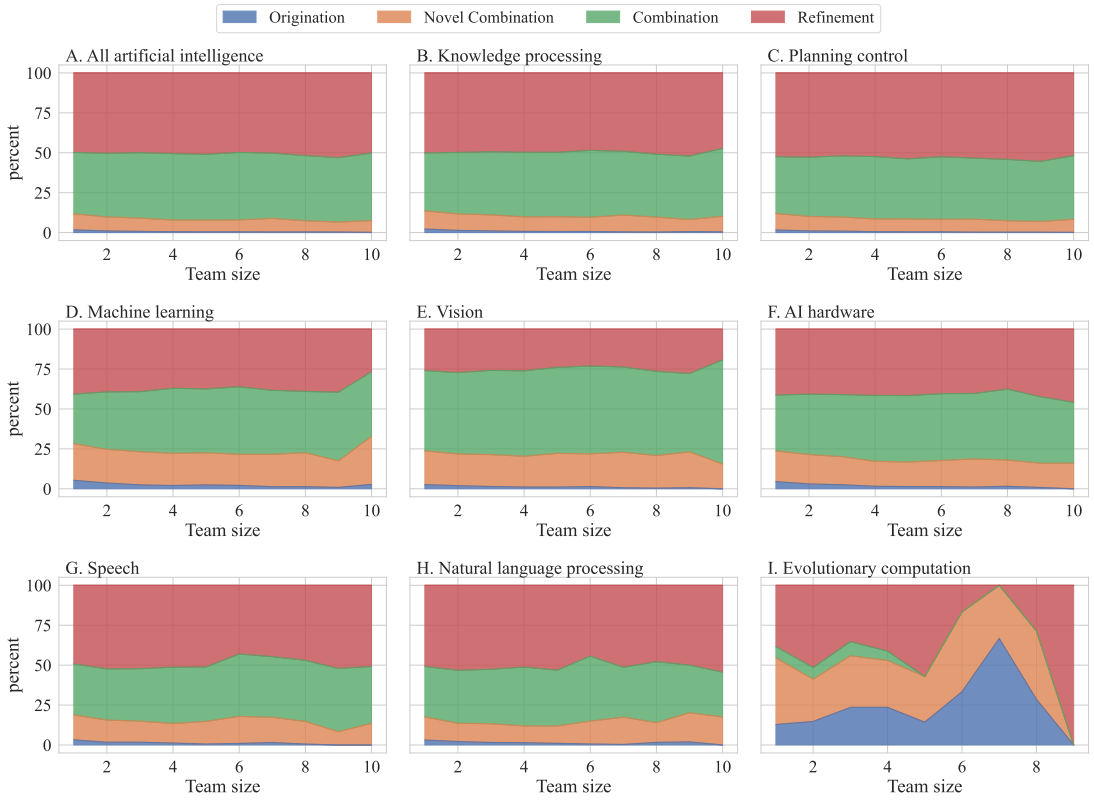


Figure M.2: Percent of AI patents by different team sizes categorized by sources of technological novelty

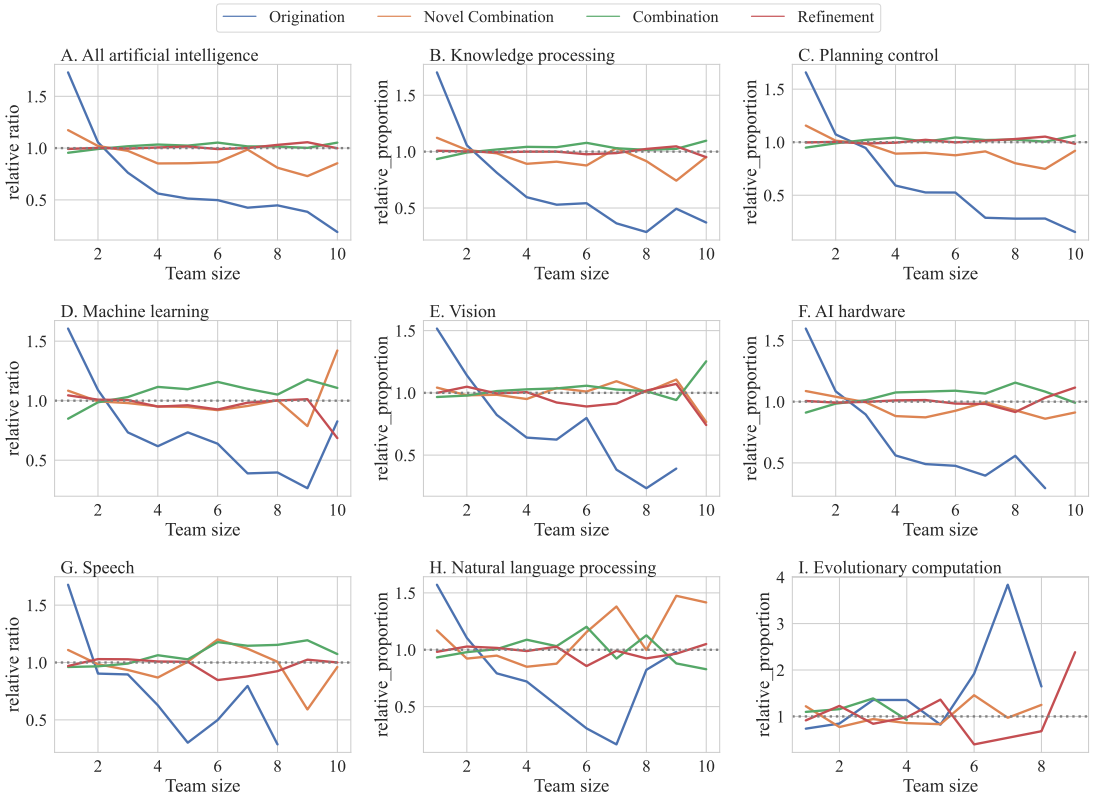


Figure M.3: Relative ratio of the percentage of AI patents categorized by technological novelty. The dotted gray lines in the panels represent a fictional relative ratio equal to one, indicating the percentage of a category if patents are evenly distributed across team sizes.

APPENDIX N

LARGER TEAMS, MORE TECHNICAL COMPONENTS AND MORE FIELDS INVOLVED IN AI INVENTIONS

This section presents the relationship between the team size and the number of technology codes (also referred to as n-tuples) in AI patents. Figures N.1 and N.2 illustrate that the larger the team is, the more technical components and fields are involved in the patents produced by the team. In other words, when it comes to invention, there is strength in numbers.

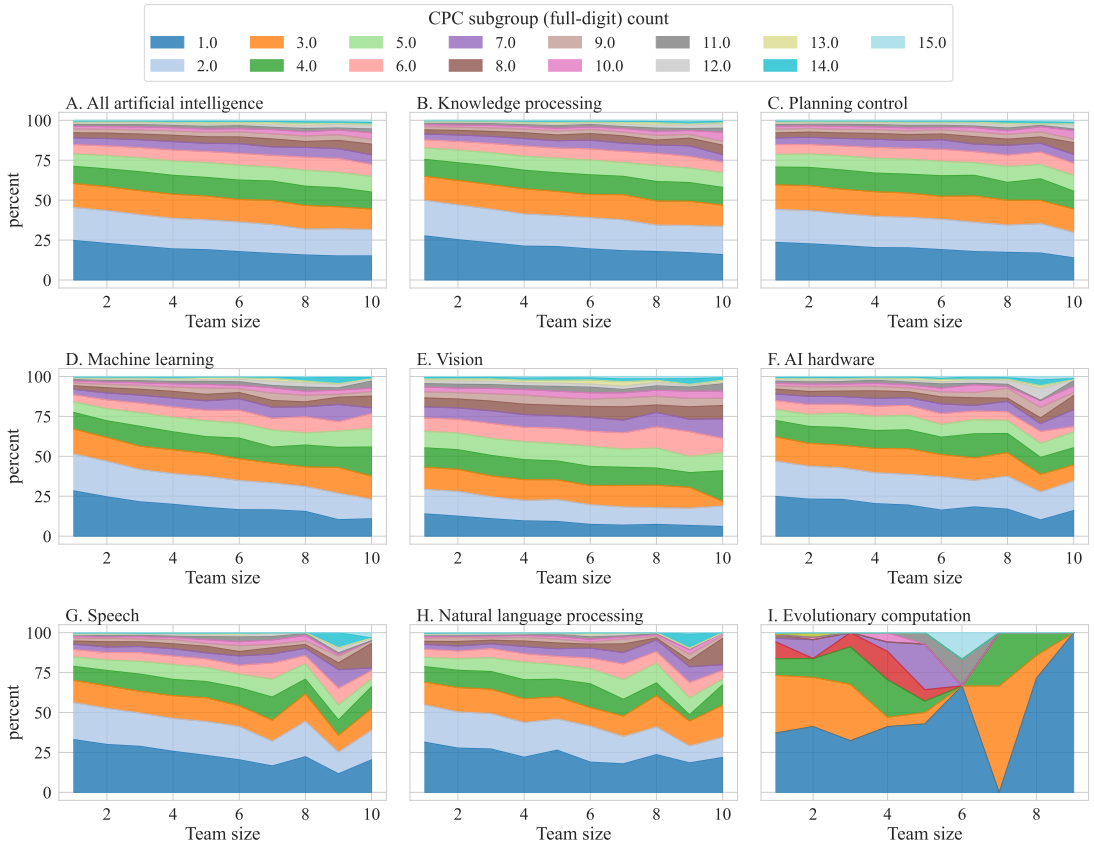


Figure N.1: Percentage of AI patents with different number of CPC subgroups (full-digit codes) by team size.

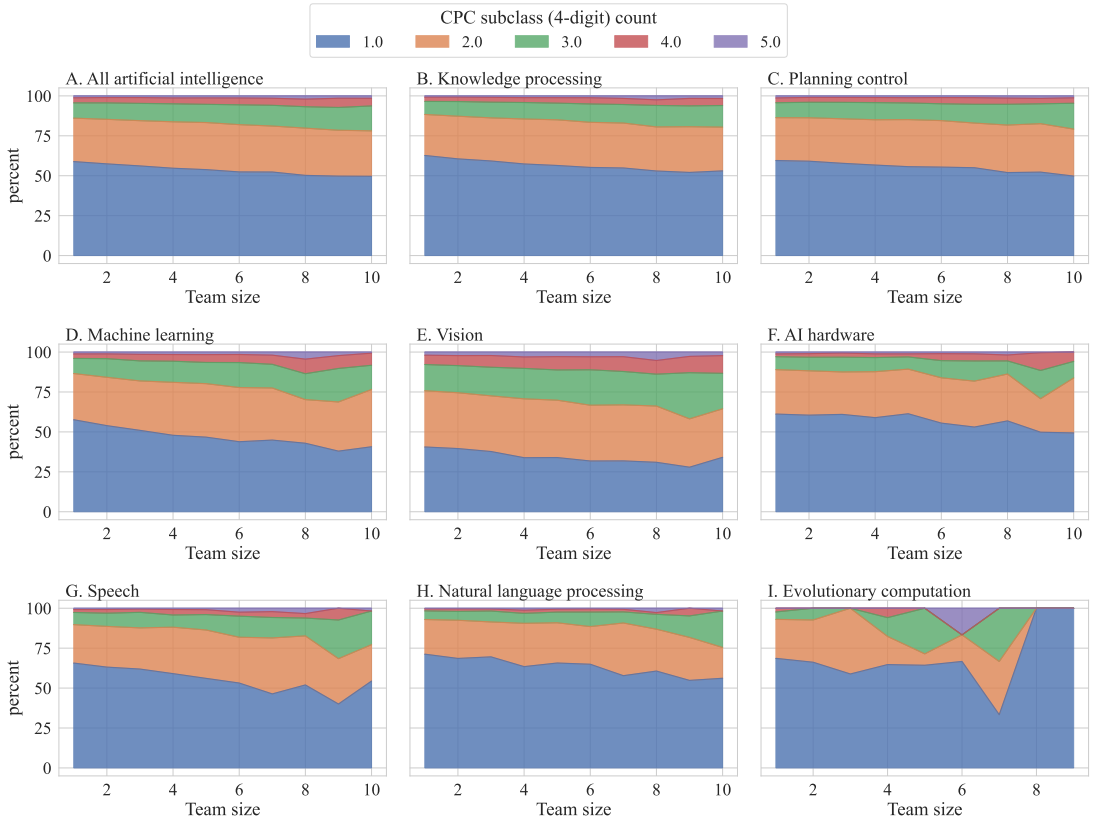


Figure N.2: Percentage of AI patents with different number of CPC subclasses (4-digit codes) by team size.

APPENDIX O

DISTRIBUTION OF AI PATENTS' CPC SUBCLASSES AND SUBGROUPS

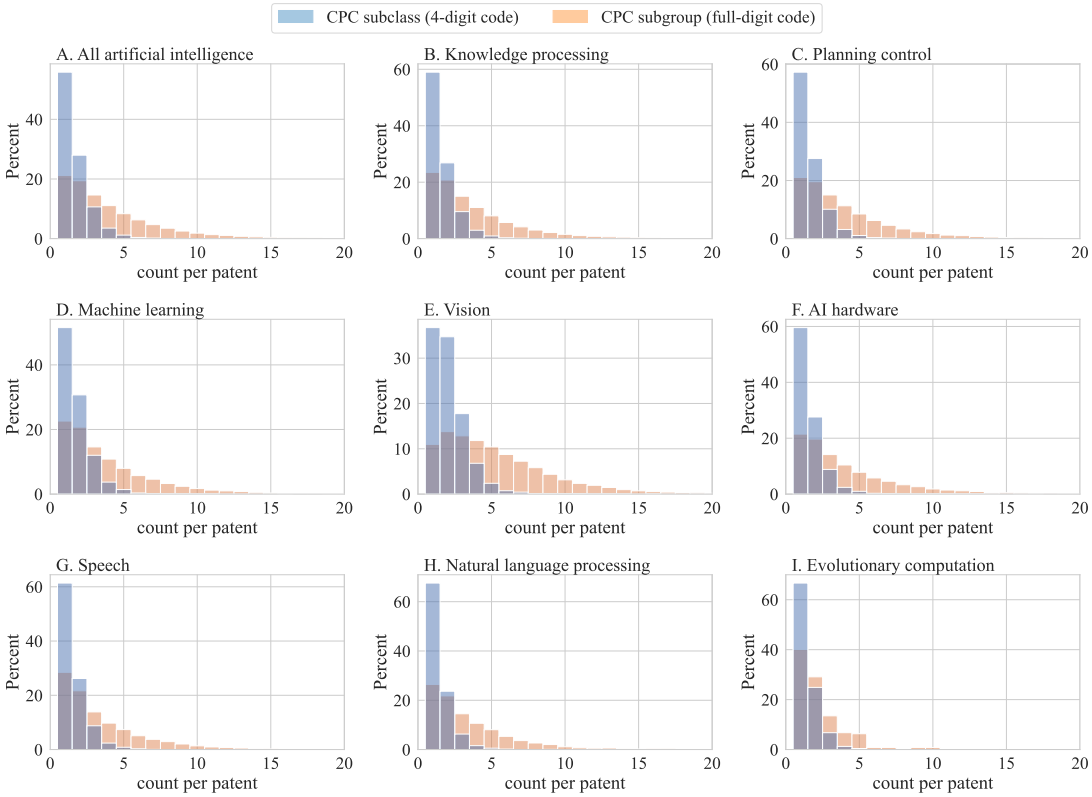


Figure O.1: Distribution of AI patents' CPC subclasses and subgroups

APPENDIX P

FULL RANGE HIT RATE OF AI PATENTS OR PUBLICATIONS COMPARED TO BACKGROUND

The following Figure P.1 illustrates the percentage of each category in AI patents or AI publications that is ranked in the top percentiles in terms of annual citations. The x-axes represent the top percentiles in terms of annual citations or the so-called background percentages, and the y-axes represent the percentage of each category illustrated as lines in different colors that are ranked in the corresponding top percentiles in terms of citations. For instance, in Panel D, the red line represents origination publications that reference only new journals. There is a point on the red line with the coordinate (0.8, 0.6). It indicates that 60% of origination AI publications are ranked in the top 80% percent in terms of annual citations. 60% is smaller than 80%, indicating that origination publications have a lower chance of becoming top 80% in terms of citations.

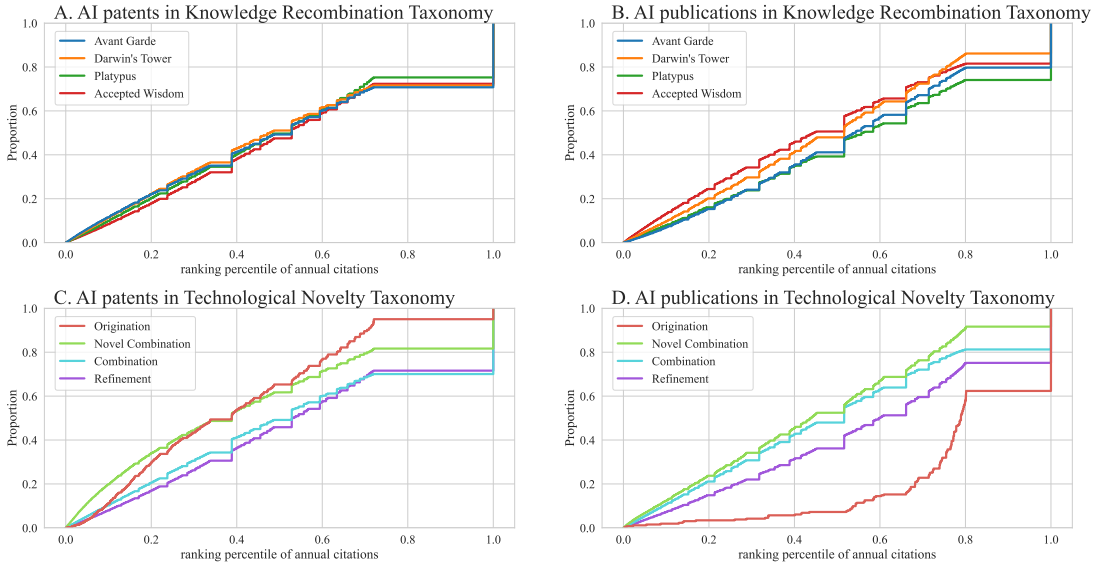


Figure P.1: Full range hit rate of AI patents and publications compared to the background. The x-axis in each panel represents the background hit rate in terms of average annual citations. The y-axis in each panel represents the hit rate in each category. A line higher than $y = x$ indicates that the category such a line represents has a higher hit rate than the background.

APPENDIX Q

SUMMARY STATISTICS OF THE SEVEN MAIN TYPES OF AI ASSIGNEES

The following Table Q.1 presents the summary statistics of the seven main types of assignees in AI patenting. As the table shows, U.S. companies or corporations own the largest share of AI patents (71%). Nevertheless, U.S. federal government agencies on average own the most patents (31 AI patents each).

assignee type	**count**	**%count**	**patent count**	**%patents**	**average patents**	**median patents**	**std**	**CV(%)**
US Company or Corporation	17,677	63.58%	181,143	70.51%	10.25	1	238.15	2324
Foreign Company or Corporation	8,911	32.05%	65,607	25.54%	7.36	1	63.75	866
US Individual	660	2.37%	870	0.34%	1.32	1	0.95	72
Foreign Individual	318	1.14%	406	0.16%	1.28	1	1.36	106
US Federal Government	37	0.13%	1,138	0.44%	30.76	3	86.76	282
Foreign Government	37	0.13%	284	0.11%	7.68	1	15.36	200
US State Government	1	0.003%	7	0.003%	7	7	n/a	n/a

Table Q.1: Summary statistics of the seven main types of AI assignees

APPENDIX R

SUMMARY STATISTICS OF THE NUMBER OF CPC SECTIONS, CPC SUBSECTIONS, CPC GROUPS, AND CPC SUBGROUPS OF AI PATENTS

	count	mean	std	min	25%	50%	75%	max
CPC section count	256,892	1.38	0.57	1	1	1	2	6
CPC subsection count	256,892	1.56	0.84	1	1	1	2	17
CPC group count	256,892	1.90	1.21	1	1	2	2	30
CPC subgroup count	256,892	5.84	6.70	1	2	4	7	157

Table R.1: Summary statistics of four levels of CPC codes

APPENDIX S

DISTRIBUTION OF THE NUMBER OF CPC SECTIONS, CPC SUB-SECTIONS, CPC GROUPS, AND CPC SUBGROUPS OF EACH AI PATENT

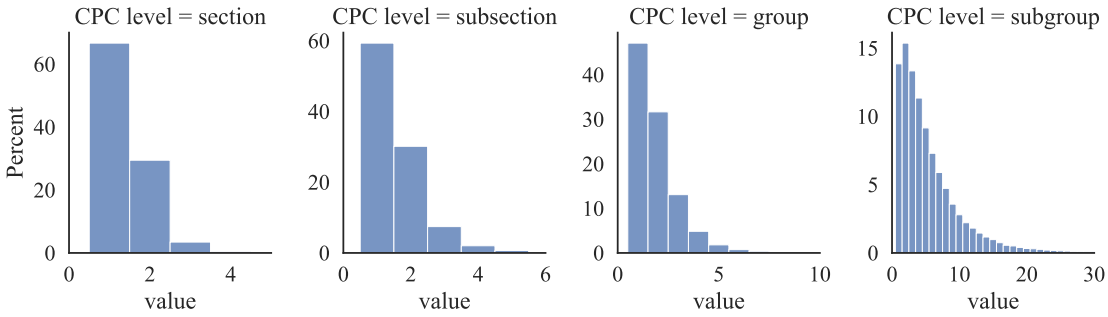


Figure S.1: Distribution of the number of CPC sections, CPC sub-sections, CPC groups, and CPC subgroups of each AI patent

APPENDIX T

TIME SERIES OF THE NUMBER OF ASSIGNEES, NUMBER OF NEW ASSIGNEES, AND PERCENT OF NEW ASSIGNEES IN FIVE TECHNOLOGY DOMAINS

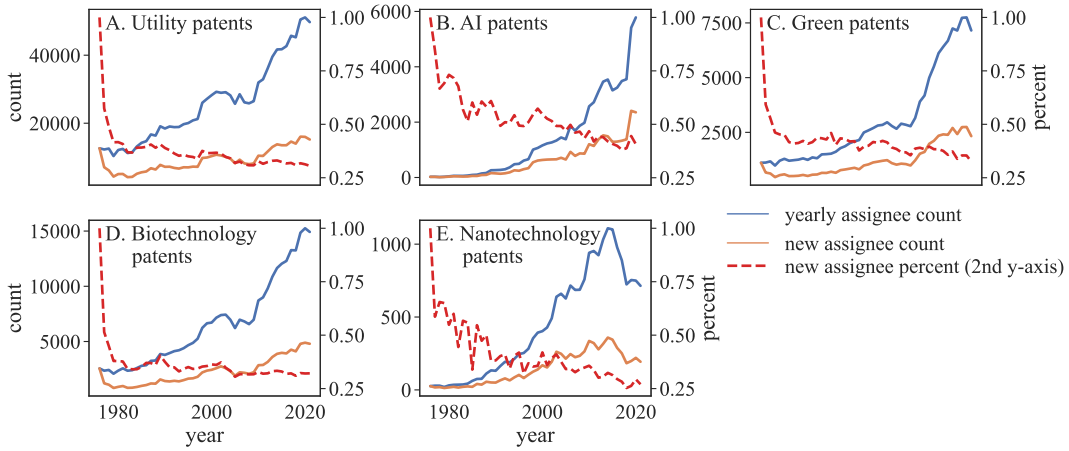


Figure T.1: Time series of the number of assignees (blue), number of new assignees (orange), and percent of new assignees in five technology domains (dashed red line, measured by the right y-axes).

APPENDIX U

SUMMARY STATISTICS OF SERIAL AI ASSIGNEES' SELECTED VARIABLES

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Active years ($AY_i = \max_{\forall p \in P_i} t_p - \min_{\forall p \in P_i} t_p + 1$)	482	15.2	10.1	4	7	12	21	45
Total patent count ($\lvert P_i \rvert$)	482	337.3	1,430.7	16	49	104.5	239.5	26,609
Annual patent count ($\lvert P_i \rvert / AY_i$)	482	17.7	44.2	4	5.2	7.8	15.3	591.3
Distinct CPC count ($\lvert TC_i \rvert$)	482	501.5	736.7	12	136.3	263.5	540.8	7,990
Distinct CPC pair count ($\lvert TCP_i \rvert$)	482	4,439.7	10,568.6	45	723	1,656.5	4,538.8	150,655
Total CPC count ($\sum_{\forall p \in P_i} \lvert TC_p \rvert$)	482	1,857.2	6,587.3	49	342	667	1,509.5	117,544
Total CPC pair count ($\sum_{\forall p \in P_i} \lvert TCP_p \rvert$)	482	11,843.2	34,101.9	64	1,210.8	2,861.5	7,869.3	410,648

Table U.1: Summary Statistics of Serial AI Assignees' Selected Variables

APPENDIX V

SELECTED VARIABLES OF TOP 10 AI ASSIGNEES

Organization	IBM	Microsoft	Google	Amazon	HP	Oracle	Samsung	Facebook	Intel	Apple
First year	1976	1994	2003	2006	1979	2002	1991	2010	1979	2007
Active years	45	27	18	15	42	19	30	11	42	14
Patent count	26,609	13,023	7,526	2,977	2,640	2,639	2,289	2,140	1,991	1,976
Average patent per year	591	482	418	198	63	139	76	195	47	141
Distinct CPC count	7,990	5,525	4,490	3,085	3,278	1,781	4,077	1,992	3,561	3,029
Distinct CPC pair count	150,655	100,465	87,863	38,072	21,222	20,292	45,597	31,119	32,983	35,059
Total CPC count	117,544	58,959	44,755	16,566	9,600	11,689	14,227	15,924	11,158	14,136
Total CPC pair count	410,648	273,599	268,822	62,826	25,590	44,255	65,697	107,191	50,178	83,821

Table V.1: Selected variables of top 10 AI assignees.

APPENDIX W

SUMMARY STATISTICS OF CEE PARAMETERS OF SERIAL AI ASSIGNEES IN 2020

	count	mean	std	min	25%	50%	75%	max	unbiased skew
patent access parameter	380	0.38	0.1	0.18	0.30	0.37	0.44	0.82	0.69
technology code access parameter	380	0.39	0.07	0.22	0.34	0.38	0.43	0.62	0.28
code pair access parameter	380	0.51	0.09	0.30	0.45	0.51	0.58	0.82	0.18
patent grant parameter	380	0.29	0.12	0.07	0.22	0.28	0.36	0.77	0.53
local exploitation parameter	380	0.50	0.18	0	0.42	0.52	0.61	1.00	-0.61
local exploration parameter	380	0.29	0.11	0	0.24	0.32	0.37	0.47	-0.71
global exploitation parameter	380	0.22	0.09	0	0.17	0.24	0.28	0.37	-0.53
global exploration parameter	380	0.09	0.11	0	0	0	0.18	0.35	0.94

Table W.1: Summary statistics of the eight knowledge access and discovery parameters of serial AI assignees in 2020.

APPENDIX X

DISTRIBUTION OF THE ABSOLUTE COUNTS OF VARIABLES OF KNOWLEDGE ACCESS AND DISCOVERY OF SERIAL AI ASSIGNEES IN 2020

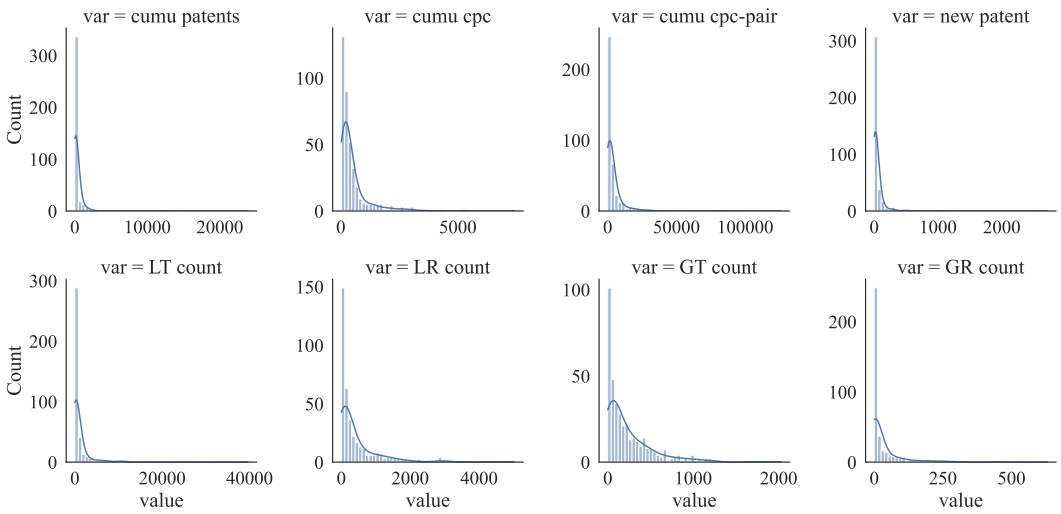


Figure X.1: Distribution of the knowledge access and discovery variables of serial AI assignees in 2020. It is notable that those variables (counts) are very skewed.

APPENDIX Y

PAIR DISTRIBUTION OF CEE PARAMETERS OF SERIAL AI ASSIGNEES IN 2020

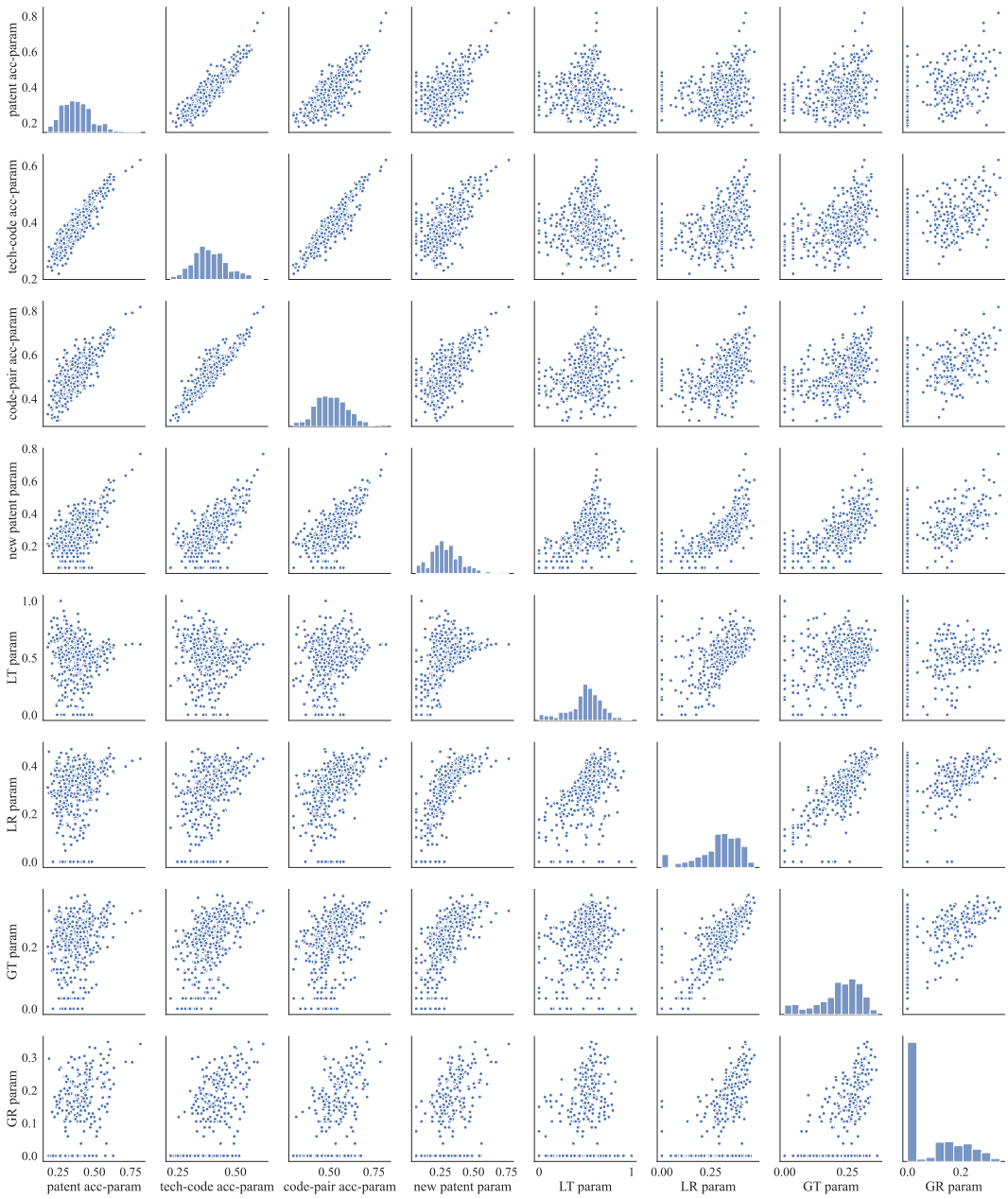


Figure Y.1: Pair distribution of CEE parameters of serial AI assignees in 2020.

APPENDIX Z

TIME SERIES OF FOUR COMPANIES' CEE PARAMETERS

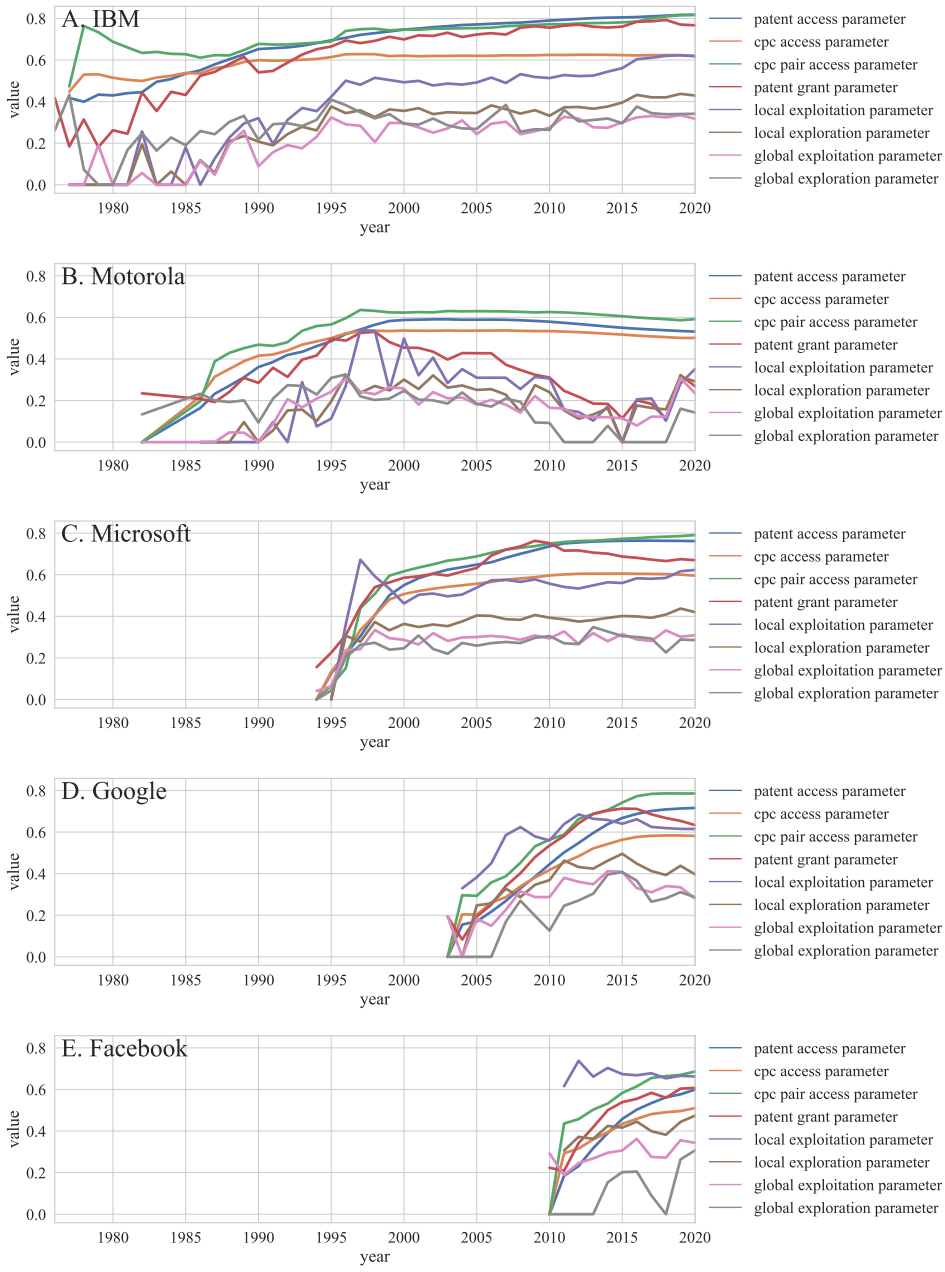


Figure Z.1: Time series of the CEE parameters of four selected AI organizations.

APPENDIX AA

GEOGRAPHIC VISUALIZATION SHOWING INVENTIVE ACTIVITIES OF AI IN EACH STATE OF THE U.S. IN 2020

The following Figure AA.1 shows geographic visualizations of the mean values of local exploration parameters of AI assignees in each state of the U.S. in 2020. The histograms at bottom right corner illustrates the distribution of the mean local exploration parameters.

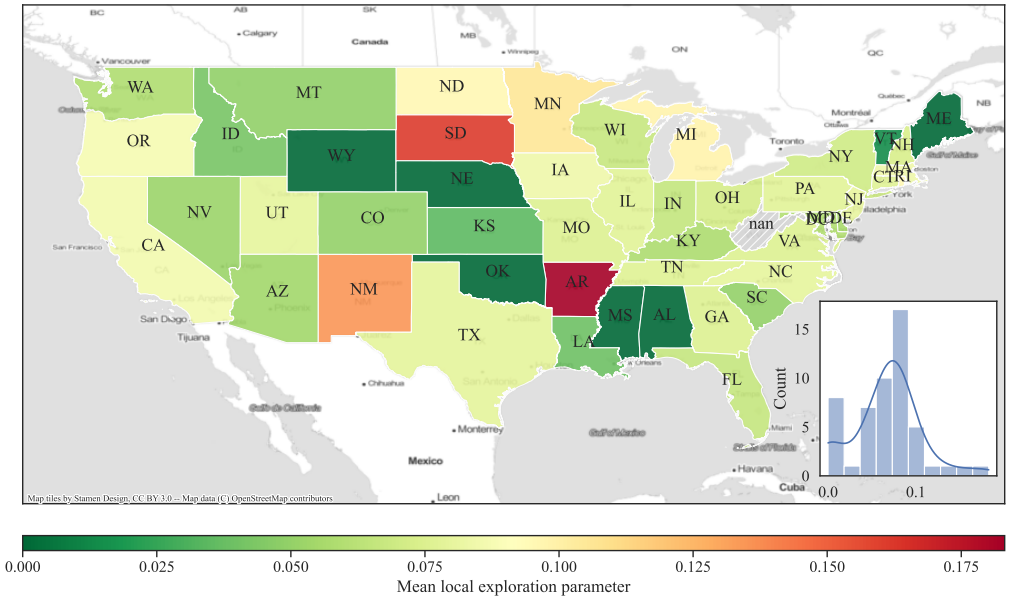


Figure AA.1: The map of mean local exploration parameters of U.S. AI companies

The following Figure AA.2 shows the number of AI assignees of each state in the U.S. in 2020.

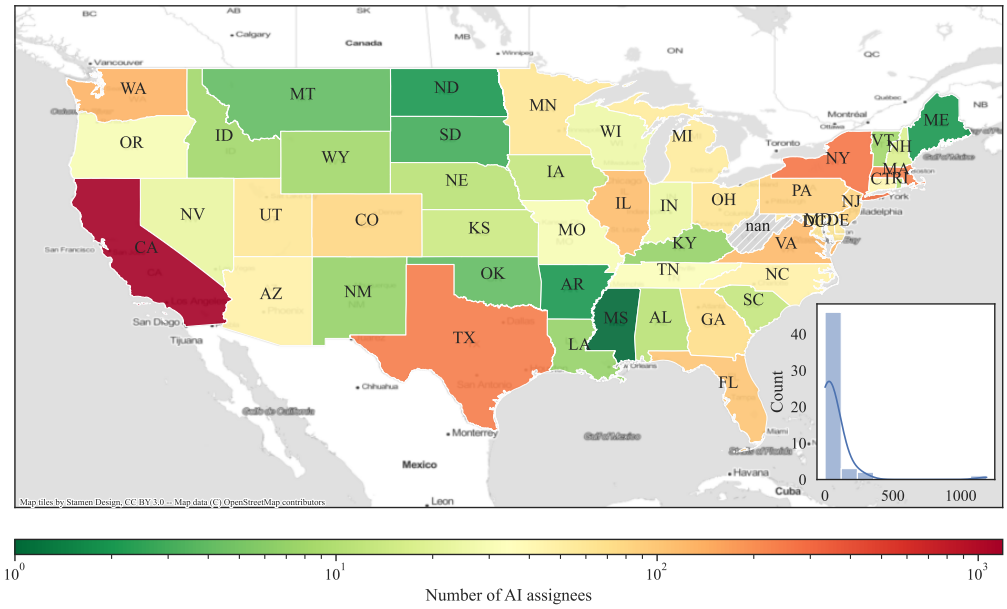


Figure AA.2: Number of AI organizations in each state in the U.S.

APPENDIX AB

GEOGRAPHIC VISUALIZATION OF THE NATIONAL MEAN GLOBAL EXPLOITATION PARAMETERS IN 2020.

The following Figure AB.1 visualizes the national mean values of global exploitation parameters of assignees that have obtained AI patents from the USPTO. The histogram at bottom left shows the distribution of such mean values of each countries.

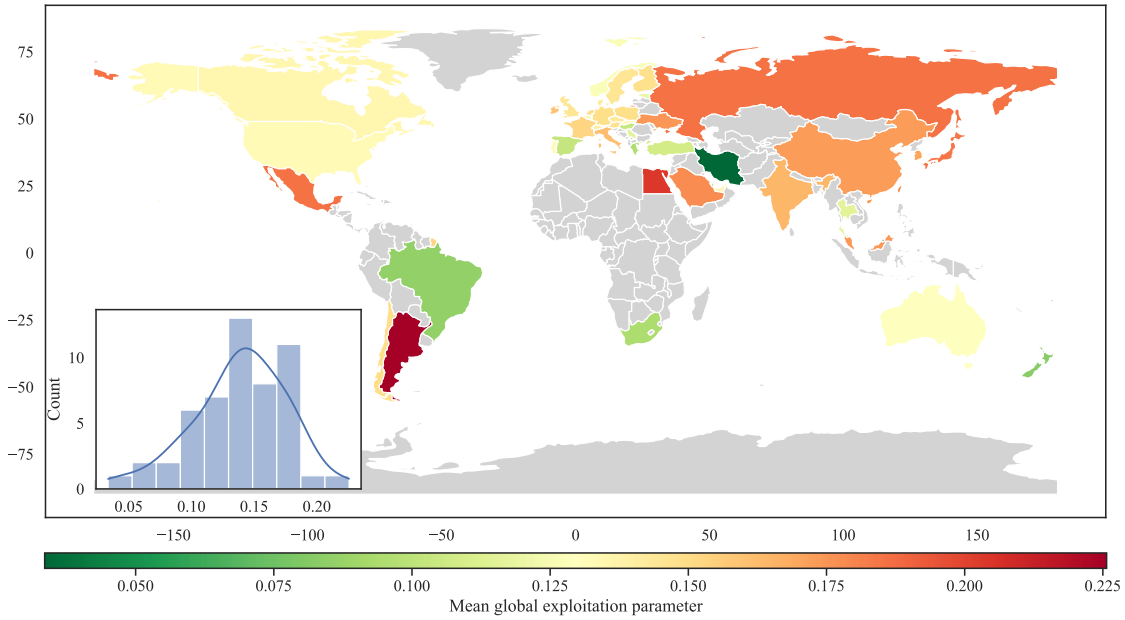


Figure AB.1: World map of global exploitation parameters

BIOGRAPHICAL SKETCH

Jieshu Wang is an interdisciplinary researcher with a keen interest in how innovation, invention, and technological changes impact society. Her primary focus is on emerging technologies, particularly in the field of artificial intelligence.

Jieshu's academic background includes civil engineering, economics, communications, and science and technology studies. Prior to her current research work, she served as a patent examiner in the Chinese patent office and then a science editor. These experiences sparked her curiosity about how science and technology advance, and the ways in which they can shape our world.

Through data-driven research, Jieshu aims to gain a deeper understanding of the mechanisms of science and technology, and to explore ways in which they can better serve society. She is committed to making meaningful contributions to the field of innovation, and to helping pave the way for a brighter technological future.