Le rôle de l'intelligence artificielle dans les questions sociétales

Diletta Abbonato

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE AUGUSTIN COURNOT ED 221

BUREAU D'ÉCONOMIE THÉORIQUE ET APPLIQUÉE UMR 7522

THÈSE

pour l'obtention du titre de Docteur en Sciences Économiques

Présentée et soutenue le 9 Septembre 2024 par

Diletta ABBONATO

The Role of Artificial Intelligence for Societal Challenges

Préparée sous la direction de Patrick LLERENA et Stefano BIANCHINI

Membres du jury :

Lili WANG	Associate Professor, Université de Maastricht	Président
Cinzia DARAIO	Professeur des Universités, Sapienza Université de Rome	Rapportrice
Björn JINDRA	Professeur des Universités, Copenhagen Business School	Rapporteur
André LORENTZ	Maître de conférences, Université de Strasbourg	Examinateur
Patrick LLERENA	Professeur des Universités, Université de Strasbourg	Directeur
Stefano BIANCHINI	Maître de conférences HDR, Université de Strasbourg	CoDirecteur

Alla mia famiglia e al loro immenso amore

L'Université de Strasbourg n'entend donner aucune approbation, ni improbation aux opinions émises dans cette thèse; elles doivent être considérées comme propres à leur auteur.

Direction de la **recherche**
et de la **valorisation**
Collège **doctoral**
Université de Strasbourg

Déclaration sur l'Honneur

Declaration of Honour

J'affirme être informé que le plagiat est une faute grave susceptible de mener à des sanctions administratives et disciplinaires pouvant aller jusqu'au renvoi de l'Université de Strasbourg et passible de poursuites devant les tribunaux de la République Française.

Je suis consciente que l'absence de citation claire et transparente d'une source empruntée à un tiers (texte, idée, raisonnement ou autre création) est constitutive de plagiat.

Au vu de ce qui précède, j'atteste sur l'honneur que le travail décrit dans mon manuscrit de thèse est un travail original et que je n'ai pas eu recours au plagiat ou à toute autre forme de fraude.

I affirm that I am aware that plagiarism is a serious misconduct that may lead to administrative and disciplinary sanctions up to dismissal from the University of Strasbourg and liable to prosecution in the courts of the French Republic.

I am aware that the absence of a clear and transparent citation of a source borrowed from a third party (text, idea, reasoning or other creation) is constitutive of plagiarism.

In view of the foregoing, I hereby certify that the work described in my thesis manuscript is original work and that I have not resorted to plagiarism or any other form of fraud.

Nom Prénom : Diletta Abbonato
Ecole doctorale : Augustin Cournot ED 221
Laboratoire : Bureau d'Économie Théorique et Appliquée UMR 7522

Acknowledgements

I would like to extend my deepest gratitude to a number of people without whom this thesis would not have been possible.

First of all, I owe my deepest thanks to my supervisors, Patrick Llerena and Stefano Bianchini. I am deeply grateful to you for having supported me morally and academically over these years. Many thanks also go to my co-authors Floriana Gargiulo and Tommaso Venturini for their availability and immense knowledge. I would also like to thank the Bureau d'Économie Théorique et Appliquée (BETA) and the Augustin Cournot Doctoral School for providing me with all the intellectual resources I needed to complete this work.

Moreover, I extend my heartfelt thanks to the members of my dissertation committee: André Lorentz, Björn Jindra, Cinzia Daraio, and Lili Wang. Your willingness to participate in my jury and provide critical insights and feedback was invaluable in refining my work and guiding my research path.

To my parents, who have given me the foundations upon which my studies have been built: your support and unwavering belief in my abilities have been the pillars of my strength. Your sacrifices have not gone unnoticed, and your love has been my constant source of encouragement. Thank you for instilling in me the curiosity and resilience that have been critical in pursuing my research. You have both taught me the value of hard work and perseverance, lessons that have carried me through the challenges of these academic years.

To my friends, especially the *late lunch* group, thank you for allowing me to enjoy my lunches at a time that did not feel like breakfast and, most importantly, thank you for all the carefree moments we shared together. Furthermore, I must acknowledge the unwavering emotional support from my group of *amici* since my return to Italy. Your encouragement and understanding have been essential in helping me through the complexities of my academic and personal life.

I am also immensely grateful to all the academic colleagues I have met over time whose insights and guidance have significantly shaped my research. Your expertise

has been invaluable to my growth as a scholar.

Finally, I extend my thanks to all those who supported me in any aspect during the completion of this project. It has been a journey of growth, learning and personal development and I am grateful to have had such a supportive community at my side.

Contents

General introduction

In the history of technological progress, some key revolutions have marked the course of human development. Scholars have identified three major technology-driven revolutions: the first industrial revolution, centered on the steam engine and mechanical production, the second marked by electricity and mass production, and the third, known as the digital revolution, marked by innovations such as semiconductors, personal computers and the Internet. These technological changes, defined as "*sets of interrelated radical breakthroughs forming a major constellation of interdependent technologies*" [Perez, 2010], were the driving forces behind the growth and transformation of society [Solow, 1957, Romer, 1990, Aghion and Howitt, 1990].

History can therefore be understood as a succession of *techno-economic paradigms*, each of which represents the optimal, most effective and profitable way of using new technology. It is '*techno*' because it starts with a technology or a small group of technologies; '*economic*' because the transformation involves a major shift in the relative price structure of existing products and services; and it is a '*paradigm*' in the sense defined by Kuhn, as it shapes and guides the standard organizational practices in technology, economics, management, and social institutions.[1]

At the core of these revolutions are General Purpose Technologies (GPTs), key technologies that have the potential to drive technical progress and economic growth, generating new opportunities and stimulating further innovations. As a result, GPTs influence a wide range of sectors, driving large-scale economic and social transformations [Bresnahan and Trajtenberg, 1995b, Perez, 2004, 2010]. GPTs are characterised by their pervasiveness across different sectors of the economy, their inherent potential for incremental and radical improvements, and their ability to stimulate complementary innovations and new ways of doing business. Such technologies lead to significant changes in the economy, often becoming an integral part of a wide

[1]Following Kuhn [1962], Dosi [1982], Perez [2004], a paradigm is defined hereafter as a specific pattern of solutions to selected techno-economic problems, based on principles derived from natural sciences, jointly with specific rules and heuristics to raise the relevant body of knowledge.

range of applications and processes. As engines of growth, GPTs play a central role in shaping the trajectory of technological progress and economic development [Helpman, 1998, Lipsey et al., 2005].

Since 2010s, the world has entered what is commonly referred to as the fourth industrial revolution (4IR henceforth) [Schwab, 2017, Philbeck and Davis, 2018].[2] The 4IR represents a shift towards an integrated and interconnected world, reducing boundaries between disciplines, industries, and geographical regions [Chen et al., 2017].

This age is expected to foster an economy characterized by enhanced cooperation and integration, building on earlier innovations such as the Internet of Things (IoT) and smart cities [Morrar et al., 2017], reshaping work, urban environments, and daily life [Ross and Maynard, 2021]. While these innovations offer transformative benefits, they also engender challenges. One example in point is the creation of new job concepts that align with technological advancements. Another is effectively managing potential disruptions in fields like communication, science, education, and behaviour [Xu et al., 2021].

More importantly, the 4IR age is characterized by the emergence and integration of Advanced Digital Technologies (ADTs), which represent a new phase in the series of major technological advances. ADTs are at the frontier of technology, using digital systems and tools to bring significant improvements in many areas. While they extend the trajectory of innovation established by previous technological advancements, ADTs also introduce their own distinct set of challenges and opportunities. Among this cluster of technologies, there are:

- *Artificial intelligence* is a class of machine-based systems that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments, and learn from experience. AI building blocks typically include four elements: machine learning, natural language processing, computer vision and speech recognition [Russell and Norvig, 2016]. As we will discuss later, AI is the major enabling technology of this revolution;

[2]The terms *4th Industrial Revolution* and *Industry 4.0* are often used interchangeably but originate from slightly different contexts. Industry 4.0 was first introduced in Germany as part of a government project to promote the computerization of manufacturing, while the term 4th Industrial Revolution was popularized by Klaus Schwab in a broader context.

- *Big data* can be thought of as those information assets characterized by high Volume, Velocity, and Variety (3Vs) that require specific technologies and analytical methods for their transformation into value. Other Vs are added from time to time, such as Veracity (data quality), Value (obtained from exploitation), and Variability (rate of change) [De Mauro et al., 2015];

- *Blockchain* is a secure protocol where a network of computers collectively verifies a transaction before it can be recorded and approved; it provides an immediate, shared, and transparent exchange of encrypted data simultaneously to multiple parties;

- *Computing infrastructures* (or ICT infrastructures) include physical and virtual resources that support the flow, storage, processing, and analysis of data. They provide the hardware and services that other systems and services are built upon; an infrastructure can be centralized within a data center or decentralized and distributed across multiple data centers;

- *Internet of Things* (IoT) is a concept describing an ecosystem of interconnected devices and services that collect, exchange and process data to adapt dynamically to a given context. IoT entails networks of physical objects – the "things" – embedded with ambient sensors and dedicated software, connected via communication protocols [Atzori et al., 2010];

- *(Advanced) robotics* encompasses agents with different capabilities to substitute for humans and replicate and automate human actions. Advances in sensors and machine learning enable robots to become more adaptive and sensitive, self-learning from the environment and improving with experience, thus engaging in a wider variety of tasks;

- *Virtual reality* (VR) involves the computer-generated simulation of a three-dimensional environment with which a person can interact in a seemingly real or physical way, using special electronic equipment fitted with sensors. *Augmented reality* (AR) is a technology that allows a computer-generated image to be superimposed on a user's view of the real world. Alternatively, the term *mixed reality* (MR) is used when elements of the real-world and the virtual environment are combined [Lanier, 2017];

- *5G* is the fifth generation of mobile networks. Compared with its predecessors, this network offers much higher connection speeds, lower response times (la-

tency) and greater capacity, making it possible to handle more high-demand applications simultaneously;

As the technologies that characterised earlier revolutions [Rosenberg, 1972, 1979], ADTs in the fourth industrial revolution are notable for their mutual dependence and complementarity, intelligent capabilities, and potential widespread impact across all sectors.

ADTs are considered as a novel form of GPT that is leading to significant and unprecedented developments in terms of size, speed, and scope. These technologies are not only influencing global issues such as climate change, migration, and geopolitical tensions, but they are also inspiring an investigation of human identity and experience within the framework of digitalization [Bianchini et al., 2023a].

The central role of AI. This dissertation focuses primarily on one of the technologies that compose the body of Industry 4.0: Artificial Intelligence (AI henceforth).

A definition that suits our purpose is the one provided by the OECD expert group on AI (AIGO) which developed a description of the AI system to define a clear-cut dimension for policy and regulation. According to AIGO (OECD, 2022), AI can be defined as "*a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with ML), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy*".

AI has emerged as a groundbreaking technology with the potential to address different societal challenges, integrating with the other key components. Moreover, with this dissertation we aim at exploring the impact of AI on societal issues and investigating its interactions with other technologies such as big data, robotics and the IoT, shedding light on the role of AI not only as a widely-applicable tool, but also as an evolving technology that shapes and upgrades itself along with technical progress and transformations in the wider social dimension.

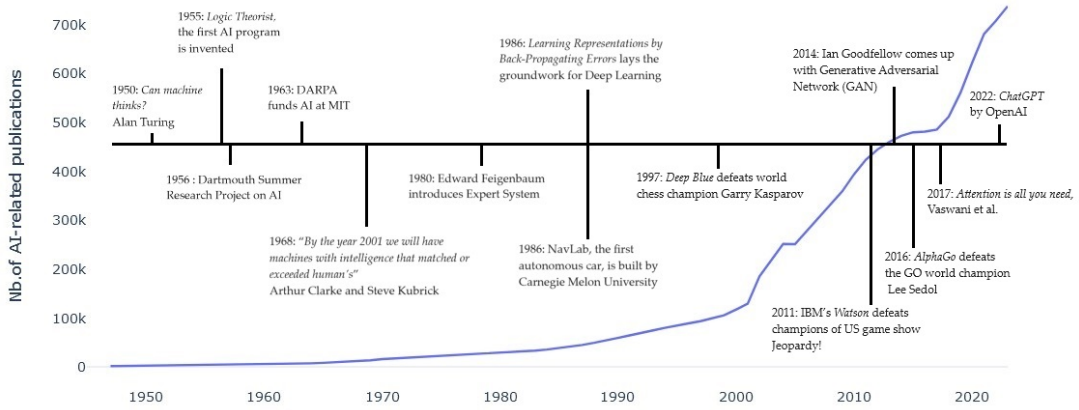
The development of AI (Figure 1) as a multidisciplinary field traces back to the XX century with the advent of electronic computers. Turing [1950] introduced it in the seminal paper on "Computing machinery and intelligence" in which he first discussed the potential of machines to think and then paved the way to future ex-

plorations. For instance, the Turing Test is a well-known measure of a machine ability to exhibit human-like intelligence. Significant contributions also came from McCarthy [1959] with the development of the Lisp programming language, known for its easy manipulation of data strings, later adopted by companies such as Google to develop further software applications. McCarthy's belief that "*every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*"[McCarthy et al., 2006], highlights a key idea in AI: machines might one day be able to fully mimic how humans think and learn. This theoretical possibility began to materialize with the work of researchers Newell and Simon [1956]. Their creation, the *Logic Theorist*, is considered to be the first artificial intelligence program, capable of mimicking the problem-solving abilities of a human being. This demonstrated that machines could not only calculate but also engage in more complex, thought-like processes, such as reasoning. Between the 1960s and 1970s, advances in computing, such as improved storage and data processing capabilities, promoted further AI research. Basic machine learning algorithms were developed during this period. Agencies such as DARPA invested in AI, initially focusing on automatic language translation and transcription. The 1980s witnessed a surge in AI funding and algorithm development, highlighted by the work of John Hopfield [1982] and David Rumelhart et al. [1985] on deep learning techniques. From the 1990s to the early 2000s, AI researchers achieved key milestones, including the defeat of the world chess champion Garry Kasparov with *Deep Blue*, a chess-playing expert system run on a purpose-built IBM supercomputer [Campbell et al., 2002]. Today, with increased computing power and data availability, AI tools such as ChatGPT, DALL-E and AlphaFold illustrate the rapid evolution of the field towards a general AI, enabling software to autonomously perform complex tasks once thought to be the exclusive domain of humans.[3]

These rapid advancements have spurred extensive academic research into its complex interactions with economic systems. A key area of debate focuses on whether AI should be classified as a General Purpose Technology (GPT) or an Invention of a

[3]General AI, also known as strong AI [Kurzweil, 2005] or human-level AI [Roser, 2023] is an artificial intelligence software that possesses human-like intelligence and has the capability to learn on its own. The objective is to enable the software to execute tasks that it may not have been specifically trained or developed for. Existing AI technologies operate within predefined parameters. General AI is an intellectual endeavour aimed at creating AI systems that have independent self-regulation, a level of self-awareness, and the capacity to acquire new abilities. The achievement of General AI with human-level abilities is still a theoretical concept and a primary objective of research [Amazon, 2023].

Figure 1: AI timeline and number of publications over time



Notes: Nb. of AI-related publications, source OpenAlex – own elaboration

Method of Inventions (IMI). This discussion involves scholars such as Cockburn et al. [2018], Klinger et al. [2018], Agrawal et al. [2019c], Klinger et al. [2020], Bianchini et al. [2022], Vannuccini and Prytkova [2023]. A growing body of literature suggests that AI can reshape the way we produce knowledge, both within and between many scientific fields increasing impacts on several avenues, from scientific discovery to productivity of scientists [Cockburn et al., 2018, Galindo-Rueda, 2020, OECD, Bianchini et al., 2022, Borsato and Lorentz, 2023].[4] As a GPT, the pervasive nature of AI in multiple sectors radically changes the technological landscape, similar to GPTs of the past such as electricity or the Internet. On the other hand, as IMI, the importance of AI lies in the enhancement and automation of the innovation process itself, making it a key tool for the acceleration of scientific discovery and the expansion of combinable knowledge. The ongoing debate about the definition of AI reflects its evolving nature and expanding applications, while some scholars advocate a narrow, technical definition focusing on ML and data processing, others call for a broader perspective that includes AI in societal, economic, and ethical dimensions [Klinger et al., 2020, Chubb et al., 2021]. A large part of this dissertation (chapter 1 and 2) will contribute to this literature on the diffusion and impact of AI/ML in the scientific system.

[4]Yet, different opinions persist on the widespread diffusion of AI-based technologies. We refer the interested reader to Vannuccini and Prytkova [2023] and the empirical evidence in McElheran et al. [2023].

The socio-economic impact of AI. AI adoption started in the physical sciences and then spread into the life sciences, social sciences, arts and humanities [Gefen et al., 2021]. Today, domains of application span a wide range of industries, ranging from healthcare to transportation with key impacts on labour-market dynamics and other societal issues - e.g., climate change, income inequality, sustainable development.

For example, in healthcare, AI-based systems are improving early diagnosis, personalizing medicine and improving patient outcomes [Jiang et al., 2017, Bohr and Memarzadeh, 2020]. ML algorithms can analyze vast medical data to identify patterns and predict diseases with increased accuracy [Uddin et al., 2019, Ngiam and Khor, 2019] while AI-enabled robots in surgeries enhance precision and reduce errors [Park et al., 2022]. In addition to this, AI algorithms are becoming crucial for analysing medical images such as X-rays, CT scans and MRI scans, helping diagnose and detect diseases such as skin cancer and lung disease, and assessing cardiovascular risk [Esteva et al., 2017]. More recently, a study published in *Nature* by researchers from MIT showed the use of AI to discover a class of compounds that can kill a drug-resistant bacterium that causes more than 10,000 deaths in the United States every year [Wong et al., 2023]. In addressing climate change, AI plays a significant role in developing sustainable solutions. AI algorithms allow for the efficient management of renewable energy sources, reducing dependence on fossil fuels [Reichstein et al., 2019, Song and Roh, 2021, Bianchini et al., 2023b]. It also forecasts climate patterns and informs disaster planning, helping decrease greenhouse gas emissions and save energy [Ramli et al., 2008]. For instance, Google's DeepMind has significantly enhanced the energy efficiency of wind farms by accurately predicting wind patterns [GoogleDeepMind, 2023].

Moreover, a number of strands of the literature emphasises the role of AI in affecting industrial dynamics through the increasing implementation of cutting-edge techniques for the analysis of large datasets and automation of tasks. As also highlighted by Brynjolfsson and McAfee [2014], and Borsato and Lorentz [2023], AI applications such as automated production processes and optimization techniques enhance efficiency and competitiveness, driving economic growth through innovative products and services [Agrawal et al., 2019a]. Acemoglu et al. [2022] analysed how over 300,000 U.S. firms adopted advanced technologies between 2016 and 2018. This research focus on five key areas - i.e, AI, robotics, specialised software for specific business functions, dedicated equipment for automated tasks, and cloud-based

Table 1: Main AI narratives

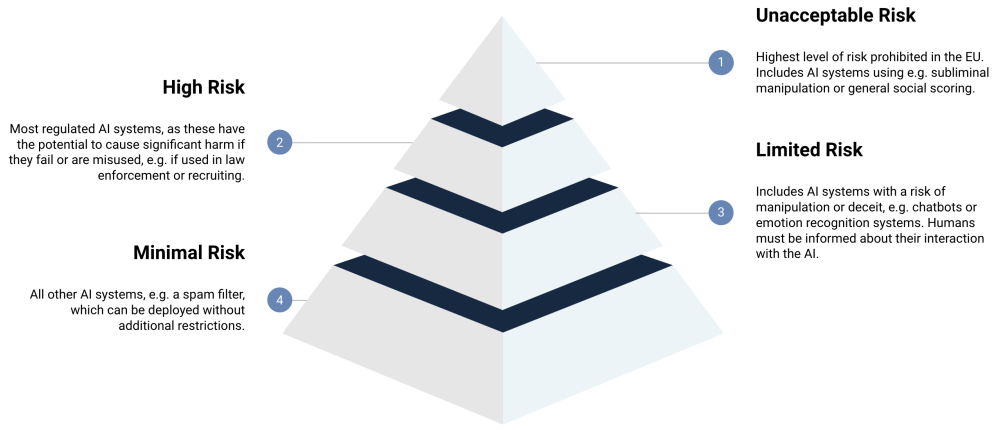
Field	Hopes	Fear	Debate
Health	Immortality	Dehumanization	Man conquering immortality while on the other side humans lose their essence, ditching values and emotions.
Employment	Freedom	Job replacement/Obsolescence	Humans will be liberated from tedious or tiring tasks, be they physical or cognitive. The opposite representation is the risk linked to this technical turning point.
Sociology	Gratification	Alienation	AI and robots fulfill every human desire, but on the other hand, the opposite scenario predicts that individuals will only interact with technologies rather than with other people.
Surveillance	Security	Uprising	The optimistic scenario predicts that new tools will enable nations and communities to ensure security for all, while on the other hand there is the iconic narrative of sci-fi where AI will take over humans.

Notes: Own elaboration based on Cave and Dihal [2019]

computing systems and applications shows that the integration of these technologies correspond to a shift in labour demand towards workers capable of managing and operating advanced systems. This shift, according to the authors, reveal a trend towards a more technologically advanced, interconnected, and data-centric business landscape. Looking further, AI-driven systems, e.g, recommendation engines, have transformed sectors such as e-commerce [Bughin et al., 2018]. There is also evidence highlighting the positve effects of AI technologies on increasing the innovative performance and profitability of firms [Khin and Ho, 2018, Leusin, 2022, Rammer et al., 2021]. AI has also been a catalyst in the digital platform economy, with companies like Uber and Airbnb creating new business opportunities [Brynjolfsson and McAfee, 2017].

While AI affects industry dynamics in terms of growth and sustainability, it presents dual-edged (ambiguous) effects on the job market [Brynjolfsson et al., 2018, Acemoglu and Restrepo, 2019a,b, Aghion et al., 2019, Domini et al., 2021, Bordot, 2022, Mondolo, 2022]. The potential for task automation may lead to job displacement in some sectors – especially those involving routine or repetitive tasks [Brynjolfsson and McAfee, 2014], yet it also creates new job opportunities and enhances productivity, contributing to economic growth [Brynjolfsson and McAfee, 2017]. The impact varies according to AI-adoption rates, the nature of automated tasks, and workforce adaptation [Autor et al., 2020].

Figure 2: The four risk classes of the EU AI Act



Notes: Source Trail-ml

Furthermore, beyond its tangible impact on different sectors, the role of AI extends in tackling social challenges, introducing significant ethical considerations and dilemmas. A primary concern is its potential to perpetuate bias and discrimination. AI systems, trained on large datasets, may amplify inherent biases if data contains prejudiced or discriminatory information [O'Neil, 2016].

This risk is exacerbated since these systems often rely on extensive personal data to work efficiently. The protection and appropriate use of data are imperative to keep confidence in AI technologies. Robust regulations and mechanisms are essential to safeguard individual privacy rights while allowing for AI-beneficial applications. The *AI Act*, adopted by the European Union in 2023, is a first step forward in this direction. As a comprehensive framework, it regulates AI development, marketing, and use across several sectors, excluding the military. The AI Act is characterized by its risk-based requirements for AI systems, prohibitions on certain harmful AI practices, and extraterritorial reach, potentially influencing AI governance globally, as much as the EU's General Data Protection Regulation (GDPR). The regulatory framework defines four levels of risk for AI systems represented in a pyramid model (Figure 2) that ranks AI systems according to their potential risks and the level of regulatory oversight required. At the highest level of the pyramid, AI systems are classified as "Unacceptable Risk" if they pose an extreme threat to security and fundamental rights. This category includes systems that employ subliminal manipulation or en-

able social scoring, which could have a severe impact on individuals' opportunities and access to resources. Such systems are prohibited within the European Union. Subsequently, there are high-risk AI systems, which include applications in critical areas such as healthcare, law enforcement and judicial decision-making. These systems are subject to strict compliance requirements because of their potential to cause significant damage if they fail or are misused. Systems at the next level of the pyramid are those that present a limited risk. These include technologies such as chatbots or emotion recognition systems that have the potential to manipulate or deceive. Although not inherently harmful, these systems require clear transparency to ensure that users are fully aware that they are interacting with AI, thereby safeguarding informed decision-making. At the base of the pyramid, AI applications with minimal risk, such as spam filters or automated video recommendation systems, pose negligible risks to society and therefore enjoy the most lenient regulatory standards, which promote innovation while maintaining essential security measures.

In the realm of data privacy and security, the challenges are amplified by the advanced collection and analysis of personal data by AI technologies. Protecting sensitive information amidst the growing scale and complexity of data processing requires robust security measures, privacy-enhancing technologies, and comprehensive legislation. Encryption, access controls, anonymization methods, secure data storage systems, regular security audits, and data protection training are key components in mitigating risks associated with personal data [Murdoch, 2021, Garrido et al., 2022, Jordan et al., 2022]. Alongside the imperative of securing data, there's an equally crucial challenge of ensuring fairness in AI algorithms. Ongoing research in algorithmic focuses on developing mathematical frameworks and methodologies to deal with bias in AI [Dwork et al., 2012]. Achieving unbiased AI algorithms is in fact not just an ethical necessity but it has economic implications. Biased algorithms can perpetuate inequalities, limit opportunities for marginalized groups, and erode trust in AI systems, hindering their adoption [Mittelstadt et al., 2016]. Prioritizing fairness in algorithm design and training processes is therefore essential to engender a more inclusive and equitable society.

The application of ADTs, especially AI, for the social good has been a topic of considerable debate. Critical analyses as in Taylor [2016] explore the complexity of treating big data as a public good, highlighting the conflicts between rights, duties, and claims (see also Savona [2019]). Likewise, Moore [2019] assesses the ambiguous narrative of "AI for social good", emphasizing the need for greater clarity and

reflection, particularly in the public debate and data science. In a special issue introduced by Cowls [2021], the complexities of defining what a "social good" is in the context of AI-based technologies are investigated, probing the ethical stances and power dynamics involved. Furthermore, Floridi et al. [2020] look at the potential of AI in addressing social issues and contributing to societal well-being, including environmentally sustainable developments. These works collectively highlight the importance of a nuanced understanding of the societal and ethical implications of AI and data technologies, challenging oversimplified narratives and advocating for a reflective approach in assessing their diverse impacts. The above opens the discussion about the role of AI in terms of economic development at large, and it is pertinent to consider the United Nations' Agenda for Sustainable Development, an holistic framework [UN-General-Assembly, 2015, TWI2050, 2019]. This agenda establishes a comprehensive plan to fuel progress across multiple domains anchored to the 17 Sustainable Development Goals (SDGs). Thus, the SDGs framework encourages a thorough evaluation of the impact of ADTs, including AI, on several interconnected phenomena [Chui et al., 2018, Goralski and Tan, 2020, Vinuesa et al., 2020, Cowls, 2021, Guenat et al., 2022, Bianchini et al., 2023a]. The literature on this strand has repeatedly stressed the critical role of AI-driven technologies in achieving the SDGs [Goh, 2021, Jindra and Leusin, 2022, Series, 2018]

In an era in which technological advancements are pivotal to societal evolution, this thesis aims to comprehensively delineate the multifaceted role of AI within the scientific paradigm and its broader societal implications. The research focuses on three main dimensions: (i) AI contribution to scientific research, (ii) the influence of the private sector in AI development and application, and (iii) the interplay between public perception and AI societal role.

In light of these considerations, the fulcrum of this dissertation turns around a threefold inquiry. The first aspect involves delineating how AI is affecting the scientific ecosystem (e.g., collaborations), the examination of the related integration between different scientific disciplines, and the transformative impacts thereof. The second part focuses on how collaboration between public and private sectors, especially in the field of emerging *Transformer* technology, enables scientists who use this technology to increase the impact of their research activity. The third aspect analyses the impact of social perceptions on AI integration into social fabric, considering in which way public attitudes, awareness, and concerns may influence and determine the development and deployment of AI technologies.

In **chapter one**, we explore the emerging trend of interdisciplinarity, a concept that has gained significant interest in science policy, particularly highlighted during the COVID-19 pandemic. The outbreak encouraged epidemiologists and medical researchers to not only use resources within their disciplines but also to seek new ideas and external collaborations. Among these, the integration with AI emerged as a particularly promising alliance. However, the collaborative ventures combining two of the most prominent topics in scientific and societal discussions – AI and COVID-19 – have shown varying degrees of productivity.

It becomes evident that the multidisciplinary nature of AI–COVID-19 research necessitates the formation of diverse, complementary teams comprising researchers from different fields. Our study aims to investigate the factors that contribute to successful collaborations between domain experts and AI specialists. While previous research indicates that the most successful collaborations occur through interdisciplinary efforts within closely related fields, there is limited investigation into the tangible outcomes of these collaborations.

To address this gap, our analysis focuses on the adoption of AI techniques in COVID-19 research. We used data from three distinct databases: CORD-19, Semantic Scholar, and Altmetric. This analysis is supported by a set of newly designed metrics that evaluate interdisciplinarity in relation to AI on two levels: team diversity (including the participation of AI experts in COVID-19 research) and epistemological diversity (measuring the actual knowledge utilized in each research article).

Our findings are threefold. First, we observe that both forms of diversity – team and epistemological – are positively associated with various forms of impact, such as citation counts, media attention, and interdisciplinary outreach. Second, a notable trend is the negative association between the involvement of AI experts and the impact of research, suggesting challenges in collaboration between domain and AI experts in producing influential science. Lastly, our analysis indicates that epistemological diversity holds more significance for impact beyond the academic sphere. By mapping the diffusion of AI in scientific research and its impact, this chapter aims to contribute to a better understanding of how computational technologies are valuable in addressing both current and future societal challenges.

Chapter two contributes to the understanding of the relationships between academia and industry in the development of AI / ML models and the impact these models have on scientific discovery, with a particular focus on Transformer technology – a groundbreaking subset of deep learning.

The inception of Transformers, introduced in the seminal paper "Attention is All You Need" [Vaswani et al., 2017] at the 2017 Neural Information Processing Systems conference, reshaped the trajectory of AI across various fields. These architectures are now ubiquitous in a myriad of scientific applications, spanning natural language processing, computer vision, reinforcement learning, biology, and beyond [Lin et al., 2022, Han et al., 2023, Wang et al., 2023]. While AI is revolutionizing scientific discovery and productivity [Cockburn et al., 2018, Bianchini et al., 2022], the specific technology driving these advancements remains unclear. Moreover, the historical role of the private sector in these developments is not well-understood.

Noteworthy collaborations, such as the Dartmouth conference in 1956 – achieved through the collaborative efforts of IBM, Bell Labs, and MIT – and contemporary projects such as AlphaGO, resulting from collaboration between Google and the Universities of Stanford and Oxford, highlight the crucial role of public-private partnerships in the evolution of AI.

In this exploration, we examine the dynamic interplay of these collaborations in fostering technological innovation and scientific growth. The analysis unfolds through a series of methodological approaches including the utilization of a robust dataset comprising 113 transformative publications categorized into textual, vision, and other applications, along with extensive citation data highlighting the pervasive influence of Transformers in scientific research.

Our study leverages Difference-in-Differences (DiD) models to quantify the impacts of Transformer adoption on scientific output, evidencing that papers developed through academia-industry collaborations not only receive more citations but also exhibit higher novelty compared to their counterparts. This empirical evidence underscores the significant role of private sector involvement in enhancing the impact and disruptiveness of scientific research using Transformer technologies.

Furthermore, we discuss the broader implications of these findings within the context of ongoing debates about the privatization of AI research and the monopolization tendencies of big tech firms. Our findings strongly suggest that such partnerships are a *condicio sine qua non* for achieving the full potential of AI-driven scientific advancements. Strategic collaborations between academia and industry are

not merely beneficial but essential, serving as the cornerstone for advancing state-of-the-art technology and ensuring that these advancements catalyze broad-based scientific enrichment.

The discourse extends into a critical evaluation of the role of such partnerships in mitigating the risks associated with concentrated technological power, advocating for policies that promote collaborative innovation while safeguarding equitable access to technology.

In **chapter three** we investigate the societal dimension of AI, recognizing the public as active participants in its development and emphasizes the dynamic interaction between technology and society in the age of 4IR. The rapidly evolving digital landscape, catalyzed by the 4IR, underscores the integration of cutting-edge technologies like AI, robotics, and blockchain into societal frameworks. This technological integration not only reshapes economic structures but deeply influences cultural norms and social interactions. As societies grapple with these changes, understanding the public perception and societal impact of these technologies becomes crucial for fostering public trust and aligning technological advancements with societal needs and values. The discourse surrounding 4IR is rich with contrasting narratives, ranging from utopian visions of enhanced capabilities and efficiencies to dystopian fears concerning privacy erosion, job displacement, and social disengagement. Literature shows a dichotomy in public sentiment that oscillates between optimism for technological empowerment and anxiety about loss of control and identity. This polarization is evident in various domains, from healthcare benefits to concerns over surveillance and algorithmic biases.

Using a multi-country dataset including tweets and media articles, in this chapter we employed sentiment analysis and machine learning models to explore the public discourse related to 4IR technologies. The analysis spans six European countries, capturing diverse public opinions shaped by cultural and economic contexts. The methodology integrates sentiment analysis to gauge public emotions and machine learning classifiers to identify prevalent themes in discussions related to AI, robotics, and other 4IR technologies.

The findings reveal a polarization in public opinion, with a significant decline in neutral perspectives and a rise in distinctly positive or negative sentiments. This trend suggests a societal shift towards more definitive stances on 4IR technologies, possibly driven by increased awareness and engagement. The sentiment analysis

indicates a general optimism about the potential benefits of these technologies, particularly in enhancing quality of life and economic opportunities. However, concerns about privacy and the ethical use of technology persist, reflecting widespread apprehensions about data misuse and the implications of autonomous systems. By understanding the diverse public sentiments and ethical considerations, policymakers can craft more inclusive and forward-thinking technology policies. Moreover, the results underscore the need for robust digital education and public awareness programs to bridge the knowledge gap and mitigate the risks associated with misinformation and polarized opinions.

Introduction générale

Dans l'histoire du progrès technologique, certaines révolutions clés ont marqué le cours du développement humain. Les chercheurs ont identifié trois grandes révolutions entraînées par la technologie : la première révolution industrielle, centrée sur la machine à vapeur et la production mécanique, la seconde marquée par l'électricité et la production de masse, et la troisième, connue sous le nom de révolution numérique, marquée par des innovations telles que les semi-conducteurs, les ordinateurs personnels et l'Internet. Ces changements technologiques, définis comme "*sets of interrelated radical breakthroughs forming a major constellation of interdependent technologies*" [Perez, 2010], ont été les forces motrices derrière la croissance et la transformation de la société [Solow, 1957, Romer, 1990, Aghion and Howitt, 1990].

L'histoire peut donc être comprise comme une succession de *techno-economic paradigms*, chacun représentant la manière optimale, la plus efficace et rentable d'utiliser les nouvelles technologies. C'est '*techno*' parce que cela commence par une technologie ou un petit groupe de technologies ; '*économique*' parce que la transformation implique un changement majeur dans la structure des prix relatifs des produits et services existants ; et c'est un '*paradigm*' dans le sens défini par Kuhn, car il façonne et guide les pratiques organisationnelles standard dans les domaines de la technologie, de l'économie, de la gestion et des institutions sociales.[5]

Au cœur de ces révolutions se trouvent les *General Purpose Technologies* (GPTs), des technologies clés qui ont le potentiel de stimuler le progrès technique et la croissance économique, générant de nouvelles opportunités et stimulant d'autres innovations. En conséquence, les GPTs influencent un large éventail de secteurs, entraînant des transformations économiques et sociales à grande échelle [Bresnahan and Trajtenberg, 1995b, Perez, 2004, 2010]. Les GPTs se caractérisent par leur om-

[5]Suivant Kuhn [1962], Dosi [1982], Perez [2004], un paradigme est défini désormais comme un modèle spécifique de solutions aux problèmes techno-économiques sélectionnés, basé sur des principes dérivés des sciences naturelles, conjointement avec des règles spécifiques et des heuristiques pour élever le corps pertinent de connaissances.

niprésence à travers différents secteurs de l'économie, leur potentiel inhérent pour des améliorations incrémentielles et radicales, et leur capacité à stimuler des innovations complémentaires et de nouvelles manières de faire des affaires. Ces technologies entraînent des changements significatifs dans l'économie, devenant souvent une partie intégrante d'une large gamme d'applications et de processus. En tant que moteurs de croissance, les GPTs jouent un rôle central dans le façonnement de la trajectoire du progrès technologique et du développement économique [Helpman, 1998, Lipsey et al., 2005].

Depuis les années 2010, le monde est entré dans ce qui est communément appelé la quatrième révolution industrielle (4IR désormais) [Schwab, 2017, Philbeck and Davis, 2018].[6] La 4IR représente un déplacement vers un monde intégré et interconnecté, réduisant les frontières entre les disciplines, les industries et les régions géographiques [Chen et al., 2017].

Cette ère devrait favoriser une économie caractérisée par une coopération et une intégration accrues, s'appuyant sur des innovations antérieures telles que l'Internet des Objets (IoT) et les villes intelligentes [Morrar et al., 2017], remodelant le travail, les environnements urbains et la vie quotidienne [Ross and Maynard, 2021]. Bien que ces innovations offrent des avantages transformateurs, elles engendrent également des défis. Un exemple est la création de nouveaux concepts d'emploi qui s'alignent sur les avancées technologiques. Un autre est la gestion efficace des perturbations potentielles dans des domaines tels que la communication, la science, l'éducation et le comportement [Xu et al., 2021].

Plus important encore, l'ère de la 4IR est caractérisée par l'émergence et l'intégration des Technologies Numériques Avancées (ADTs), qui représentent une nouvelle phase dans la série des grandes avancées technologiques. Les ADTs sont à la frontière de la technologie, utilisant des systèmes et des outils numériques pour apporter des améliorations significatives dans de nombreux domaines. Bien qu'elles prolongent la trajectoire de l'innovation établie par les avancées technologiques précédentes, les ADTs introduisent également leur propre ensemble distinct de défis et d'opportunités. Parmi ce groupe de technologies, il y a :

[6]Les termes *4th Industrial Revolution* et *Industry 4.0* sont souvent utilisés de manière interchangeable mais proviennent de contextes légèrement différents. Industry 4.0 a été introduit pour la première fois en Allemagne dans le cadre d'un projet gouvernemental visant à promouvoir l'informatisation de la fabrication, tandis que le terme 4th Industrial Revolution a été popularisé par Klaus Schwab dans un contexte plus large.

- *Intelligence artificielle* est une classe de systèmes basés sur la machine qui peuvent, pour un ensemble donné d'objectifs définis par l'homme, faire des prédictions, des recommandations ou des décisions influençant des environnements réels ou virtuels, et apprendre de l'expérience. Les éléments constitutifs de l'IA incluent généralement quatre éléments : l'apprentissage automatique, le traitement du langage naturel, la vision par ordinateur et la reconnaissance vocale [Russell and Norvig, 2016]. Comme nous le discuterons plus tard, l'IA est la technologie habilitante majeure de cette révolution;

- *Big data* peut être considéré comme ces actifs d'information caractérisés par un Volume élevé, une Vélocité et une Variété (3Vs) qui nécessitent des technologies spécifiques et des méthodes analytiques pour leur transformation en valeur. D'autres Vs sont ajoutés de temps en temps, tels que la Véracité (qualité des données), la Valeur (obtenue de l'exploitation), et la Variabilité (taux de changement) [De Mauro et al., 2015];

- *Blockchain* est un protocole sécurisé où un réseau d'ordinateurs vérifie collectivement une transaction avant qu'elle puisse être enregistrée et approuvée; il offre un échange immédiat, partagé et transparent de données cryptées simultanément à plusieurs parties;

- *Infrastructures de calcul* (ou infrastructures TIC) comprennent des ressources physiques et virtuelles qui soutiennent le flux, le stockage, le traitement et l'analyse des données. Elles fournissent le matériel et les services sur lesquels d'autres systèmes et services sont construits; une infrastructure peut être centralisée dans un centre de données ou décentralisée et distribuée à travers plusieurs centres de données;

- *Internet des Objets* (IoT) est un concept décrivant un écosystème d'appareils et de services interconnectés qui collectent, échangent et traitent des données pour s'adapter dynamiquement à un contexte donné. IoT implique des réseaux d'objets physiques – les "choses" – équipés de capteurs ambiants et de logiciels dédiés, connectés via des protocoles de communication [Atzori et al., 2010];

- *Robotique avancée* englobe des agents avec différentes capacités pour substituer les humains et répliquer et automatiser les actions humaines. Les progrès dans les capteurs et l'apprentissage automatique permettent aux robots de devenir

plus adaptatifs et sensibles, auto-apprenants de l'environnement et s'améliorant avec l'expérience, s'engageant ainsi dans une variété plus large de tâches;

- *Réalité virtuelle* (VR) implique la simulation générée par ordinateur d'un environnement tridimensionnel avec lequel une personne peut interagir de manière apparemment réelle ou physique, en utilisant un équipement électronique spécial équipé de capteurs. *Réalité augmentée* (AR) est une technologie qui permet de superposer une image générée par ordinateur sur la vue réelle de l'utilisateur. Alternativement, le terme *réalité mixte* (MR) est utilisé lorsque des éléments du monde réel et de l'environnement virtuel sont combinés [Lanier, 2017];

- *5G* est la cinquième génération de réseaux mobiles. Comparée à ses prédécesseurs, ce réseau offre des vitesses de connexion beaucoup plus élevées, des temps de réponse (latence) plus courts et une plus grande capacité, permettant de gérer simultanément plus d'applications à forte demande;

Comme les technologies qui ont caractérisé les révolutions précédentes [Rosenberg, 1972, 1979], les ADTs dans la quatrième révolution industrielle se distinguent par leur interdépendance et complémentarité mutuelles, leurs capacités intelligentes et leur impact potentiel généralisé dans tous les secteurs.

Les ADTs sont considérées comme une nouvelle forme de GPT qui conduit à des développements significatifs et sans précédent en termes de taille, de vitesse et de portée. Ces technologies influencent non seulement des problèmes mondiaux tels que le changement climatique, la migration et les tensions géopolitiques, mais elles inspirent également une exploration de l'identité et de l'expérience humaines dans le cadre de la digitalisation [Bianchini et al., 2023a].

Le rôle central de l'IA. Cette dissertation se concentre principalement sur l'une des technologies qui composent le corps de l'Industrie 4.0 : l'Intelligence Artificielle (IA désormais).

Une définition qui répond à nos objectifs est celle fournie par le groupe d'experts de l'OCDE sur l'IA (AIGO), qui a développé une description du système d'IA pour définir une dimension claire pour la politique et la réglementation. Selon l'AIGO (OCDE, 2022), l'IA peut être définie comme "*un système basé sur la machine capable d'influencer l'environnement en produisant une sortie (prédictions, recommandations ou décisions) pour un ensemble donné d'objectifs. Il utilise des données et des entrées basées sur la machine et/ou l'humain pour (i) percevoir des environnements réels*

et/ou virtuels; (ii) abstraire ces perceptions en modèles par une analyse de manière automatisée (par exemple, avec ML), ou manuellement; et (iii) utiliser l'inférence de modèle pour formuler des options pour les résultats. Les systèmes d'IA sont conçus pour fonctionner avec des niveaux d'autonomie variables".

L'IA a émergé comme une technologie révolutionnaire avec le potentiel de répondre à différents défis sociétaux, s'intégrant avec les autres composants clés. De plus, avec cette dissertation, nous visons à explorer l'impact de l'IA sur les problèmes sociétaux et à étudier ses interactions avec d'autres technologies telles que les big data, la robotique et l'IoT, mettant en lumière le rôle de l'IA non seulement en tant qu'outil largement applicable, mais aussi en tant que technologie évolutive qui se façonne et se perfectionne au gré du progrès technique et des transformations dans la dimension sociale plus large.

Le développement de l'IA (Figure 1) en tant que domaine multidisciplinaire remonte au XXe siècle avec l'avènement des ordinateurs électroniques. Turing [1950] l'a introduit dans l'article fondateur sur "Computing machinery and intelligence" dans lequel il a d'abord discuté du potentiel des machines à penser puis a ouvert la voie à de futures explorations. Par exemple, le test de Turing est une mesure bien connue de la capacité d'une machine à exhiber une intelligence semblable à celle de l'homme. Des contributions significatives sont également venues de McCarthy [1959] avec le développement du langage de programmation Lisp, connu pour sa manipulation facile des chaînes de données, plus tard adopté par des entreprises telles que Google pour développer d'autres applications logicielles. La croyance de McCarthy que "*chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut en principe être si précisément décrit qu'une machine peut être faite pour le simuler*"[McCarthy et al., 2006], souligne une idée clé en IA : les machines pourraient un jour être capables d'imiter entièrement la façon dont les humains pensent et apprennent. Cette possibilité théorique a commencé à se matérialiser avec le travail des chercheurs Newell and Simon [1956]. Leur création, le *Logic Theorist*, est considéré comme le premier programme d'intelligence artificielle, capable d'imiter les capacités de résolution de problèmes d'un être humain. Cela a démontré que les machines pouvaient non seulement calculer mais aussi s'engager dans des processus de pensée plus complexes, tels que le raisonnement. Entre les années 1960 et 1970, les progrès en informatique, tels que l'amélioration des capacités de stockage et de traitement des données, ont favorisé la recherche en IA. Les algorithmes de base d'apprentissage automatique ont été développés durant cette période. Des

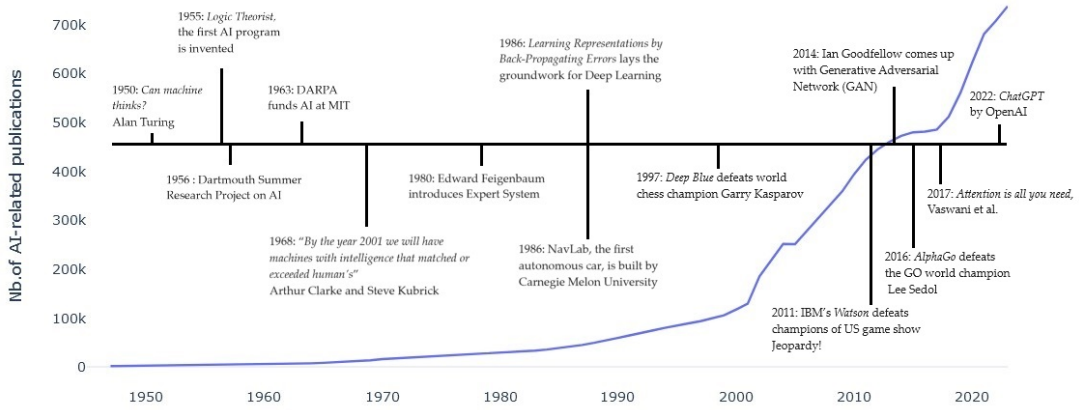
agences telles que la DARPA ont investi dans l'IA, en se concentrant initialement sur la traduction et la transcription automatiques des langues. Les années 1980 ont vu une augmentation du financement et du développement des algorithmes en IA, soulignée par les travaux de John Hopfield [1982] et David Rumelhart et al. [1985] sur les techniques d'apprentissage profond. Depuis les années 1990 jusqu'au début des années 2000, les chercheurs en IA ont atteint des jalons clés, incluant la défaite du champion du monde d'échecs Garry Kasparov avec *Deep Blue*, un système expert jouant aux échecs fonctionnant sur un superordinateur conçu spécialement par IBM [Campbell et al., 2002]. Aujourd'hui, avec l'augmentation de la puissance de calcul et de la disponibilité des données, des outils d'IA tels que ChatGPT, DALL-E et AlphaFold illustrent l'évolution rapide du domaine vers une IA générale, permettant aux logiciels de réaliser de manière autonome des tâches complexes autrefois considérées comme le domaine exclusif des humains.[7]

Ces avancements rapides ont suscité de nombreuses recherches académiques sur ses interactions complexes avec les systèmes économiques. Un domaine clé de débat se concentre sur la question de savoir si l'IA doit être classifiée comme une Technologie à Usage Général (GPT) ou comme une Invention de Méthode d'Inventions (IMI). Ce débat implique des chercheurs tels que Cockburn et al. [2018], Klinger et al. [2018], Agrawal et al. [2019c], Klinger et al. [2020], Bianchini et al. [2022], Vannuccini and Prytkova [2023]. Un corpus croissant de littérature suggère que l'IA peut remodeler la manière dont nous produisons des connaissances, à la fois au sein de nombreux champs scientifiques et entre eux, augmentant les impacts sur plusieurs voies, de la découverte scientifique à la productivité des scientifiques [Cockburn et al., 2018, Galindo-Rueda, 2020, OECD, Bianchini et al., 2022, Borsato and Lorentz, 2023].[8] En tant que GPT, la nature omniprésente de l'IA dans de multiples secteurs change radicalement le paysage technologique, similairement aux GPT du passé tels que l'électricité ou l'Internet. D'autre part, en tant qu'IMI, l'importance de l'IA réside

[7] L'IA générale, également connue sous le nom d'IA forte [Kurzweil, 2005] ou d'IA de niveau humain [Roser, 2023] est un logiciel d'intelligence artificielle qui possède une intelligence semblable à celle de l'homme et a la capacité d'apprendre de manière autonome. L'objectif est de permettre au logiciel d'exécuter des tâches pour lesquelles il n'a pas été spécifiquement formé ou développé. Les technologies d'IA existantes fonctionnent dans des paramètres prédéfinis. L'IA générale est un effort intellectuel visant à créer des systèmes d'IA qui disposent d'une autorégulation indépendante, d'un niveau de conscience de soi et de la capacité d'acquérir de nouvelles compétences. La réalisation d'une IA générale avec des capacités de niveau humain est encore un concept théorique et un objectif principal de recherche [Amazon, 2023].

[8] Toutefois, des opinions différentes persistent sur la diffusion généralisée des technologies basées sur l'IA. Nous renvoyons le lecteur intéressé à Vannuccini and Prytkova [2023] et aux preuves empiriques dans McElheran et al. [2023].

Figure 1: Chronologie de l'IA et nombre de publications au fil du temps



Notes : Nb. de publications liées à l'IA, source OpenAlex – élaboration propre

dans l'amélioration et l'automatisation du processus même d'innovation, en faisant un outil clé pour l'accélération de la découverte scientifique et l'expansion des connaissances combinables. Le débat en cours sur la définition de l'IA reflète sa nature évolutive et ses applications en expansion, tandis que certains chercheurs plaident pour une définition étroite et technique axée sur l'apprentissage machine et le traitement des données, d'autres appellent à une perspective plus large incluant l'IA dans les dimensions sociétales, économiques et éthiques [Klinger et al., 2020, Chubb et al., 2021]. Une grande partie de cette dissertation (chapitre 1 et 2) contribuera à cette littérature sur la diffusion et l'impact de l'IA/ML dans le système scientifique.

L'impact socio-économique de l'IA. L'adoption de l'IA a commencé dans les sciences physiques puis s'est étendue aux sciences de la vie, aux sciences sociales, aux arts et aux lettres [Gefen et al., 2021]. Aujourd'hui, les domaines d'application couvrent un large éventail d'industries, allant de la santé aux transports, avec des impacts clés sur la dynamique du marché du travail et d'autres problèmes sociétaux - par exemple, le changement climatique, les inégalités de revenus, le développement durable.

Par exemple, dans le domaine de la santé, les systèmes basés sur l'IA améliorent le diagnostic précoce, la médecine personnalisée et les résultats pour les patients [Jiang et al., 2017, Bohr and Memarzadeh, 2020]. Les algorithmes de ML peuvent analyser d'immenses données médicales pour identifier des motifs et prédire des maladies avec

une précision accrue [Uddin et al., 2019, Ngiam and Khor, 2019] tandis que les robots activés par l'IA en chirurgie augmentent la précision et réduisent les erreurs [Park et al., 2022]. De plus, les algorithmes d'IA deviennent cruciaux pour l'analyse des images médicales telles que les radiographies, les scanners CT et les IRM, aidant à diagnostiquer et détecter des maladies telles que le cancer de la peau et les maladies pulmonaires, et à évaluer les risques cardiovasculaires [Esteva et al., 2017]. Plus récemment, une étude publiée dans *Nature* par des chercheurs du MIT a montré l'utilisation de l'IA pour découvrir une classe de composés capables de tuer une bactérie résistante aux médicaments qui cause plus de 10 000 décès aux États-Unis chaque année [Wong et al., 2023].

Dans la lutte contre le changement climatique, l'IA joue un rôle significatif dans le développement de solutions durables. Les algorithmes d'IA permettent une gestion efficace des sources d'énergie renouvelable, réduisant la dépendance aux combustibles fossiles [Reichstein et al., 2019, Song and Roh, 2021, Bianchini et al., 2023b]. Elle prévoit également les modèles climatiques et informe la planification en cas de catastrophe, aidant à réduire les émissions de gaz à effet de serre et à économiser de l'énergie [Ramli et al., 2008]. Par exemple, DeepMind de Google a considérablement amélioré l'efficacité énergétique des parcs éoliens en prédisant avec précision les motifs du vent [GoogleDeepMind, 2023].

De plus, de nombreux aspects de la littérature soulignent le rôle de l'IA dans l'affectation des dynamiques industrielles par la mise en œuvre croissante de techniques de pointe pour l'analyse de grands ensembles de données et l'automatisation des tâches. Comme également mis en évidence par Brynjolfsson and McAfee [2014], et Borsato and Lorentz [2023], les applications de l'IA telles que les processus de production automatisés et les techniques d'optimisation augmentent l'efficacité et la compétitivité, stimulant la croissance économique par le biais de produits et services innovants [Agrawal et al., 2019a]. Acemoglu et al. [2022] a analysé comment plus de 300 000 entreprises américaines ont adopté des technologies avancées entre 2016 et 2018. Cette recherche se concentre sur cinq domaines clés - à savoir, l'IA, la robotique, les logiciels spécialisés pour des fonctions commerciales spécifiques, les équipements dédiés aux tâches automatisées, et les systèmes informatiques et applications basés sur le cloud qui montrent que l'intégration de ces technologies correspond à un changement dans la demande de main-d'œuvre vers des travailleurs capables de gérer et d'opérer des systèmes avancés. Ce changement, selon les auteurs, révèle une tendance vers un paysage commercial plus avancé technologiquement, interconnecté

Table 2: Principaux récits relatifs à l'IA

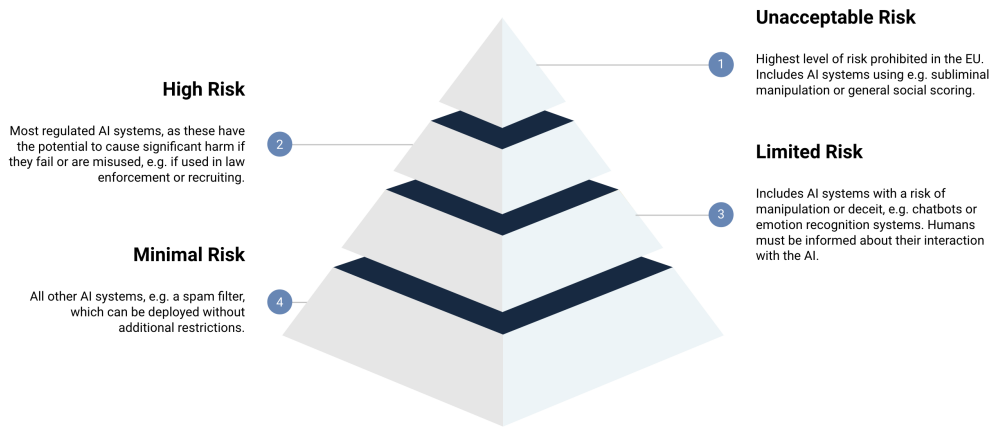
Domaine	Espoirs	Peurs	Débat
Santé	Immortalité	Déshumanisation	L'homme conquiert l'immortalité tandis que d'un autre côté, les humains perdent leur essence, abandonnant valeurs et émotions.
Emploi	Liberté	Remplacement des emplois/Obsolescence	Les humains seront libérés des tâches fastidieuses ou fatigantes, qu'elles soient physiques ou cognitives. La représentation opposée est le risque lié à ce tournant technique.
Sociologie	Gratification	Aliénation	L'IA et les robots satisfont tous les désirs humains, mais en revanche, le scénario opposé prédit que les individus n'interagiront qu'avec les technologies plutôt qu'avec d'autres personnes.
Surveillance	Sécurité	Soulèvement	Le scénario optimiste prédit que de nouveaux outils permettront aux nations et aux communautés d'assurer la sécurité de tous, tandis que de l'autre côté, il y a le récit emblématique de la science-fiction où l'IA prendra le contrôle des humains.

Notes: Élaboration personnelle basée sur Cave and Dihal [2019]

et centré sur les données. En regardant plus loin, les systèmes pilotés par l'IA, par exemple, les moteurs de recommandation, ont transformé des secteurs tels que le commerce électronique [Bughin et al., 2018]. Il existe également des preuves mettant en lumière les effets positifs des technologies de l'IA sur l'augmentation de la performance innovante et de la rentabilité des entreprises [Khin and Ho, 2018, Leusin, 2022, Rammer et al., 2021]. L'IA a également été un catalyseur dans l'économie des plateformes numériques, avec des entreprises comme Uber et Airbnb créant de nouvelles opportunités commerciales [Brynjolfsson and McAfee, 2017].

Bien que l'IA influence les dynamiques industrielles en termes de croissance et de durabilité, elle présente des effets ambigus (à double tranchant) sur le marché du travail [Brynjolfsson et al., 2018, Acemoglu and Restrepo, 2019a,b, Aghion et al., 2019, Domini et al., 2021, Bordot, 2022, Mondolo, 2022]. Le potentiel d'automatisation des tâches peut entraîner un déplacement des emplois dans certains secteurs – en particulier ceux impliquant des tâches routinières ou répétitives [Brynjolfsson and McAfee, 2014], mais il crée également de nouvelles opportunités d'emploi et améliore la productivité, contribuant à la croissance économique [Brynjolfsson and McAfee, 2017]. L'impact varie selon les taux d'adoption de l'IA, la nature des tâches automatisées, et l'adaptation de la main-d'œuvre [Autor et al., 2020].

Figure 2: Les quatre classes de risque de l'Acte IA de l'UE



Notes: Source Trail-ml

De plus, au-delà de son impact tangible sur différents secteurs, le rôle de l'IA s'étend à la résolution de défis sociaux, introduisant des considérations éthiques significatives et des dilemmes. Une préoccupation principale est son potentiel à perpétuer les biais et la discrimination. Les systèmes d'IA, entraînés sur de grands ensembles de données, peuvent amplifier les biais inhérents si les données contiennent des informations préjugées ou discriminatoires [O'Neil, 2016].

Ce risque est exacerbé car ces systèmes reposent souvent sur une vaste quantité de données personnelles pour fonctionner efficacement. La protection et l'utilisation appropriée des données sont impératives pour maintenir la confiance dans les technologies de l'IA. Des réglementations et des mécanismes robustes sont essentiels pour protéger les droits à la vie privée des individus tout en permettant des applications bénéfiques de l'IA. L'*Acte IA*, adopté par l'Union Européenne en 2023, constitue un premier pas dans cette direction. En tant que cadre global, il réglemente le développement, la commercialisation et l'utilisation de l'IA dans plusieurs secteurs, à l'exception du militaire. L'Acte IA se caractérise par ses exigences basées sur le risque pour les systèmes d'IA, les interdictions de certaines pratiques d'IA nuisibles et sa portée extraterritoriale, influençant potentiellement la gouvernance de l'IA à l'échelle mondiale, autant que le Règlement Général sur la Protection des Données (RGPD) de l'UE. Le cadre réglementaire définit quatre niveaux de risque pour les systèmes d'IA représentés dans un modèle en pyramide (Figure 2) qui classe les

systèmes d'IA selon leurs risques potentiels et le niveau de surveillance réglementaire requis. Au plus haut niveau de la pyramide, les systèmes d'IA sont classés comme "Risque Inacceptable" s'ils posent une menace extrême à la sécurité et aux droits fondamentaux. Cette catégorie inclut les systèmes qui utilisent la manipulation subliminale ou permettent le scoring social, qui pourraient avoir un impact sévère sur les opportunités des individus et l'accès aux ressources. De tels systèmes sont interdits dans l'Union Européenne. Ensuite, il y a les systèmes d'IA à haut risque, qui incluent des applications dans des domaines critiques tels que la santé, l'application de la loi et la prise de décision judiciaire. Ces systèmes sont soumis à des exigences strictes de conformité en raison de leur potentiel de causer des dommages significatifs en cas de défaillance ou de mauvais usage. Les systèmes au niveau suivant de la pyramide sont ceux qui présentent un risque limité. Cela inclut des technologies telles que les chatbots ou les systèmes de reconnaissance des émotions qui ont le potentiel de manipuler ou de tromper. Bien qu'ils ne soient pas intrinsèquement nuisibles, ces systèmes nécessitent une transparence claire pour garantir que les utilisateurs sont pleinement conscients qu'ils interagissent avec de l'IA, préservant ainsi la prise de décision informée. À la base de la pyramide, les applications d'IA à risque minimal, telles que les filtres anti-spam ou les systèmes automatisés de recommandation vidéo, présentent des risques négligeables pour la société et bénéficient donc des normes réglementaires les plus indulgentes, qui favorisent l'innovation tout en maintenant des mesures de sécurité essentielles.

Dans le domaine de la protection des données et de la sécurité, les défis sont amplifiés par la collecte et l'analyse avancées des données personnelles par les technologies de l'IA. Protéger les informations sensibles face à l'augmentation de l'échelle et de la complexité du traitement des données nécessite des mesures de sécurité robustes, des technologies renforçant la confidentialité et une législation complète. Le chiffrement, les contrôles d'accès, les méthodes d'anonymisation, les systèmes de stockage de données sécurisés, les audits de sécurité réguliers et la formation à la protection des données sont des composants clés pour atténuer les risques associés aux données personnelles [Murdoch, 2021, Garrido et al., 2022, Jordan et al., 2022]. À côté de l'impératif de sécurisation des données, il existe un défi tout aussi crucial d'assurer l'équité dans les algorithmes de l'IA. La recherche en cours sur l'algorithmique se concentre sur le développement de cadres mathématiques et de méthodologies pour traiter les biais dans l'IA [Dwork et al., 2012]. Atteindre des algorithmes d'IA non biaisés est en fait non seulement une nécessité éthique mais cela a également des

implications économiques. Les algorithmes biaisés peuvent perpétuer les inégalités, limiter les opportunités pour les groupes marginalisés et éroder la confiance dans les systèmes d'IA, entravant leur adoption [Mittelstadt et al., 2016]. Prioriser l'équité dans la conception et les processus de formation des algorithmes est donc essentiel pour favoriser une société plus inclusive et équitable.

L'application des Technologies Numériques Avancées (ADTs), en particulier de l'IA, pour le bien social a été un sujet de débat considérable. Des analyses critiques, comme dans Taylor [2016], explorent la complexité de traiter les grandes données comme un bien public, mettant en évidence les conflits entre droits, devoirs et revendications (voir aussi Savona [2019]).

De même, Moore [2019] évalue le discours ambigu de "l'IA pour le bien social", en soulignant la nécessité de plus de clarté et de réflexion, en particulier dans le débat public et la science des données. Dans un numéro spécial introduit par Cowls [2021], les complexités de la définition de ce qu'est un "bien social" dans le contexte des technologies basées sur l'IA sont étudiées, en examinant les positions éthiques et les dynamiques de pouvoir impliquées. En outre, Floridi et al. [2020] examine le potentiel de l'IA à répondre aux problèmes sociaux et à contribuer au bien-être de la société, y compris les développements durables sur le plan environnemental. Ces travaux soulignent collectivement l'importance d'une compréhension nuancée des implications sociétales et éthiques de l'IA et des technologies de données, remettant en question les discours simplifiés et plaidant pour une approche réfléchie dans l'évaluation de leurs impacts divers. Ce qui précède ouvre la discussion sur le rôle de l'IA en termes de développement économique en général, et il est pertinent de considérer l'Agenda des Nations Unies pour le Développement Durable, un cadre holistique [UN-General-Assembly, 2015, TWI2050, 2019]. Cet agenda établit un plan global pour stimuler les progrès dans de multiples domaines, ancré dans les 17 Objectifs de Développement Durable (ODD). Ainsi, le cadre des ODD encourage une évaluation approfondie de l'impact des Technologies Numériques Avancées (ADTs), y compris l'IA, sur plusieurs phénomènes interconnectés [Chui et al., 2018, Goralski and Tan, 2020, Vinuesa et al., 2020, Cowls, 2021, Guenat et al., 2022, Bianchini et al., 2023a]. La littérature sur ce sujet a maintes fois souligné le rôle crucial des technologies pilotées par l'IA dans la réalisation des ODD [Goh, 2021, Jindra and Leusin, 2022, Series, 2018].

Dans une ère où les avancées technologiques sont essentielles à l'évolution sociétale, cette thèse vise à décrire de manière exhaustive le rôle multifacette de l'IA au sein

du paradigme scientifique et ses implications sociétales plus larges. La recherche se concentre sur trois dimensions principales : (i) la contribution de l'IA à la recherche scientifique, (ii) l'influence du secteur privé dans le développement et l'application de l'IA, et (iii) l'interaction entre la perception publique et le rôle sociétal de l'IA.

À la lumière de ces faits, le pivot de cette dissertation tourne autour d'une enquête tripartite. Le premier aspect concerne la délimitation de la manière dont l'IA affecte l'écosystème scientifique (par exemple, les collaborations), l'examen de l'intégration connexe entre les différentes disciplines scientifiques, et les impacts transformateurs qui en découlent. La deuxième partie se concentre sur la manière dont la collaboration entre les secteurs public et privé, en particulier dans le domaine de la technologie émergente des *Transformers*, permet aux scientifiques utilisant cette technologie d'augmenter l'impact de leur activité de recherche. Le troisième aspect analyse l'impact des perceptions sociales sur l'intégration de l'IA dans le tissu social, en considérant de quelle manière les attitudes publiques, la sensibilisation et les préoccupations peuvent influencer et déterminer le développement et le déploiement des technologies de l'IA.

Dans le **chapitre un**, nous explorons la tendance émergente de l'interdisciplinarité, un concept qui a suscité un intérêt significatif dans la politique scientifique, particulièrement mis en évidence lors de la pandémie de COVID-19. L'épidémie a encouragé les épidémiologistes et les chercheurs médicaux à non seulement utiliser les ressources au sein de leurs disciplines, mais aussi à rechercher de nouvelles idées et des collaborations externes. Parmi celles-ci, l'intégration avec l'IA est apparue comme une alliance particulièrement prometteuse. Cependant, les projets collaboratifs combinant deux des sujets les plus en vue dans les discussions scientifiques et sociétales – l'IA et la COVID-19 – ont montré des degrés de productivité variés.

Il devient évident que la nature multidisciplinaire de la recherche sur l'IA et la COVID-19 nécessite la formation d'équipes diverses et complémentaires comprenant des chercheurs de différents domaines. Notre étude vise à examiner les facteurs qui contribuent à des collaborations réussies entre experts de domaines spécifiques et spécialistes de l'IA. Bien que des recherches précédentes indiquent que les collaborations les plus fructueuses se produisent grâce à des efforts interdisciplinaires au sein de domaines étroitement liés, il y a peu d'investigation sur les résultats tangibles de ces collaborations.

Pour combler cette lacune, notre analyse se concentre sur l'adoption des techniques d'IA dans la recherche sur la COVID-19. Nous avons utilisé des données provenant de trois bases de données distinctes : CORD-19, Semantic Scholar et Altmetric. Cette analyse est soutenue par un ensemble de nouvelles métriques conçues pour évaluer l'interdisciplinarité en relation avec l'IA à deux niveaux : la diversité des équipes (incluant la participation des experts en IA dans la recherche sur la COVID-19) et la diversité épistémologique (mesurant les connaissances effectivement utilisées dans chaque article de recherche).

Nos résultats sont triples. Premièrement, nous observons que les deux formes de diversité – diversité des équipes et diversité épistémologique – sont positivement associées à diverses formes d'impact, telles que le nombre de citations, l'attention des médias et la portée interdisciplinaire. Deuxièmement, une tendance notable est l'association négative entre l'implication des experts en IA et l'impact de la recherche, suggérant des défis dans la collaboration entre les experts du domaine et les experts en IA pour produire une science influente. Enfin, notre analyse indique que la diversité épistémologique a plus de poids pour l'impact au-delà de la sphère académique. En cartographiant la diffusion de l'IA dans la recherche scientifique et son impact, ce chapitre vise à contribuer à une meilleure compréhension de la manière dont les technologies computationnelles sont précieuses pour relever les défis sociétaux actuels et futurs.

Chapitre deux contribue à la compréhension des relations entre le monde académique et l'industrie dans le développement des modèles d'IA / ML et l'impact de ces modèles sur la découverte scientifique, avec un accent particulier sur la technologie des Transformers – un sous-ensemble révolutionnaire de l'apprentissage profond.

L'apparition des Transformers, introduite dans l'article fondateur "Attention is All You Need" [Vaswani et al., 2017] lors de la conférence Neural Information Processing Systems de 2017, a redéfini la trajectoire de l'IA dans divers domaines. Ces architectures sont désormais omniprésentes dans une multitude d'applications scientifiques, couvrant le traitement du langage naturel, la vision par ordinateur, l'apprentissage par renforcement, la biologie, et bien plus encore [Lin et al., 2022, Han et al., 2023, Wang et al., 2023]. Bien que l'IA révolutionne la découverte scientifique et la productivité [Cockburn et al., 2018, Bianchini et al., 2022], la technologie spécifique à l'origine de ces avancées reste floue. De plus, le rôle historique du secteur privé dans ces développements n'est pas bien compris.

Des collaborations remarquables, telles que la conférence de Dartmouth en 1956 – réalisée grâce aux efforts collaboratifs d'IBM, Bell Labs et MIT – et des projets contemporains tels que AlphaGO, résultant de la collaboration entre Google et les universités de Stanford et d'Oxford, mettent en évidence le rôle crucial des partenariats public-privé dans l'évolution de l'IA.

Dans cette exploration, nous examinons l'interaction dynamique de ces collaborations dans la promotion de l'innovation technologique et de la croissance scientifique. L'analyse se déroule à travers une série d'approches méthodologiques, incluant l'utilisation d'un ensemble de données robuste comprenant 113 publications transformatrices classées en applications textuelles, visuelles et autres, ainsi que des données de citations exhaustives mettant en évidence l'influence omniprésente des Transformers dans la recherche scientifique.

Notre étude utilise des modèles de Différence-en-Différences (DiD) pour quantifier les impacts de l'adoption des Transformers sur la production scientifique, montrant que les articles développés par le biais de collaborations entre le monde académique et l'industrie non seulement reçoivent plus de citations, mais présentent également une plus grande nouveauté par rapport à leurs homologues. Ces preuves empiriques soulignent le rôle significatif de l'implication du secteur privé dans l'amélioration de l'impact et du caractère disruptif de la recherche scientifique utilisant les technologies des Transformers.

De plus, nous discutons des implications plus larges de ces résultats dans le contexte des débats en cours sur la privatisation de la recherche en IA et les tendances à la monopolisation des grandes entreprises technologiques. Nos résultats suggèrent fortement que de tels partenariats sont une *condicio sine qua non* pour atteindre le plein potentiel des avancées scientifiques impulsées par l'IA. Les collaborations stratégiques entre le monde académique et l'industrie ne sont pas seulement bénéfiques, mais essentielles, servant de pierre angulaire pour faire progresser les technologies de pointe et garantir que ces avancées catalysent un enrichissement scientifique généralisé.

Le discours s'étend à une évaluation critique du rôle de ces partenariats dans la mitigation des risques associés à la concentration du pouvoir technologique, plaidant pour des politiques qui promeuvent l'innovation collaborative tout en garantissant un accès équitable à la technologie.

Dans le **chapitre trois**, nous enquêtons sur la dimension sociétale de l'IA, reconnaissant le public comme des participants actifs à son développement et met-

tant en avant l'interaction dynamique entre la technologie et la société à l'ère de la 4IR. Le paysage numérique en évolution rapide, catalysé par la 4IR, souligne l'intégration de technologies de pointe comme l'IA, la robotique et la blockchain dans les cadres sociétaux. Cette intégration technologique ne redéfinit pas seulement les structures économiques, mais influence profondément les normes culturelles et les interactions sociales. Alors que les sociétés s'efforcent de s'adapter à ces changements, comprendre la perception publique et l'impact sociétal de ces technologies devient crucial pour favoriser la confiance du public et aligner les avancées technologiques avec les besoins et les valeurs sociétales. Le discours entourant la 4IR est riche en récits contrastés, allant des visions utopiques de capacités et d'efficacités accrues aux craintes dystopiques concernant l'érosion de la vie privée, la perte d'emploi et le désengagement social. La littérature montre une dichotomie dans le sentiment public qui oscille entre l'optimisme pour l'autonomisation technologique et l'anxiété concernant la perte de contrôle et d'identité. Cette polarisation est évidente dans divers domaines, des avantages pour la santé aux préoccupations concernant la surveillance et les biais algorithmiques.

En utilisant un ensemble de données multi-pays comprenant des tweets et des articles de presse, dans ce chapitre, nous avons employé l'analyse de sentiment et des modèles d'apprentissage automatique pour explorer le discours public relatif aux technologies de la 4IR. L'analyse couvre six pays européens, capturant des opinions publiques diverses façonnées par des contextes culturels et économiques. La méthodologie intègre l'analyse de sentiment pour évaluer les émotions du public et des classificateurs d'apprentissage automatique pour identifier les thèmes prédominants dans les discussions liées à l'IA, à la robotique et à d'autres technologies de la 4IR.

Les résultats révèlent une polarisation de l'opinion publique, avec une baisse significative des perspectives neutres et une augmentation des sentiments nettement positifs ou négatifs. Cette tendance suggère un déplacement sociétal vers des positions plus définies sur les technologies de la 4IR, probablement motivé par une sensibilisation et un engagement accrus. L'analyse de sentiment indique un optimisme général quant aux avantages potentiels de ces technologies, notamment pour améliorer la qualité de vie et les opportunités économiques. Cependant, les préoccupations concernant la vie privée et l'utilisation éthique de la technologie persistent, reflétant des appréhensions généralisées quant à l'utilisation abusive des données et aux implications des systèmes autonomes. En comprenant les divers sentiments publics et les considérations éthiques, les décideurs politiques peuvent élaborer des politiques tech-

nologiques plus inclusives et tournées vers l'avenir. De plus, les résultats soulignent la nécessité de programmes d'éducation numérique robustes et de sensibilisation du public pour combler le fossé des connaissances et atténuer les risques associés à la désinformation et aux opinions polarisées.

Chapter 1

Interdisciplinary research in artificial intelligence: Lessons from COVID-19

This chapter was co-authored with

Stefano Bianchini, Floriana Gargiulo and Tommaso Venturini

Summary of the chapter

Artificial intelligence (AI) is viewed as one of the most promising technologies for solving global challenges. Recent years have seen a push for teamwork between experts from different fields and AI specialists, but the results of these collaborations have yet to be studied. We focus on about 15,000 papers at the intersection of AI and COVID-19 – reasonably one of the major challenges of recent decades – and show that interdisciplinary collaborations between medical professionals and AI specialists have largely resulted in publications with low visibility and impact. Our findings suggest that impactful research depends less on the overall interdisciplinary of author teams and more on the diversity of knowledge they actually harnessed in their research. We conclude that team composition can significantly influence the successful integration of new computational technologies into science and that obstacles still exist to effective interdisciplinary collaborations in the realm of AI.[1]

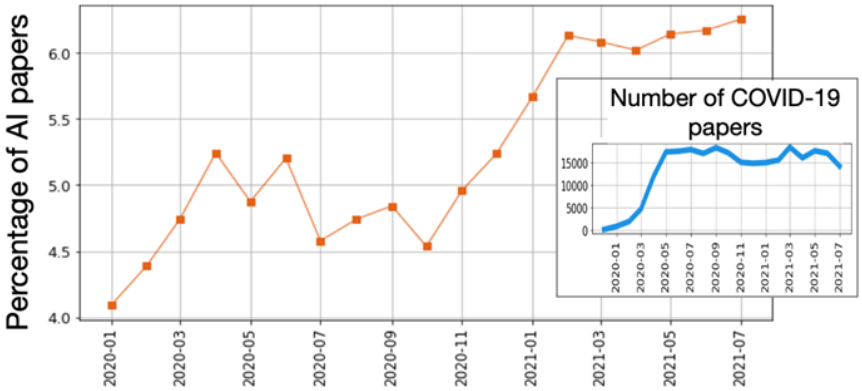
[1]The chapter has been accepted for publication and is forthcoming in *Quantitative Science Studies*.

The paper follows a conventional structure: In section 1.1, the concepts of interdisciplinarity, the impact of COVID-19 and AI. Next, in section 1.2, we provide information on the data sources and analytical methods used. Moving forward, section 1.3 show statistics about the data and the outcome from our econometric model. While, section 1.4 presents a discussion of the findings and their limitation and offers concluding remarks. Moreover additional control and model are available in appendix (section 1.5).

1.1 Introduction

Interdisciplinarity has become a buzzword in science policy. And with good reason. Disciplines have for decades – in some cases, centuries – facilitated scientific progress by providing scholars with the scaffolding of a coherent paradigm and the possibility of standing on the shoulders of their predecessors. However, disciplinary boundaries have often proved to be a stumbling block to innovation, as growing specialization makes it ever harder (though ever more necessary) to venture into unexplored research territories and combine intellectual tools originating from different traditions [Jones, 2009]. These entrenched boundaries are especially problematic when facing unprecedented research challenges that require fresh thinking and unrestrained experimentation. Such a situation presented itself recently with the outbreak of the COVID-19 pandemic. The urgency and gravity of the situation prompted researchers in epidemiology and medical science not only to mobilize all the resources available within their disciplines, but to look beyond them for new ideas and external collaborations. Among them, the alliance with artificial intelligence (AI) emerged as one of the most promising (Fig. 1.1).

Figure 1.1: COVID-19 publications mentioning AI technology



Notes: Fraction of COVID-19 papers mentioning AI technologies. Inset: Total number of COVID-19 papers. After an initial period of exponential growth, scientific production related to the COVID-19 virus stabilized in May 2020. At the same time, AI research dedicated to COVID-19 virus remained relatively marginal until summer 2020 when it began to record constant linear growth, so that by July 2021 it accounted for nearly 7% of total COVID-19 scientific production. Source: Own elaboration on CORD-19 data.

Although AI techniques have a long history, the field has recently been revived by the escalating power of computational technologies and the growing availability of data on social and natural phenomena. This has led to the development of new machine learning approaches, which have yielded remarkable results within and beyond data science [Cardon et al., 2018, Frank et al., 2019]. Recent studies have shown that AI/ML techniques are indeed changing the "way of doing science", from agenda setting and hypothesis formulation to experimentation, knowledge sharing, and public involvement, with a considerable impact on scientific practices [Cockburn et al., 2018, Agrawal et al., 2018, Xu et al., 2021, Bianchini et al., 2022, Birhane et al., 2023, Van Noorden and Perkel, 2023, Koehler and Sauermann, 2024].

The coronavirus pandemic hits at the peak of this cycle of AI hype and, unsurprisingly, many scholars quickly embraced the idea of adopting AI techniques to tackle the challenges presented by COVID-19 [DeGrave et al., 2021, Khan et al., 2021, Roberts et al., 2021].[2] Opportunities for collaborative funding have emerged globally to bring various scientific communities together, and researchers from different

[2]It is worth noting that AI is seen by many as a technological solution to meet contemporary global challenges, such as sustainable development, green transition, global health and others (see, e.g., Schwalbe and Wahl [2020], Vinuesa et al. [2020]). Yet, it is equally important to note that the benefits brought by technology are such only under proper AI governance frameworks [Truby, 2020].

backgrounds have come together to try to harness the potential of AI in COVID-19 research [Ahuja et al., 2020, Luengo-Oroz et al., 2020]. Some of these collaborations offered substantial contributions to the fight against the pandemic. By manually screening some of the most cited and visible online papers in our dataset, we did find some interesting use of AI for COVID-19 research, particularly to make sense of large archives of literature or data (cf. for example, Mistry et al. [2021], Salari et al. [2020], Wynants et al. [2020]), and some reflexive assessment of the efficacy of AI and big data approaches (cf. for example, Wang et al. [2020a], Agbehadji et al. [2020]). Many other publications at the AI/COVID-19 intersection, however, never gained much visibility or scientific traction. What can explain these contrasting outcomes?

Previous research shows that (large) interdisciplinary teams produce more cited research and high-impact papers [Wuchty et al., 2007, Fortunato et al., 2018], and that diversity – not only epistemic, but also institutional and ethnic – is beneficial for producing novel, valuable ideas [Taylor and Greve, 2006]. Teams comprising researchers with different backgrounds, methodological approaches, and experience have access to a broader pool of knowledge, which allows them to produce more creative outputs than those produced by less collaborative science [Stephan, 2012, Uzzi et al., 2013, Gargiulo et al., 2022]. This can be explained by the functional diversity of teams, that is, differences in the way scientists encode problems and attempt to solve them; as Hong and Page [2004] put it succinctly: "diversity trumps ability". Collaborative projects also serve to boost visibility by exposing scientific findings to a wider and more diverse readership [Leahey, 2016]. In the case of COVID-19 research, this suggests that collaborations between AI experts and clinicians may result in successful research outcomes, as domain specialists could provide their "on-the-ground" knowledge to identify promising areas for investigation, while technology experts could offer access to the latest computational methods.

Team diversity, however, is not without its disadvantages. Teams that are too large and heterogeneous often suffer from lower consensus-building, cognitive diversity, higher coordination costs, and emotional conflict. As diversity increases, it may become more difficult to convert specialized expertise into scientific outputs [Lee et al., 2015]. Studies show that team performances depend more on how the team interacts than on the characteristics of its members [Woolley et al., 2010], and that most successful collaborations seem to be achieved through efforts that, while interdisciplinary, combine relatively close fields [Yegros-Yegros et al., 2015].[3] Difficulties,

[3]A comprehensive review of the rich literature on the impact of interdisciplinary research is beyond

therefore, could have arisen in collaborations between AI and COVID-19 experts due to differences in their areas of expertise, and this could have resulted in less impactful and visible scientific outcomes compared to teams consisting of only AI or clinical specialists.

In this chapter, we examine the impact of interdisciplinarity by investigating a large corpus of scientific publications at the intersection of COVID-19 and AI (about 15,000 papers retrieved from the COVID-19 Open Research Dataset, CORD-19 – version 2021-08-09 – and supplemented by other metadata from Altmetric and OpenAlex), and studying which forms of interdisciplinarity are more strongly associated with scientific impact. In the remainder, we first describe the metrics of interdisciplinarity used in our study, and then link these metrics to three indicators of scientific "success", namely the number of citations, online visibility, and outreach to other disciplines.

1.2 Material and methods

1.2.1 Data

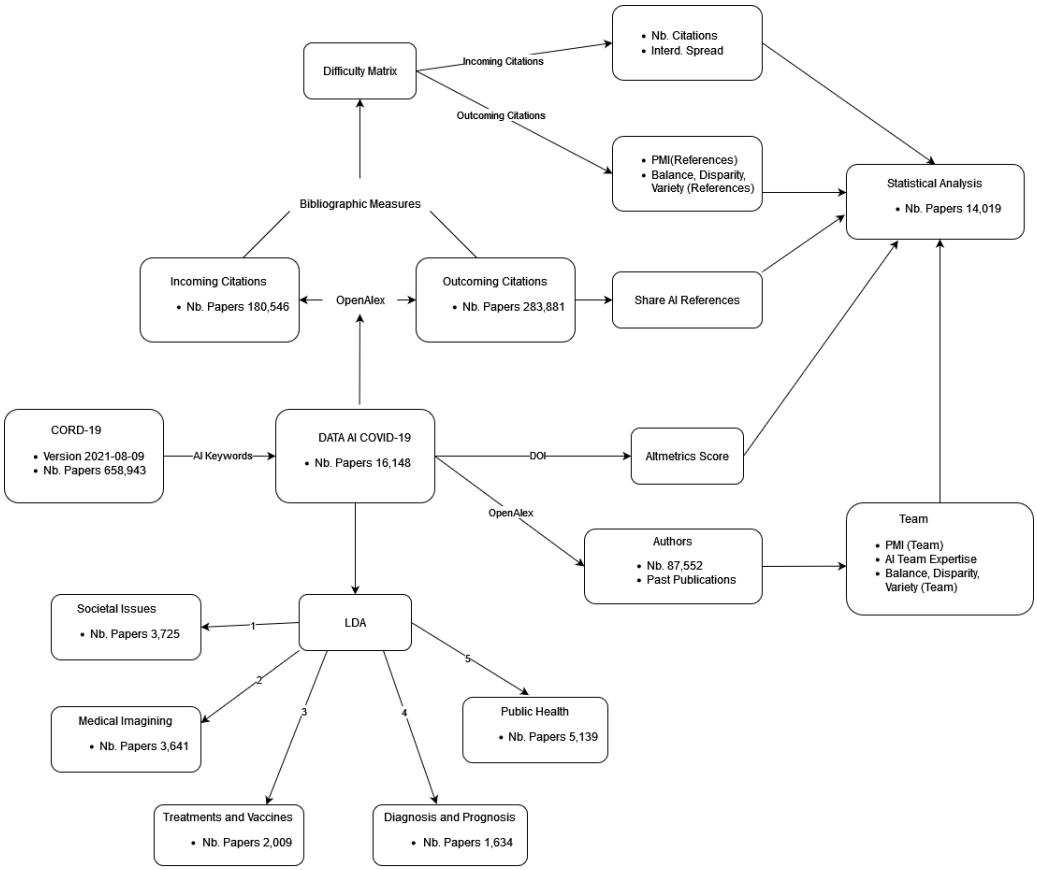
Our analysis combines data from three different databases – CORD-19, OpenAlex, and Altmetric – and is based on the pre-processing protocol (Fig. 1.2).

The COVID-19 Open Research Dataset (CORD-19) is a growing corpus of publications on COVID-19 and other coronavirus infections [Wang et al., 2020b]. It includes, in the period that we considered (from 01/12/2019 to 31/08/2021), around 600K documents from different sources, including WHO, PubMed central, bioRxiv and medRxiv. Within this large corpus, we focused specifically on a subset of publications that included, in their abstract or title, at least one keyword related to AI. Our list of around 300 AI keywords (see Appendix) was created by merging the terms mentioned in the Wikipedia AI Glossary for AI with other 'AI vocabularies' [Baruffaldi et al., 2020, Bianchini et al., 2022, Gargiulo et al., 2023].

For each paper in this subset, we retrieved additional metadata from OpenAlex. We discarded all documents with missing information and obtained a final corpus

the scope of this article. However, it is interesting to note that the question is still open and debated in the scientific community. For instance, while some studies on environmental sciences and biomedicine suggest long-term benefits of interdisciplinary approaches, particularly in terms of introducing novel ideas [Steele and Stier, 2000, Schilling and Green, 2011, Wang et al., 2015, Larivière et al., 2015, Okamura, 2019], others indicate that interdisciplinary research may reduce both scientific productivity [Leahey et al., 2017] and impact [Levitt and Thelwall, 2008].

Figure 1.2: Data preparation pipeline



of 16,148 AI publications on COVID-19 (COVID-19+AI dataset). We retrieved the metadata for all the references cited by the publications in our corpus (circa 300K unique papers) and for all the papers that cite them (c. 200K papers). OpenAlex metadata included the DOI, which we used for retrieving the 'attention score' for each paper in the COVID-19+AI dataset from the website Altmetric.com. The score provides a measure of online visibility for scholarly contents (e.g., mentions on the news, in blogs, and on Twitter; article page-views and downloads; GitHub repository watchers). Finally, we used the author identifier in OpenAlex to retrieve the previous publications of all 87,552 authors present in our corpus (around 150K papers) and the institutions to which they are affiliated.

1.2.2 Measuring interdisciplinarity

The concept of interdisciplinarity is multifaceted, often ambiguous, and there is no consensus on the definition and operationalization of interdisciplinary research (cf.

for example, Porter et al. [2007], Huutoniemi et al. [2010], Leydesdorff and Rafols [2011], Yegros-Yegros et al. [2015], Wang and Schneider [2020], Fontana et al. [2020], Fontana et al. [2022]). Here, we use different measures of interdisciplinarity that consider the diversity of team members and references cited in a paper.

Each document, i, in our data is characterized by a set of authors (A_i), a set of references and citations (R_i, C_i), a set of AI keywords, if any, (W_i), the journal where it is published (J_i), and its altmetric score (M_i). Each author, a, in our corpus is associated with his/her list of papers (P_a) and with his/her three most recent papers (P_a^3).

Using a measure inspired by pairwise mutual information and based on the co-occurrence of journals in the reference lists of all articles, we compute a matrix, D, of distances between all journals in the dataset (the more two journals are regularly cited together, the smaller is their distance). To build the distance matrix, we first calculate the mutual co–citation network among journals (where two journals are linked if they appear simultaneously in a reference list). Self-loops are removed. The network is weighted and the weights, w_{ij}, correspond to the number of co–occurrences. Normalizing these weights, we define a connection probability among journals in the following way:

$$p_{ij} = \frac{w_{ij}}{\sum_{i>j} w_{ij}} \tag{1.1}$$

The structure of this network, however, is biased by the heterogeneity in terms of the number of publications among the journals: some important relationships among small journals could be hidden by their relative size compared to large journals. For this reason, instead of using the weighted adjacency matrix of this network for calculating journal similarity, we introduce a measure based on point–wise mutual information (PMI), that is:

$$pmi_{ij} = max\{0, \frac{1}{log_2 w_{ij}} log_2(\frac{w_{ij}}{p_i p_j})\} \tag{1.2}$$

where $p_i = \sum_j w_{ij}$. This measure is a similarity ranging between 0 and 1. Hence, we obtain the distance as $D_{ij} = 1 - pmi_{ij}$.

Using this notion of distance, we define two types of interdisciplinarity metrics: the first is related to team composition (measuring the disciplinary span of the previous papers by the contributors of a paper); the second is related to the knowledge

mobilized in the paper (measuring the disciplinary span in papers' references). For each dimension (team and knowledge), we introduce a further distinction between interdisciplinarity metrics specifically related to AI, and the more general interdisciplinarity, providing us with four main different metrics:

- *AI Team Expertise* is the fraction of previous AI publications for each author, averaged over the entire team:

$$AI\ Team\ Expertise_i = \frac{1}{\#\mathcal{A}_i} \sum_{a \in \mathcal{A}_i} \frac{\#\{j \in \mathcal{P}_a | \mathcal{W}(j) \neq \{\}\}}{\#\mathcal{P}_a}$$

- *Share AI References* is the fraction of cited references related to AI:

$$Share\ AI\ References_i = \frac{\#\{j \in \mathcal{R}_i | \mathcal{W}(j) \neq \{\}\}}{\#\mathcal{R}_i}$$

- *PMI (Team)* is the average disciplinary dispersion (in term of journal distances) of team authors:

$$PMI\ (Team)_i = \frac{1}{\#\mathcal{A}_i} \sum_{a \in \mathcal{A}_i} \left(\frac{1}{3} \sum_{k \neq l \in \mathcal{P}_a^3} \mathbf{D}_{J(k)J(l)} \right)$$

- *PMI (References)* is the average distance among all the journals cited in the references:

$$PMI\ (References)_i = \frac{1}{\#(\mathcal{R}_i \times \mathcal{R}_i)} \sum_{(u,v) \in (\mathcal{R}_i \times \mathcal{R}_i)} \mathbf{D}_{J(u)J(v)}$$

The first two metrics measure the share of AI in the author teams and knowledge mobilized by the publications, respectively. The last two measure levels of general interdisciplinarity in the teams and the knowledge mobilized by the publications.

In the scientometric literature, the indicator we use to characterize interdisciplinarity is similar to the disparity measure known as the Rao-Stirling indicator [Stirling, 2007]. This measure uses another way to manage the bias of a distance based on the matrix defined by 1.1, inserting the relative frequencies of the journals in the calculation of the index:

$$\Delta(i) = \sum_{ij} w_{ij} p_i p_j \tag{1.3}$$

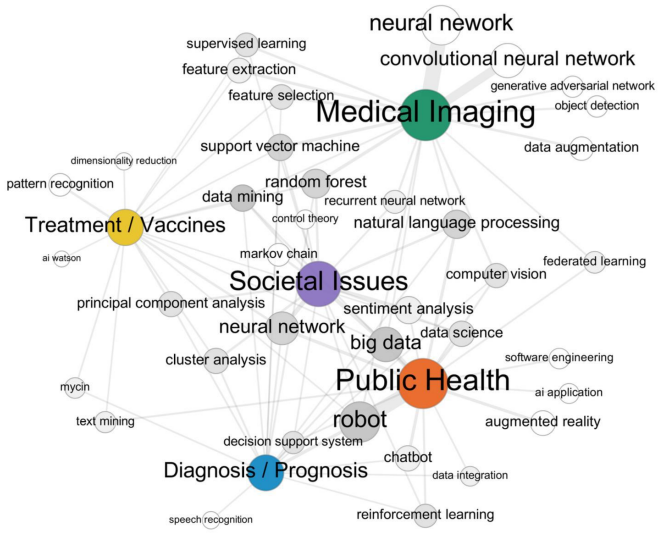
Table 1.1: Descriptive statistics

Variable	Mean	Std	Min	Median	Max
Nb. Citations	26.24	83.27	0	9	3,654
Attention Score	17.73	174.19	0	0.85	9,829.67
Interdisciplinary Spread	0.36	0.38	0	0	0.99
AI Team Expertise	0.41	0.25	0	0.38	1
Share AI References	0.23	0.25	0	0.13	1
PMI (Team)	0.34	0.35	0.01	0.27	1
PMI (Reference)	0.47	0.37	0	0.69	0.93
Balance (Team)	0.33	0.33	0	0.31	1
Balance (References)	0.44	0.37	0	0.57	1
Disparity (Team)	0.01	0.03	0	0.00	0.50
Disparity (References)	0.01	0.02	0	0.00	0.50
Variety (Team)	1.11	0.82	0	1.33	3
Variety (References)	2.67	2.57	0	3	14
AI Collaborator	0.99	0.08	0	1	1
Nb. Authors	5.64	5.77	1	4	138
Past Impact	316.67	1,151.75	0	44	38,148
Academic Age	10.77	7.33	1	9.66	89
Nb. Countries	1.53	0.99	1	1	26
Nb. References	45.41	67.50	1	30	1,620
Nb. Affiliations	2.38	3.03	1	1	78

Thus, for the sake of completeness, we also calculate the Rao-Stirling *disparity*, D_i, for teams and knowledge composition. Our metric, as well as the Rao-Stirling disparity, takes into account the relative distance among the journals/disciplines, avoiding to count as really different two journals/disciplines that are very similar in terms of contents. However, to make our study more robust, we also extend the analysis to two other dimensions traditionally used to define interdisciplinarity: *variety* and *balance*. The variety, V_i, is the count of the number of different journals where the authors previously published (for team) and of the different journals cited in the references (for knowledge). The balance, B_i, is the Gini index of the frequency associated to each journal for authors previous publications (teams) and for the references (knowledge).

Finally, we define three indicators of "success" for the publications in our corpus, namely: the *number of citations*, N_i, the *altmetric score*, M_i, and the *interdisciplinary spread*, I_i – i.e., how a paper is cited in a diverse set of disciplines - defined as:

Figure 1.3: AI application areas for COVID-19 research



Notes: Co-occurrence of AI keywords (gray nodes) and COVID-19 topics (colored nodes). Edges are weighted by the number of articles using each keyword in each topic. Nodes are sized according to their popularity (number of articles). Keywords are colored according to their degree, from white keywords specific to a single topic to dark gray keywords used in multiple topics. The consistency score of the LDA model is 0.53.

$$\mathcal{J}(i) = \frac{1}{\#(\mathcal{C}_i \times \mathcal{C}_i)} \sum_{(u,v)\in(\mathcal{C}_i\times\mathcal{C}_i)} \mathbf{D}_{J(u)J(v)}$$

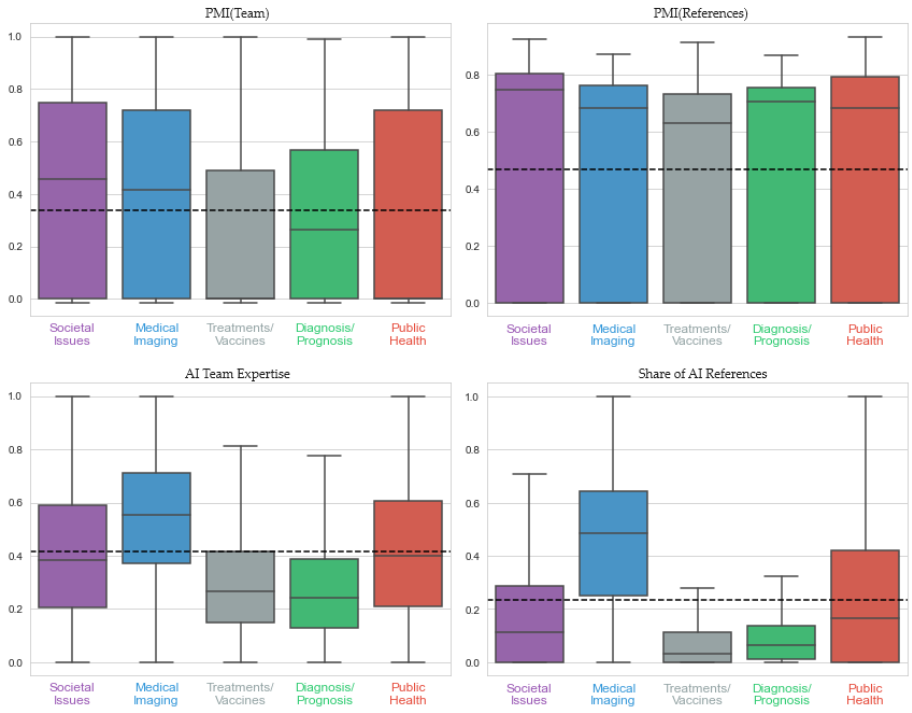
Descriptive statistics of the variables used for this study are reported in Tab.1.1

1.2.3 AI applications

By running a LDA (Latent Dirichlet Allocation) topic modelling on the abstracts of the papers in our corpus, we obtained five distinct areas in which AI/ML techniques have been applied (Fig. 1.3):

- *Societal Issues* (including epidemiology and infodemics), with some recurrent terms such as social medium, infectious disease, mental health, reproduction number, social distance, etc.;
- *Medical Imaging*: chest X-ray, chest scan, tomography, etc.;

Figure 1.4: Interdisciplinarity metrics in the different axes of COVID-19 research



Notes: General and AI-related interdisciplinarity. The dotted line represents the mean.

- *Diagnosis and Prognosis*: clinical trials, risk factors, mechanical ventilation, etc.;
- *Treatments and Vaccines*: molecular docking, spike protein, gene expression, drug discovery, etc.;
- *Public Health*: public health, contact tracing, health system, face mask, etc..

A closer reading of the terms characterizing each topic suggests that AI has found a multitude of applications [Bullock et al., 2020, Naudé, 2021, Yang et al., 2020, Piccialli et al., 2021]. In the case of societal issues, AI seems to have been used mainly for predicting the spread of disease over time and space, modeling public policy interventions (e.g., social distancing) and risk assessment, and fighting misinformation and disinformation on social media. In the case of medical imaging, what we essentially see is the deployment of deep learning models (e.g., CNN) to detect signs of COVID-19 from X-ray images and computed tomography (CT) scans. Another area of application, particularly of machine learning and deep learning, is the identification of possible treatments and vaccines, as well as the re-purposing of existing

drugs. Finally, AI appears to support the management of the public health system, for example, robotics providing assistance in the delivery of healthcare tasks.

Each application area may have required specific skills and know-how from researchers with diverse backgrounds and experiences, as well as the (re)combination of different types of knowledge. Unsurprisingly, our corpus reveals a high level of general interdisciplinarity both in the teams and in the knowledge mobilized by the publications across all research topics – with a slightly higher knowledge heterogeneity in societal issues and diagnosis/prognosis (Fig. 1.4 top).

In the case of AI, we observe very different scenarios at the topic level. Indeed, the share of teams with more AI experts is markedly higher in medical imaging and public health research, whereas teams working on vaccines, treatments, and prognosis seem to rely very little on AI knowledge (Fig. 1.4 bottom).

1.3 Results

1.3.1 What determines 'success'

We model the various impact measures – i.e., the number of citations received by the publication, the Altmetric attention score, and the interdisciplinary spread – as a function of the different interdisciplinarity metrics discussed earlier and a set of control variables, namely: *AI Collaborator* (=1 if the team includes at least one AI researcher); *Top AI Collaborator* (=1 if the team includes an AI researcher with past number of citations in the top 10° percentile of the citation distribution); *Academic Age* (average academic age of team members, in logs); *Past Impact* (average H-Index of team members based on past publications, in logs); *Nb. Countries* (number of participating countries within a team, in logs); and *Nb. References* (number of cited references, in logs). We also included a complete set of fixed effects for the month of publication and the dominant topic. The number of citations is a count variable and was modeled using a negative binomial regression. The continuous variables – attention score and interdisciplinarity spread – were modeled using ordinary least square regressions.

As shown in Table 1.2 and 1.3, the most notable result to emerge from our model is that collaborations with researchers experienced in AI (*AI Collaborator*) do not have a significant impact, and those involving a high share of researchers with established track records of AI publications (*AI Team Expertise*) receive, ceteris paribus, fewer citations, have less online visibility, and struggle to reach distant disciplines. Only

those teams that include a top AI researcher (*Top AI Collaborator*) present a positive impact on citations received by their publication, albeit that this impact is not strong. Similarly, the ratio of AI-related references (*Share AI References*) has a null or negative impact on the Altmetric attention score and interdisciplinary spread. All in all, research interdisciplinarity limited to AI does not seem to have *any* influence on the impact of COVID-19 publications, and when it does, this influence is negative.

What appears to ensure the impact of a publication is, above all else, the interdisciplinarity of the knowledge mobilized via its references, that is the actual epistemological diversity of the research conducted by a team. Regardless of how we operationalize this diversity, we find a systematic positive effect on all impact measures (except for disparity in the models on the number of citations and attention score). The effect is consistently higher than that of more classic features, such as past impact or the number of affiliated countries. The overall diversity of team members generally has a much less strong, significant and in many cases negative effect.

1.3.2 Robustness checks

The results discussed in the preceding section may depend on some methodological and arbitrary choices. We made sure that our main findings are robust to alternative specifications. First, we performed the same modeling exercise using OpenAlex 'concepts' instead of journals. We replicated the models by considering level-0 concepts (e.g., computer science) associated to each journal and level-1 concepts (e.g., machine learning), which are more granular. A journal can be associated with more than one concept; in this case, we considered the concept with the highest 'confidence score' provided by OpenAlex. Second, we considered the 5 most recent papers by each author instead of 3. Third, we re-estimated the models for the number of citations with a quasi-Poisson instead of a negative binomial regression. Finally, we excluded all publications that are still pre-print as of December 31, 2023. All results are available in Appendix.

Table 1.2: Determinants of 'success' – Nb. Citations and Attention Score

	Nb. Citations				*Attention Score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI Team Expertise	-0.219***	-0.217***	-0.199***	-0.201***	-0.498***	-0.509***	-0.479***	-0.481***
	(0.062)	(0.062)	(0.062)	(0.061)	(0.057)	(0.057)	(0.057)	(0.057)
Share AI References	0.268***	0.274***	0.271***	0.363***	-0.352***	-0.345***	-0.365***	-0.304***
	(0.057)	(0.057)	(0.057)	(0.057)	(0.053)	(0.053)	(0.053)	(0.053)
PMI (Team)	-0.386***				-0.065			
	(0.067)				(0.062)			
PMI (References)	0.482***				0.287***			
	(0.045)				(0.042)			
Balance (Team)		-0.014				-0.123***		
		(0.047)				(0.044)		
Balance (References)		0.220***				0.285***		
		(0.044)				(0.041)		
Disparity (Team)			0.474***				0.318***	
			(0.081)				(0.075)	
Disparity (References)			-0.182				-0.375***	
			(0.113)				(0.104)	
Variety (Team)				-0.052***				0.018*
				(0.012)				(0.011)
Variety (References)				0.014***				0.004***
				(0.001)				(0.001)
AI Collaborator	0.026	0.013	0.009	-0.024	-0.304**	-0.304**	-0.293**	-0.302**
	(0.134)	(0.134)	(0.134)	(0.133)	(0.123)	(0.123)	(0.123)	(0.123)
Top AI Collaborator	0.471***	0.474***	0.481***	0.458***	0.109	0.109	0.105	0.106
	(0.075)	(0.075)	(0.075)	(0.074)	(0.070)	(0.070)	(0.070)	(0.070)
Past Impact [log]	0.189***	0.184***	0.185***	0.186***	0.193***	0.194***	0.193***	0.193***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Academic Age [log]	-0.327***	-0.326***	-0.333***	-0.318***	-0.278***	-0.276***	-0.281***	-0.280***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.020)	(0.021)	(0.021)	(0.020)
Nb. Countries [log]	0.722***	0.726***	0.731***	0.678***	0.212***	0.213***	0.214***	0.197***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)
Nb. References [log]	0.195***	0.200***	0.208***	0.154***	0.086***	0.088***	0.110***	0.081***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)
Log Likelihood	-55,085	-55,109	-55,121	-54,983				
AIK	110,250	110,299	110,322	110,046				
Adjusted R^2					0.192	0.191	0.188	0.190
F Statistic					86.540***	86.027***	84.352***	85.419***
# Observations	14,019	14,019	14,019	14,019	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on two indicators of 'success': the number of citations received by the publication (Columns 1–4) and the Altmetric attention score (Column 5–8). Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.3: Determinants of 'success' – Interdisciplinarity Spread

	Interd. Spread			
	(1)	(2)	(3)	(4)
AI Team Expertise	-0.010	-0.042***	-0.068***	0.089***
	(0.013)	(0.013)	(0.014)	(0.013)
Share AI References	-0.010	0.050***	0.094***	0.097***
	(0.012)	(0.012)	(0.013)	(0.012)
PMI (Team)	0.281***			
	(0.014)			
PMI (References)	0.594***			
	(0.010)			
Balance (Team)		0.281***		
		(0.010)		
Balance (References)		0.427***		
		(0.009)		
Disparity (Team)			1.029***	
			(0.018)	
Disparity (References)			0.740***	
			(0.025)	
Variety (Team)				0.206***
				(0.003)
Variety (References)				0.004***
				(0.0002)
AI Collaborator	0.021	0.002	0.001	0.008
	(0.028)	(0.028)	(0.030)	(0.029)
Top AI Collaborator	0.018	0.024	0.030*	0.022
	(0.016)	(0.016)	(0.016)	(0.016)
Past Impact [log]	-0.002*	0.006***	0.015***	0.003**
	(0.001)	(0.001)	(0.001)	(0.001)
Academic Age [log]	-0.004	-0.017***	-0.035***	-0.022***
	(0.005)	(0.005)	(0.005)	(0.005)
Nb. Countries [log]	0.118***	0.109***	0.094***	0.093***
	(0.007)	(0.007)	(0.007)	(0.007)
Nb. References [log]	0.007***	0.008***	0.014***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.551	0.554	0.476	0.515
F Statistic	441.300***	446.700***	327.200***	383.400***
# Observations	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the interdisciplinary spread. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks $***$, $**$ and $*$ indicate significance at the 1%, 5%, and 10% levels, respectively.

1.4 Discussion

The COVID-19 pandemic sparked a global research effort to address this unprecedented event. The scientific system responded promptly to the early stages of the virus and the international scientific community called upon its diverse expertise to assess the clinical and pathogenic characteristics of the disease and to formulate therapeutic and epidemiological strategies to cope with it. Policymakers were also quick to seek advice from ethicists, sociologists, and economists on how best to deal with the crisis [Fry et al., 2020, Chahrour et al., 2020]. Against this backdrop, AI applications represented a promising approach to face many of the challenges posed by the pandemic. A number of studies focusing on the application of AI-based approach to COVID-19 research have identified various barriers and shortcomings. They include poor data quality and flow, as well as the lack of global standards and database interoperability (e.g., genetic sequences, protein structures, medical imagery and epidemiological data); the inability of algorithms to work without sufficient knowledge of the domain; overly exacting computational, architectural, and infrastructural requirements; and the legal and ethical opacity associated with privacy and intellectual property [Bullock et al., 2020, Luengo-Oroz et al., 2020, Naudé, 2020, Khan et al., 2021, Piccialli et al., 2021].

In this paper, we have analyzed the role played by different forms of interdisciplinarity, both at the team level and in the research conducted, and their repercussions on various measures of scientific impact. Our research was, in part, motivated by the fact that policy initiatives around the world have emerged – and continue to emerge – aimed at encouraging collaboration between the AI community and specialists in various domains. However, we have no direct evidence of the effectiveness of these initiatives.

Our study provides an important takeaway message for academic decision-makers: collaborations involving AI researchers did not necessarily result in more impactful science. As our analysis revealed, the visibility, relevance and spread of the publications we considered all seem to be linked to the diversity of references rather than that of authors. What generates high-impact science, in other words, is not the *possible* interdisciplinarity associated with team diversity, but the *actual* epistemological diversity hardwired into a paper.

1.5 Appendix

Table 1.4: AI search terms

abductive logic programming	boolean satisfiability problem	developmental robotics	kl one	ontology learning
abductive reasoning	brain technology	dialogue system	knowledge acquisition	open mind common sense
abstract data type	branching factor	dimensionality reduction	knowledge engineering	openai
action language	brute-force search	discrete system	knowledge extraction	opencog
action model learning	capsule neural network	distributed artificial intelligence	knowledge interchange format	partial order reduction
action selection	case based reasoning	dynamic epistemic logic	knowledge representation and reasoning	partially observable markov decision process
activation function	chatbot	eager learning	knowledge-based system	particle swarm optimization
adaptive algorithm	cloud robotics	ebert test	lazy learning	path finding
adaptive neuro fuzzy inference system	cluster analysis	echo state network	lisp	pattern recognition
admissible heuristic	cobweb	embodied agent	logic programming	predicate logic
adversarial neural	cognitive architecture	embodied cognitive science	long short term memory	predictive analytics
affective computing	cognitive computing	ensemble averaging	machine learning	principal component analysis
agent architecture	cognitive science	error driven learning	machine listening	principle of rationality
ai accelerator	combinatorial optimization	ethics of artificial intelligence	machine perception	probabilistic programming
ai application	committee machine	evolutionary algorithm	machine translation	prolog
ai applications	commonsense knowledge	evolutionary computation	machine vision	propositional calculus
ai complete	commonsense reasoning	evolving classification function	markov chain	qualification problem
aiml	computational chemistry	existential risk from artificial general intelligence	markov decision process	quantum computing
alphago	computational complexity theory	expert system	mathematical optimization	query language
ambient intelligence	computational creativity	fast and frugal trees	mechanism design	radial basis function network
answer set programming	computational cybernetics	feature extraction	mechatronics	random forest
anytime algorithm	computational humor	feature learning	meta learning	reasoning system
application programming interface	computational intelligence	feature selection	metabolic network reconstruction and simulation	recurrent neural
approximate string matching	computational learning theory	federated learning	metaheuristic	recurrent neural network
approximation error	computational linguistics	first order logic	model checking	region connection calculus
argumentation framework	computational mathematics	forward chaining	modus ponens	reinforcement learning
artificial general intelligence	computational neuroscience	friendly artificial intelligence	modus tollens	reservoir computing
artificial immune system	computational number theory	fuzzy control system	monte carlo tree search	resource description framework
artificial intelligence	computational problem	fuzzy logic	multi agent system	restricted boltzmann machine
artificial neural network	computational statistics	fuzzy rule	multi swarm optimization	rete algorithm
association for the advancement of artificial intelligence	computer automated design	fuzzy set	mycin	robot
asymptotic computational complexity	computer vision	general game playing	naive bayes classifier	robotics
attributional calculus	concept drift	generative adversarial network	naive semantics	rule-based system
augmented reality	connectionism	genetic algorithm	name binding	satisfiability
automata theory	consistent heuristic	genetic operator	named entity recognition	search algorithm
automated planning and scheduling	constrained conditional model	glowworm swarm optimization	named graph	self-management
automated reasoning	constraint logic programming	graph database	natural language	semantic analysis
autonomic computing	constraint programming	graph theory	natural language generation	semantic network
autonomous car	constructed language	graph traversal	natural language processing	semantic query sensor fusion
autonomous robot	control theory	halting problem	natural language programming	semantic reasoner
backpropagation	convolutional	hyper heuristic	network motif	semantic search
backpropagation through time	convolutional neural	ieee computational intelligence society	neural machine translation	semi supervised learning
backward chaining	convolutional neural network	image detection	neural network	sentiment analysis
bag of words model	darkforest	image recognition	neural networking	separation logic
bag of words model in computer vision	dartmouth workshop	incremental learning	neural networks	similarity learning
batch normalization	data augmentation	inference engine	neural turing machine	situation calculus
bayesian programming	data fusion	information integration	neuro fuzzy	speech recognition
bees algorithm	data integration	intelligence amplification	neuromorphic engineering	statistical learning
behavior informatics	data mining	intelligence explosion	nlp	supervised learning
behavior tree	data science	intelligent agent	nondeterministic algorithm	tensorflow
belief desire intention software model	datalog	intelligent control	nouvelle ai	text mining
bias-variance tradeoff	decision boundary	intelligent machine	np completeness	trajectory forecasting
big data	decision support system	intelligent personal assistant	np hardness	transfer learning
big o notation	deep learning	issue tree	object detection	unsupervised learning
binary tree	deepmind technologies	junction tree algorithm	occam's razor	
blackboard system	default logic	keras	offline learning	
boltzmann machine	description logic	kernel method	online machine learning	;

Table 1.5: List of non-AI related bigrams per topic/aggregate

Aggregate	*Societal Issues*		*Medical Imagining*		*Treatments and Vaccine*		*Diagnosis and Prognosis*		*Public Health*	
public health	2748	social medium	1600	chest xray	1973	molecular docking	588	mental health	653	health care	885
social medium	2283	public health	1412	xray image	1762	main protease	435	systematic metanalysis	565	public health	738
chest xray	2011	confirmed case	841	ct image	1228	immune response	388	risk factor	539	mental health	522
xray image	1782	social distancing	795	ct scan	1117	spike protein	324	intensive car	419	social medium	433
mental health	1614	infectious disease	601	chest ct	1022	molecular dynamic	305	controlled trial	405	internet thing	416
health care	1607	united states	563	transfer learning	948	signaling pathway	286	clinical trial	396	contact tracing	400
result show	1576	number case	403	computed tomography	903	drug discovery	274	logistic regression	337	social distancing	398
learning algorithm	1446	mental health	390	learning model	842	amino acid	271	health care	335	digital technology	371
learning model	1424	using learning	375	learning algorithm	697	drug repurposing	271	mechanical ventilation	332	digital health	337
using learning	1411	case death	359	experimental result	675	clinical trial	261	viral infection	254	language processing	303
social distancing	1363	air quality	345	learning method	652	healthcare system	312	cross-sectional study	305	health system	292
infectious disease	1334	march 2020	307	tomography ct	639	binding affinity	243	care unit	302	virtual reality	271
ct image	1285	health organization	304	using learning	596	gene expression	236	significant difference	296	face mask	267
chest ct	1257	world health	301	chest x-rays	522	innate immune	226	included study	296	clinical trial	262
ct scan	1236	learning model	289	learning approach	505	virtual screening	220	confidence interval	295	learning algorithm	256
learning method	1139	reproduction number	282	learning technique	465	chinese medicine	214	study conducted	295	medical student	252
confirmed case	1120	learning algorithm	276	publicly available	463	recognition receptor	213	analysis performed	285	higher education	251
computed tomography	1082	language processing	258	sensitivity specificity	449	antiviral drug	202	public health	280	infectious disease	243
learning approach	1061	learning approach	254	chain reaction	408	immune system	192	statistically significant	280	distance learning	227
transfer learning	1048	learning technique	251	cxr image	395	component analysis	191	material method	280	supply chain	226

Table 1.6: Level-0 concepts and 5 recent works per author (Nb. Citat. and Attention)

	Nb. Citations				*Attention Score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI Team Expertise	-0.203***	-0.206***	-0.237***	-0.216***	-0.514***	-0.481***	-0.505***	-0.497***
	(0.062)	(0.062)	(0.061)	(0.062)	(0.057)	(0.057)	(0.057)	(0.057)
Share AI References	0.270***	0.287***	0.320***	0.276***	-0.369***	-0.336***	-0.322***	-0.341***
	(0.057)	(0.057)	(0.057)	(0.057)	(0.053)	(0.053)	(0.053)	(0.053)
Balance 5 (Team)	0.146***				-0.074*			
	(0.044)				(0.040)			
Balance (References)	0.087**				0.199***			
	(0.041)				(0.038)			
Disparity 5 (Team)		1.288***				1.962***		
		(0.460)				(0.428)		
Disparity (References)		3.182***				0.708		
		(0.520)				(0.485)		
Variety 5 (Team)			-0.071***				-0.029**	
			(0.015)				(0.014)	
Variety (References)			0.089***				0.035***	
			(0.006)				(0.006)	
PMI 5 (Team)				-0.106**				-0.190***
				(0.046)				(0.043)
PMI (References)				0.394***				0.322***
				(0.044)				(0.041)
AI Collaborator	0.008	0.024	0.051	0.004	-0.293**	-0.284**	-0.288**	-0.300**
	(0.134)	(0.134)	(0.133)	(0.133)	(0.123)	(0.123)	(0.123)	(0.123)
Top AI Collaborator	0.480***	0.475***	0.477***	0.473***	0.109	0.106	0.106	0.108
	(0.075)	(0.075)	(0.075)	(0.075)	(0.070)	(0.070)	(0.070)	(0.070)
Past Impact [log]	0.185***	0.186***	0.191***	0.185***	0.194***	0.195***	0.195***	0.194***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Academic Age [log]	-0.327***	-0.328***	-0.334***	-0.327***	-0.281***	-0.280***	-0.281***	-0.278***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.020)	(0.020)	(0.020)	(0.020)
Nb. Countries [log]	0.733***	0.746***	0.699***	0.725***	0.216***	0.212***	0.202***	0.211***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)
Nb. References [log]	0.205***	0.206***	0.173***	0.193***	0.097***	0.102***	0.086***	0.087***
	(0.009)	(0.008)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Log Likelihood	-55,120.640	-55,121.180	-55,025.290	-55,090.670				
AIK	110,321.300	110,322.400	110,130.600	110,261.300				
Adjusted R^2					0.189	0.189	0.189	0.191
F Statistic					84.741***	84.562***	84.984***	85.722***
# Observations	14,019	14,019	14,019	14,019	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on two indicators of 'success': the number of citations received by the publication (Columns 1–4) and the Altmetric attention score (Column 5–8). Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.7: Level-0 concepts and 5 recent works per author (Spread)

	Interd. Spread			
	(1)	(2)	(3)	(4)
AI Team Expertise	-0.028**	-0.016	0.001	-0.056***
	(0.013)	(0.016)	(0.013)	(0.013)
Share AI References	-0.059***	-0.022	0.022*	-0.028**
	(0.012)	(0.015)	(0.012)	(0.012)
Balance 5 (Team)	0.293***			
	(0.009)			
Balance (References)	0.406***			
	(0.009)			
Disparity 5 (Team)		2.066***		
		(0.121)		
Disparity (References)		3.430***		
		(0.137)		
Variety 5 (Team)			0.130***	
			(0.003)	
Variety (References)			0.053***	
			(0.001)	
PMI 5 (Team)				0.212***
				(0.010)
PMI (References)				0.512***
				(0.009)
AI Collaborator	0.013	0.056	0.020	0.016
	(0.029)	(0.035)	(0.028)	(0.028)
Top AI Collaborator	0.023	0.014	0.017	0.020
	(0.016)	(0.020)	(0.016)	(0.016)
Past Impact [log]	0.007***	0.008***	0.009***	0.005***
	(0.001)	(0.002)	(0.001)	(0.001)
Academic Age [log]	-0.018***	-0.013**	-0.022***	-0.017***
	(0.005)	(0.006)	(0.005)	(0.005)
Nb. Countries [log]	0.123***	0.114***	0.100***	0.114***
	(0.007)	(0.008)	(0.007)	(0.007)
Nb. References [log]	0.026***	0.049***	0.009***	0.012***
	(0.002)	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.421	0.147	0.456	0.459
F Statistic	262.656***	62.889***	301.715***	305.413***
# Observations	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the interdisciplinary spread. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ∗ ∗ ∗, ∗∗ and ∗ indicate significance at the 1%, 5%and 10% levels, respectively.

Table 1.8: Level-1 concepts and 5 recent works per author (Nb. Citat. and Attention)

	Nb. Citations				*Attention Score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI Team Expertise	-0.221***	-0.212***	-0.242***	-0.221***	-0.505***	-0.485***	-0.505***	-0.498***
	(0.062)	(0.061)	(0.061)	(0.061)	(0.057)	(0.057)	(0.057)	(0.057)
Share AI References	0.268***	0.320***	0.353***	0.276***	-0.348***	-0.334***	-0.288***	-0.346***
	(0.057)	(0.057)	(0.057)	(0.057)	(0.053)	(0.053)	(0.053)	(0.053)
Balance lvl1 5 (Team)	-0.035				-0.103**			
	(0.045)				(0.041)			
Balance lvl1 (References)	0.268***				0.266***			
	(0.043)				(0.039)			
Disparity lvl1 5 (Team)		-0.900***				-0.135		
		(0.279)				(0.259)		
Disparity lvl1 (References)		5.512***				1.722***		
		(0.268)				(0.253)		
Variety lvl1 5 (Team)			-0.111***				-0.043***	
			(0.012)				(0.011)	
Variety lvl1 (References)			0.068***				0.034***	
			(0.004)				(0.003)	
PMI lvl1 5 (Team)				-0.435***				-0.217***
				(0.053)				(0.049)
PMI lvl1 (References)				0.614***				0.372***
				(0.050)				(0.046)
AI Collaborator	0.019	-0.020	0.026	0.092	-0.297**	-0.294**	-0.300**	-0.297**
	(0.134)	(0.133)	(0.132)	(0.134)	(0.123)	(0.123)	(0.123)	(0.123)
Top AI Collaborator	0.474***	0.498***	0.455***	0.473***	0.109	0.107	0.093	0.106
	(0.075)	(0.075)	(0.074)	(0.075)	(0.070)	(0.070)	(0.070)	(0.070)
Past Impact [log]	0.185***	0.189***	0.185***	0.183***	0.194***	0.192***	0.192***	0.194***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Academic Age [log]	-0.325***	-0.331***	-0.314***	-0.322***	-0.276***	-0.276***	-0.273***	-0.276***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)	(0.020)	(0.020)
Nb. Countries [log]	0.731***	0.715***	0.671***	0.723***	0.216***	0.195***	0.191***	0.210***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)
Nb. References [log]	0.202***	0.199***	0.159***	0.191***	0.092***	0.098***	0.070***	0.084***
	(0.009)	(0.008)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Log Likelihood	-55,108.450	-55,030.780	-54,946.760	-55,075.410				
AIK	110,296.900	110,141.600	109,973.500	110,230.800				
Adjusted R^2					0.190	0.190	0.194	0.192
F Statistic					85.584***	85.215***	87.249***	86.237***
# Observations	14,019	14,019	14,019	14,019	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on two indicators of 'success': the number of citations received by the publication (Columns 1–4) and the Altmetric attention score (Column 5–8). Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.9: Level-1 concepts and 5 recent works per author (Spread)

	Interd. Spread			
	(1)	(2)	(3)	(4)
AI Team Expertise	-0.00004	0.003	0.024*	-0.033**
	(0.014)	(0.016)	(0.014)	(0.013)
Share AI References	0.004	0.010	0.064***	0.008
	(0.013)	(0.015)	(0.013)	(0.012)
Balance lvl1 5 (Team)	0.375***			
	(0.010)			
Balance lvl1 (References)	0.424***			
	(0.009)			
Disparity lvl1 5 (Team)		2.487***		
		(0.074)		
Disparity lvl1 (References)		4.015***		
		(0.073)		
Variety lvl1 5 (Team)			0.144***	
			(0.003)	
Variety lvl1 (References)			0.031***	
			(0.001)	
PMI lvl1 5 (Team)				0.295***
				(0.011)
PMI lvl1 5 (References)				0.501***
				(0.011)
AI Collaborator	-0.004	0.037	-0.007	0.005
	(0.029)	(0.035)	(0.030)	(0.028)
Top AI Collaborator	0.011	0.010	-0.003	0.020
	(0.017)	(0.020)	(0.017)	(0.016)
Past Impact [log]	0.005***	0.003	0.002	0.003**
	(0.001)	(0.002)	(0.001)	(0.001)
Academic Age [log]	-0.016***	-0.011*	-0.012**	-0.011**
	(0.005)	(0.006)	(0.005)	(0.005)
Nb. Countries [log]	0.128***	0.090***	0.102***	0.130***
	(0.007)	(0.008)	(0.007)	(0.007)
Nb. References [log]	0.018***	0.041***	0.004**	0.010***
	(0.002)	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.520	0.309	0.509	0.550
F Statistic	390.800***	161.500***	373.600***	439.700***
# Observations	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the interdisciplinary spread. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.10: Models without pre-prints (Nb. Cit. and Attention)

	Nb. Citations				*Attention Score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI Team Expertise	-0.255*** (0.063)	-0.222*** (0.063)	-0.225*** (0.062)	-0.256*** (0.062)	-0.497*** (0.061)	-0.471*** (0.061)	-0.475*** (0.061)	-0.487*** (0.061)
Share AI References	0.200*** (0.057)	0.187*** (0.058)	0.277*** (0.057)	0.189*** (0.057)	-0.279*** (0.055)	-0.293*** (0.056)	-0.235*** (0.056)	-0.286*** (0.055)
Balance (Team)	-0.055 (0.048)				-0.104** (0.046)			
Balance (References)	0.354*** (0.045)				0.236*** (0.044)			
Disparity (Team)		0.711*** (0.083)				0.283*** (0.080)		
Disparity (References)		-0.385*** (0.114)				-0.304*** (0.110)		
Variety (Team)			-0.025** (0.012)				0.006 (0.011)	
Variety (References)			0.014*** (0.001)				0.004*** (0.001)	
PMI (Team)				-0.329*** (0.069)				-0.149** (0.066)
PMI (References)				0.573*** (0.046)				0.286*** (0.044)
AI Collaborator	0.049 (0.139)	0.041 (0.139)	0.003 (0.138)	0.054 (0.139)	-0.288** (0.134)	-0.278** (0.134)	-0.285** (0.134)	-0.291** (0.134)
Top AI Collaborator	0.427*** (0.074)	0.438*** (0.074)	0.412*** (0.074)	0.424*** (0.074)	0.136* (0.073)	0.135* (0.073)	0.134* (0.073)	0.136* (0.073)
Past Impact [log]	0.196*** (0.007)	0.197*** (0.007)	0.197*** (0.007)	0.199*** (0.007)	0.190*** (0.007)	0.189*** (0.007)	0.189*** (0.007)	0.191*** (0.007)
Academic Age [log]	-0.346*** (0.023)	-0.359*** (0.023)	-0.339*** (0.023)	-0.346*** (0.023)	-0.269*** (0.022)	-0.275*** (0.022)	-0.271*** (0.022)	-0.272*** (0.022)
Nb. Countries [log]	0.279*** (0.035)	0.292*** (0.035)	0.232*** (0.035)	0.273*** (0.035)	0.467*** (0.034)	0.477*** (0.034)	0.458*** (0.034)	0.464*** (0.034)
Nb. References [log]	0.197*** (0.009)	0.212*** (0.009)	0.156*** (0.009)	0.192*** (0.009)	0.083*** (0.008)	0.100*** (0.008)	0.073*** (0.009)	0.080*** (0.008)
Log Likelihood	-49,867.000	-49,896.000	-49,757.000	-49,839.000				
AIK	99,815.00	99,872.000	99,593.000	99,757.000				
Adjusted R^2					0.181	0.179	0.181	0.182
F Statistic					71.040***	70.100***	71.120***	71.510***
# Observations	12,333	12,333	12,333	12,333	12,333	12,333	12,333	12,333

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on two indicators of 'success': the number of citations received by the publication (Columns 1–4) and the Altmetric attention score (Column 5–8). Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.11: Models without pre-prints (Spread)

	Interd. Spread			
	(1)	(2)	(3)	(4)
AI Team Expertise	0.035***	0.066***	0.085***	-0.020
	(0.013)	(0.014)	(0.013)	(0.012)
Share AI References	0.030***	0.079***	0.078***	-0.015
	(0.011)	(0.013)	(0.012)	(0.011)
Balance (Team)	0.282***			
	(0.010)			
Balance (References)	0.459***			
	(0.009)			
Disparity (Team)		1.065***		
		(0.018)		
Disparity (References)		0.752***		
		(0.025)		
Variety (Team)			0.216***	
			(0.002)	
Variety (References)			0.004***	
			(0.0002)	
PMI (Team)				0.313***
				(0.014)
PMI (References)				0.623***
				(0.009)
AI Collaborator	-0.012	0.002	0.002	0.006
	(0.028)	(0.031)	(0.029)	(0.028)
Top AI Collaborator	0.002	0.012	0.003	-0.006
	(0.015)	(0.017)	(0.016)	(0.015)
Past Impact [log]	0.009***	0.017***	0.006***	-0.001
	(0.001)	(0.002)	(0.001)	(0.001)
Academic Age [log]	-0.022***	-0.040***	-0.027***	-0.008*
	(0.005)	(0.005)	(0.005)	(0.005)
Nb. Countries [log]	0.022***	0.040***	0.022***	0.021***
	(0.007)	(0.008)	(0.007)	(0.007)
Nb. References [log]	0.007***	0.014***	0.007***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.621	0.530	0.575	0.624
F Statistic	520.000***	356.900***	428.200***	526.500***
# Observations	12,333	12,333	12,333	12,333

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the interdisciplinary spread. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.12: Models with interaction terms (Nb. Cit. and Attention)

	Nb. Citations				*Attention Score*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI Team Expertise	-0.091	-0.138*	-0.389***	-0.078	-0.444***	-0.553***	-0.626***	-0.415***
	(0.084)	(0.078)	(0.070)	(0.084)	(0.078)	(0.072)	(0.065)	(0.077)
Share AI References	0.277***	0.271***	0.365***	0.273***	-0.343***	-0.367***	-0.303***	-0.349***
	(0.057)	(0.057)	(0.057)	(0.057)	(0.053)	(0.053)	(0.053)	(0.053)
Balance (Team)	-0.031				-0.132***			
	(0.048)				(0.044)			
Balance (References)	0.321***				0.339***			
	(0.064)				(0.059)			
Disparity (Team)		0.466***				0.332***		
		(0.082)				(0.076)		
Disparity (References)		-0.017				-0.589***		
		(0.176)				(0.164)		
Variety (Team)			-0.054***				0.017	
			(0.012)				(0.011)	
Variety (References)			0.009***				0.0001	
			(0.001)				(0.001)	
PMI (Team)				-0.396***				-0.073
				(0.068)				(0.063)
PMI (References)				0.599***				0.360***
				(0.067)				(0.062)
AI Collaborator	0.009	0.008	0.007	0.020	-0.310**	-0.287**	-0.276**	-0.313**
	(0.134)	(0.134)	(0.133)	(0.134)	(0.123)	(0.124)	(0.123)	(0.123)
Top AI Collaborator	0.470***	0.479***	0.461***	0.466***	0.109	0.107	0.104	0.108
	(0.075)	(0.075)	(0.074)	(0.075)	(0.070)	(0.070)	(0.070)	(0.070)
Past Impact [log]	0.184***	0.185***	0.186***	0.189***	0.194***	0.193***	0.193***	0.193***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
Academic Age [log]	-0.323***	-0.332***	-0.322***	-0.324***	-0.274***	-0.283***	-0.282***	-0.277***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)	(0.020)	(0.020)
Nb. Countries [log]	0.732***	0.734***	0.666***	0.728***	0.214***	0.213***	0.193***	0.214***
	(0.031)	(0.031)	(0.031)	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)
Nb. References [log]	0.199***	0.208***	0.155***	0.194***	0.087***	0.110***	0.081***	0.085***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.008)
I(AI Team Expertise, Balance (References))	-0.210**				-0.112			
	(0.098)				(0.091)			
I(AI Team Expertise, Disparity (References))		-0.399				0.502*		
		(0.319)				(0.295)		
I(AI Team Expertise, Variety (References))			0.013***				0.011***	
			(0.002)				(0.002)	
I(AI Team Expertise, PMI (References))				-0.265**				-0.162
				(0.110)				(0.101)
Log Likelihood	-55,108.000	-55,120.000	-54,972.000	-55,082.000				
AIK	110,297.000	110,323.000	110,025.000	110,247.400				
Adjusted R^2					0.191	0.188	0.191	0.192
F Statistic					83.920***	82.330***	83.950***	84.450***
# Observations	14,019	14,019	14,019	14,019	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on two indicators of 'success': the number of citations received by the publication (Columns 1–4) and the Altmetric attention score (Column 5–8). Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.13: Models with interaction terms (Spread)

	Interd. Spread			
	(1)	(2)	(3)	(4)
AI Team Expertise	-0.002	0.016	0.030**	-0.012
	(0.017)	(0.017)	(0.015)	(0.017)
Share AI References	0.049***	0.092***	0.098***	0.010
	(0.012)	(0.013)	(0.012)	(0.012)
Balance (Team)	0.287***			
	(0.010)			
Balance (References)	0.391***			
	(0.013)			
Disparity (Team)		1.038***		
		(0.018)		
Disparity (References)		0.591***		
		(0.040)		
Variety (Team)			0.206***	
			(0.003)	
Variety (References)			0.003***	
			(0.0003)	
PMI (Team)				0.281***
				(0.014)
PMI (References)				0.592***
				(0.014)
AI Collaborator	0.006	0.006	0.018	0.022
	(0.028)	(0.030)	(0.029)	(0.028)
Top AI Collaborator	0.024	0.031*	0.021	0.018
	(0.016)	(0.017)	(0.016)	(0.016)
Past Impact [log]	0.006***	0.015***	0.003**	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
Academic Age [log]	-0.018***	-0.036***	-0.023***	-0.004
	(0.005)	(0.005)	(0.005)	(0.005)
Nb. Countries [log]	0.108***	0.094***	0.091***	0.117***
	(0.007)	(0.007)	(0.007)	(0.007)
Nb. References [log]	0.008***	0.014***	0.006***	0.007***
	(0.002)	(0.002)	(0.002)	(0.002)
I(AI Team Expertise, Balance (References))	0.075***			
	(0.020)			
I(AI Team Expertise, Disparity (References))		0.348***		
		(0.072)		
I(AI Team Expertise, Variety (References))			0.004***	
			(0.001)	
I(AI Team Expertise, PMI (References))				0.004
				(0.023)
Adjusted R^2	0.554	0.477	0.518	0.551
F Statistic	436.300***	320.100***	377.100***	430.200***
# Observations	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the interdisciplinary spread. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Table 1.14: Robustness analysis – Quasi-Poisson (Nb. Citations)

	Nb. Citations			
	(1)	(2)	(3)	(4)
AI Team Expertise	-0.214**	-0.194*	-0.227**	-0.236**
	(0.107)	(0.106)	(0.106)	(0.106)
Share AI References	0.568***	0.527***	0.631***	0.550***
	(0.094)	(0.094)	(0.094)	(0.094)
Balance (Team)	0.154*			
	(0.091)			
Balance (References)	0.056			
	(0.086)			
Disparity (Team)		0.813***		
		(0.139)		
Disparity (References)		-0.846***		
		(0.204)		
Variety (Team)			0.007	
			(0.019)	
Variety (References)			0.007***	
			(0.001)	
PMI (Team)				-0.061
				(0.125)
PMI (References)				0.277***
				(0.085)
AI Collaborator	-0.038	-0.053	-0.026	-0.029
	(0.263)	(0.261)	(0.261)	(0.262)
Top AI Collaborator	0.458***	0.452***	0.458***	0.457***
	(0.090)	(0.090)	(0.090)	(0.090)
Past Impact [log]	0.216***	0.214***	0.218***	0.216***
	(0.011)	(0.011)	(0.011)	(0.011)
Academic Age [log]	-0.436***	-0.432***	-0.435***	-0.431***
	(0.040)	(0.039)	(0.039)	(0.039)
Nb. Countries [log]	0.580***	0.584***	0.557***	0.582***
	(0.041)	(0.041)	(0.041)	(0.041)
Nb. References [log]	0.261***	0.280***	0.205***	0.257***
	(0.017)	(0.017)	(0.018)	(0.017)
# Observations	14,019	14,019	14,019	14,019

Notes: The statistical model for evaluating the relationship of different interdisciplinary metrics on the number of citations received by the publications. Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

Chapter 2

Partnerships is all you need: The development of transformer technology and its impact on science

This chapter was co-authored with

Stefano Bianchini and Patrick Llerena

Summary of the chapter

This chapter provides insights into the relationships between academia and industry in the development of AI/ML models and their impact on scientific discovery, with a focus on Transformer technology, a groundbreaking subset of deep learning models. First, we study the development and adoption process of the 113 transformer architectures in the sciences. We show that transformers diffuse at an extremely high speed across virtually every scientific domain. Then, using a quasi-experimental design, we investigate whether the adoption of transformers results in a citation premium and the production of more novel discoveries. Our findings show that scientists who adopted transformers produced more impactful and (to a lesser extent) novel science than scientists who did not. Transformers developed through university-industry partnerships have a particularly strong impact on knowledge creation.

2.1 Introduction

In 2012, after a series of groundbreaking advancements in data availability, hardware capabilities, and algorithm innovations, deep neural networks (aka deep learning) began to outpace other computational approaches in a wide range of tasks, including but not limited to computer vision and natural language processing (NLP). Fast forward to 2017, the development of the *transformer* marked another milestone in the evolution of AI technology.

The transformer architecture was first presented in the seminal paper *Attention Is All You Need* [Vaswani et al., 2017] and quickly supplanted other AI/ML approaches, becoming the de facto standard for most applications involving machine intelligence. In very general terms, transformers are a category of deep learning models characterized by an innovative component called the "self-attention mechanism" (Section 2.1 for details) that significantly enhances the efficiency of an architecture in handling long-term dependencies in sequential data – i.e., data arranged in sequences where order matters, such time series, text streams, audio/video clips, but also genomics or weather data.

Since its appearance, the transformer architecture has not only captivated the scientific community but also influenced the landscape of AI/ML beyond traditional research sphere, extending to both the general public and various professional domains. Large Language Models (LLMs), built upon the transformer framework, are perhaps the most impressive example in point. Some architectures such as BERT, GPT-3, RoBERTa, or T5 [Gillioz et al., 2020] have literally redefined the state-of-the-art in NLP, achieving unprecedented performance across a large spectrum of language-related task, e.g., translation, summarization, sentiment analysis, just to name a few. The widespread adoption of these models has permeated into everyday applications, impacting several industries and the general public through tools such as GPT-4 and other generative AIs (AlphaCode, DALL-E 2, MusicLM, ...) [Gozalo-Brizuela and Garrido-Merchan, 2023, Feuerriegel et al., 2024]. And, of course, the success of transformers has resonated with legislators bringing (this type of) AI, again, into the forefront of policy discussions [Commission, 2021, OECD, 2023a]. Never before in history has our society been so closely touched and influenced by machine intelligence.

A growing body of research has recently explored the influence of generative AI – indeed, primarily transformer-based neural networks – across various domains, including labor market dynamics [Brynjolfsson et al., 2023, Eloundou et al., 2023],

economic growth and productivity [Chui et al., 2023, Trammell and Korinek, 2023], education [Baidoo-Anu and Ansah, 2023], safety and responsibility [Jo, 2023], creativity [Epstein et al., 2023], and science (see, e.g., Krenn et al. [2022], Birhane et al. [2023], OECD [2023b]). It is this latter domain, science, that holds our focus in this paper.

Rapid advances in the capabilities of AI/ML models, coupled with their broad accessibility to nearly every researcher, has sparked enthusiasm – and, admittedly, some apprehension – regarding their application in science. It comes as no surprise that AI is permeating the scientific landscape at an exceptionally rapid rate. By way of example, the volume of AI/ML publications has surged nearly five-fold in the past decade, with more than 200,000 papers by 2022 alone, accounting for approximately 5% of the total volume of scientific publications [Arranz et al., 2023].

A recent *Nature* survey of more than 1,600 researchers worldwide confirms that most respondents expect AI will soon be central to research practice [Noorden and Perkel, 2023]. Impressively, more than half anticipate that AI/ML tools – LLMs in particular – will be "very important" or "essential" to science. And although many, including us, worry that LLMs may amplify the proliferation of misinformation or make plagiarism easier, most seem to agree that such AI already provides faster ways to process data, saves time (for example by providing faster ways to write code or automate some mundane laboratory tasks), helps brainstorm for research ideas, and, eventually, helps write research manuscripts and communicate results to the community.

The diffusion and impact of AI in science prompted some scholars to classify deep neural networks as a *general method of invention* [Agrawal et al., 2018, Cockburn et al., 2018, Bianchini et al., 2022], a conceptual framework that blends the concepts of method of invention [Griliches, 1957] and general-purpose technology (GPT) [Bresnahan and Trajtenberg, 1995a]. In short, deep learning has been promoted as a versatile and broadly applicable tool for invention and innovation, carrying important externalities that extend to the broader economy. For instance, the semi-endogenous growth model recently proposed by Besiroglu et al. [2022] shows that the widespread adoption of deep learning techniques may cause a positive shock to the R&D elasticity of capital, leading to long-last effects on the rate of idea accumulation and, ultimately, economic growth. Through calibration with U.S. data, they show that if deep learning were widely adopted in the U.S. R&D sector, it would nearly double the rate of productivity growth observed over the past 70 years. More recently,

Koehler and Sauermann [2023] have suggested that, at least in the context of crowd science, AI/ML can also "manage" human workers performing research tasks and take on a variety of managerial functions such as division and allocation of tasks, coordination, motivation and learning support.

Given that transformers represent a specific and more powerful variant of "traditional" deep learning architectures, we should expect them to share the same traits of a general method of invention, particularly (i) a rapid uptake across scientific domains and (ii) a high impact on discovery. Yet despite the numerous applications of transformers in science and some anecdotes about their potential for scientific discovery (see Section 2.2 for a review), we lack systematic empirical evidence of their diffusion and impact. In this work, we aim to fill this gap. Specifically, we proceed in two steps. We first consolidate data from various source (Hugging Face, OpenAlex, and Semantic Scholar) to study the mechanics of the diffusion process of 130 transformer models in the sciences, aiming to answer some key questions: Who developed these models? Which architectures are most adopted by scientists? In which domains do these architectures find applications? And for what types of research problems? We then try to assess the quality and disruptive potential of nearly 32,000 papers, published in the period 2018-2022, using transformers in various application domains. Our overarching question here is: Does the adoption of transformers in research result in a citation premium and the production of more novel outcomes?

To approach as closely as possible the causal effect of transformers on scientific research, we leverage the introduction of the first transformer architecture [Vaswani et al., 2017], as an exogenous shock. In fact, although the scientific community, or at least part of it, had already recognized the potential of deep learning for research, it could not anticipate the superior performance of transformers, which led to the sudden popularity of this technology (as confirmed by the trends discussed in Section 2.3). Thus, the unanticipated rise of transformers provides us with an exogenous event which prompted some domain scientists to adopt transformers for their research, while others with similar characteristics – e.g., publishing in the same journal outlets, same seniority, etc. – did not.[1]

An interesting pattern that emerges from our data is the high involvement of private industry in the advancement of transformer technology. Of the 130 architec-

[1] Ahmed and Wahed [2020] employ a similar approach, using the 2012 edition of the ImageNet contest as an exogenous shock to provide causal evidence that large technology firms are increasingly making greater contributions to AI research.

tures considered in our study, 114 involve at least one author from industry, while 51 feature exclusively industry-affiliated authors. This goes in line with some recent research that has documented an increasing privatization of AI research [Ahmed et al., 2023]. According to the AI Index Report [Maslej et al., 2023], until 2014 the academic sector dominated AI research, but since then, the trend has changed and industry has taken over. The pattern can be largely explained by the fact that modern AI/ML research relies extensively on data and computational resources, both of which are assets more readily available in (large) private corporations than in academic labs [Ahmed and Wahed, 2020]. Especially nine tech giants – often referred to as the BIG9 – Google, Amazon, Apple, IBM, Microsoft and Facebook Meta in the United States and Baidu, Alibaba, and Tencent in China have been key players in the advancement of AI in recent years [Webb, 2019].[2]

Yet, academic institutions and research labs are by no means less important. A closer look at the nearly 70-year history of AI suggests that the most important milestones in the field have, indeed, emerged from collaborative efforts between university and industry (Table 2.1). By way of example, the Dartmouth Summer Research Project on Artificial Intelligence, the 1956 summer workshop considered to be the founding event of artificial intelligence as a research field, was possible thanks to the collaboration between Dartmouth College, Harvard, IBM and Bell Labs. Other emblematic events have seen universities and private companies join forces: the well-known Deep Blue chess computer, involving IBM and Carnegie Mellon University; IBM Watson, a joint effort of IBM, MIT and the University of Toronto; and AlphaGo, a project that brought together Google researchers and other universities, including Stanford and Oxford. And, incidentally, the first transformer model was proposed by researchers at Google and the University of Toronto.

The same Nature survey cited earlier confirmed that scientists very often collaborate with – or work at – companies developing AI/ML. More than half of the

[2]Note that other companies such as NVIDIA, which develops and innovates in graphics processors, or telecommunications giants like Cisco or Huawei, are vitally important for the AI ecosystem. However, they operate in relatively specific and narrow areas; the BIG9 have a much broader reach when it comes to the impact of AI. The growing influence of industry in AI research, particularly the dominance of a bunch of tech companies, has raised several concerns. For instance, there are some apprehensions that private companies may prioritize profits over ethical considerations [Attard-Frost et al., 2023], favor environmentally demanding approaches to machine intelligence [Marcus and Davis, 2019], reduce technological diversity in AI research [Mateos-Garcia and Klinger, 2023], and attract AI talents away from academia [Jurowetzki et al., 2023]. As Mittelstadt [2019] put it: "*The fundamental aims of developers, users, and affected parties do not necessarily align. [...] Public interests are not granted primacy over commercial interest*".

Table 2.1: Timeline of university-industry partnerships for AI development

Year	Event	Actors
1956	Birth of AI: Dartmouth Conference brings together researchers to discuss the future of the field	IBM, Bell Labs, Dartmouth College, MIT
1979	DARPA SUR (Speech Understanding Research) project	US Department of Defense, Carnegie Mellon University, SRI International
1980	STAIR (STanford AI Robot) project to develop a learning robot	Stanford University, Hewlett-Packard, General Motors
1980	XCON eXpert CONfigurer) expert system	Carnegie Mellon University, Digital Equipment Corporation
1997	Deep Blue defeats world chess champion Garry Kasparov	IBM, Carnegie Mellon University
2001	Semantic Web to facilitate collaboration between academia and industry	MIT, University of Maryland, Nokia
2005	Stanford University team wins DARPA's Autonomous Driving Grand Challenge	Stanford University, Volkswagen, Intel
2011	Watson AI system competes on Jeopardy! and wins	IBM, MIT, University of Toronto
2016	AlphaGo AI system defeats world champion Go player Lee Sedol	Google, Stanford University, University of Oxford

Notes: This table, our own elaboration, largely builds on Nilsson [2009] and Wooldridge [2021], two omnibus historical accounts of AI since its inception to present days. Our table is for illustrative purposes only and should not be considered an exhaustive list of university-industry collaborations in AI development.

respondents agree that collaborating with such companies (Google and Microsoft being the most named) is "very" or "somewhat" important in advancing the field. This brings us to the second contribution of our study, which is to examine whether public-private partnerships have contributed to the development of more impactful transformer models.[3] Specifically, we address the following question: Do papers that adopt transformers developed through university-industry collaborations receive, ceteris paribus, more citations and exhibit a higher degree of novelty?

This is a good place to provide a summary of the main findings of this paper. Regarding the diffusion process, as detailed in the initial part of our study (Section 2.3), the results suggest that transformers diffuse at an extremely high speed across virtually every scientific domain: the number of articles (beyond computer science) incorporating transformers has shown an average growth rate of about 400 percent over the past five years. We confirm the central role of some private companies and elite universities in the development of these methods. As for the impact on knowledge creation, our counterfactual analysis presented in Section 2.4 provides evidence

[3]Interestingly, this question has been neglected not only in the AI-focused literature, but also in more general studies exploring the impact of technology and instrumentation on scientific research.

that scientists who adopted transformers experienced increased citation counts, a higher likelihood of publishing novel papers, and produced more disruptive science. Finally, we find evidence that all these gains are particularly more pronounced when the transformers adopted for research are co-developed by university and industry. Our results are robust to a set of alternative econometric specifications, as shown in Section 2-5. The main findings of this paper have significant implications for the organization and management of science, which will be discussed in more detail in the final section of this manuscript.

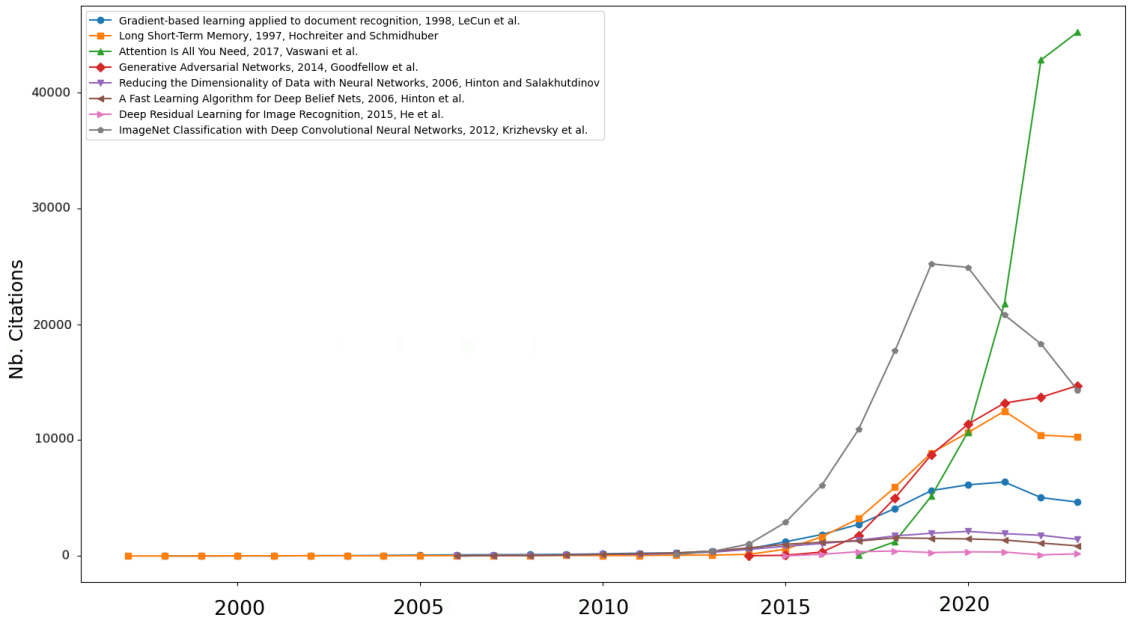
2.2 The Transformer

To understand what a transformer is, we think it is appropriate to take a step back and start with deep neural networks (DNN). Generally, a DNN is an architecture consisting of a multilayer stack of simple modules, most of which are subject to learning, and many of which compute non-linear input-output mappings [LeCun et al., 2015]. So, with multiple non-linear layers (hence the appellation "deep" learning), such architecture can implement extremely intricate functions of its inputs. As Marcus [2018] put it, deep learning is "*[a] perfectly fine way of optimizing a complex system for representing a mapping between inputs and outputs, given a sufficiently large data set*". What is interesting about deep learning is that the system, once trained, is simultaneously sensitive to particular minute details and insensitive to large irrelevant variations in the data. Of course, different DNN architectures are optimized for different tasks.

In the early 2010s, deep learning became popular for computer vision tasks. Much of this popularity was due to important breakthroughs in the so-called convolutional neural networks (CNN) that could, for the first time, surpass human performance in recognizing objects and other vision tasks [Alom et al., 2018]. There was no similar breakthrough in NLP tasks – e.g., translation, text summarizing, or text generation – until 2017. To be clear, recurrent neural networks (RNNs) were commonly used for NLP and to model sequence data; yet they had significant limitations in effectively handling long sequences, making them difficult to train with large and complex datasets. So thus, transformers changed everything.

The transformer model was introduced in 2017 by a team of researchers from Google and the University of Toronto, who published the paper *Attention is All You Need* (∼100,000 citations according to Google Scholar as of January 2024 –

Figure 2.1: Citations received by some of the most influential AI/ML articles



Notes: This graph shows the number of citations received over time by some of the most important AI/ML contributions, those regarded as methodological breakthrough. The transformer paper (green line) is by far the most influential contribution, with more than twice as many citations as the others in the most recent years.

see Figure 2.1). For non-expert readers, a transformer is a type of neural network architecture that presents three main building-blocks, namely: *positional encoding*, *attention mechanism*, and *self-attention mechanism*. Each block has its own purpose. Consider a text corpus for simplicity – although the same logic applies to any sequential data. The first block, the positional encoding, allows the model to recognize the sequential order of words within a sentence. Without this block, the model would not be able to distinguish between permutations of a sentence, such as "I walked my dog" and "My dog walked me", because the word order would be lost. The second block, the attention mechanism, allows the model to focus on the most important information within a sentence, the information actually essential for the attribution of semantic meaning. For example, in the sentence "Today the weather is nice, let's go for a swim", the attention mechanism allows the model to determine that some words such as "weather" and "swim" are central to understanding the context, while others such as "the" or "for" have less semantic weight. Finally, the self-attention mechanism, the key innovation in the transformers, can be seen as a clever trick the

model uses to handle information and discern the relationship between each word in a sentence. The block learns which words requires greater focus (or attention) to better to capture dependencies in the sentence. In the above example, the word "today" carries a crucial meaning, so it needs much attention, because it specifies that the action of going swimming is happening now, not in the past or future. Without getting into technicalities, the attention mechanism determines where to apply attention, while self-attention determines which relationships to capture. We can easily draw parallels with science. Take, for example, drug discovery, a well-known area in which AI/ML has shown great potential. When representing drug molecules, the attention mechanism prioritizes certain chemical features or regions within a molecule that are relevant for its pharmacological activity; the self-attention mechanism then captures potentially long-distance interactions between atoms, especially in larger molecules or proteins [Zhang et al., 2024].

Since the introduction of transformers, there has been a surge in the development of large-scale models, such as OpenAI's GPT (Generative Pre-trained Transformer) series and Google's BERT (Bidirectional Encoder Representations from Transformers). The number of parameters and computing required to train such models opens, according to some estimates, a new era for AI/ML [Sevilla et al., 2022]. Although transformers were initially designed for NLP tasks, their flexible architecture and ability to model complex relationships have made them suitable for various other applications, including applications in science.

An in-depth review of the many applications of transformers in science is beyond the scope of this article and, by necessity, also beyond our knowledge. However, it was enough for us to analyze a few fields of application to understand the versatility and potential of this technology. Transformers have found applications, for instance, in drug discovery and computational biology, where they have been used to predict molecular properties, model protein structures, and understand gene regulation [Vig et al., 2020]. They also have been adapted for a wide variety of time series forecasting tasks, specifically to handle long sequences of data and efficiently predict future values in various domains such as finance, weather forecasting, and energy demand management [Zhou et al., 2021]. They have been employed to improve the performance of some existing reinforcement learning models for robotic control and game playing [Xu et al., 2020, Hu et al., 2022]; to improve the design and production processes of new materials [Rane, 2023]; or to extract highlights from scientific papers [La Quatra and Cagliero, 2022, Taylor et al., 2022].

ChatGPT and its LLM cousins are the tools researchers most often cite when asked for the most useful example of AI tools in science [Noorden and Perkel, 2023]. Most researchers when asked "What do you use generative AI tools (such as ChatGPT and other large language models) for?", common responses include using them for code writing assistance, brainstorming research ideas, drafting research manuscripts, and conducting literature reviews. Similarly, when asked about the greatest benefits brought by AI to the field of science, at the top of the list we find, e.g., helping non-English speaking researchers in paper writing (through editing or translation), making coding easier and faster, and providing summaries of other research to save time during literature reviews.

The rapid adoption of the technology in various scientific domains is certainly the first signal to classify it as a general method of invention – the "general" attribute in the definition. It is therefore time to turn to empirical analysis and study the process of diffusion of transformers in science.

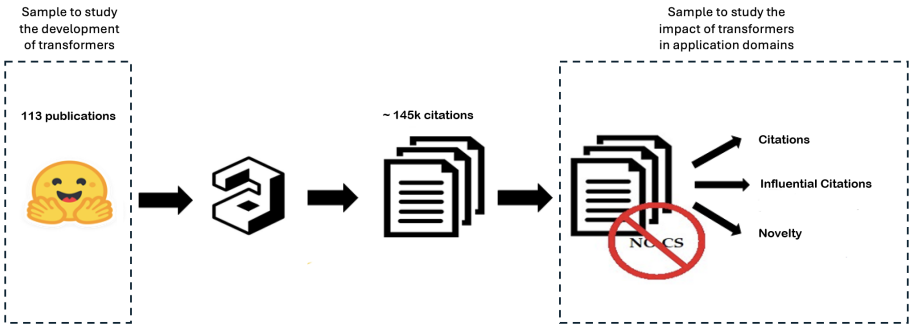
2.3 Data and methods

The sample. Our empirical analysis combines data from two sources: Hugging Face and OpenAlex. The starting point of our research are all the transformer papers available on Hugging Face in January 2023.[4] We initially identified a total of 154 publications. However, after careful manual inspection, we decided to narrow our search by making sure that each publication was actually about the implementation of transformer-based neural networks and not "traditional" deep neural networks. This process resulted in a selection of 113 articles, each proposing a new transformer architecture over the period 2018-2022. These articles are categorized into five groups, reflecting the main field of application for which the technology was developed, namely: Text, Vision, Audio, Multimodal, and Reinforcement Learning (Table 2.6 in Appendix reports the list of selected transformer models). We retrieved further metadata to obtain author information of each publication – i.e., names and affiliations. This dataset will be used to study the development of transformer technology (Figure 2.2, left).

In a second step, we retrieved from OpenAlex all the papers that cited at least one of the 113 articles on transformers just mentioned, during the period 2018-2022. Ope-

[4]Hagging Face – `https://huggingface.co/` – is an open and free platform where the AI/ML community collaborates on models, datasets, and applications. It is accessible to everyone, and thanks to its API, the implementation and use of the proposed models turns out to be very simple.

Figure 2.2: Data pipeline and samples for analysis



Notes: Our starting point was Hugging Face, where we downloaded all publications on transformer technology available in January 2023. Then, using OpenAlex, we collected the bibliographic information of all the papers citing the 113 transformers. After excluding publications classified only as Computer Science, we ended up with about 32,000 publications. Three main metrics, described in detail in Section 2.3, are used to assess scientific impact: number of citations, influential citations, and novelty.

nAlex also provides the *concepts* associated with each publication, that is abstract ideas that works are about and which, consequently, reflect the scientific domain of each publication. Each paper may be associated with multiple concepts, with a score indicating the confidence level for each concept [Priem et al., 2022].[5] Since we are interested in the adoption of transformer technology across different application domains, we followed a common approach employed in previous studies (see, e.g., Cockburn et al., 2018, Bianchini et al., 2022) and focused on publications in all areas other than computer science. This way we ensure – validated also by manual inspection – that we specifically focus on applications of transformers to solve field-specific research problems, and not on technology development. Hence, after excluding those publications that exclusively identified with the level-0 concept of "Computer Science", our final sample consists of about 32,000 publications. This sample will be used to study the impact of transformer technology in science (Figure 3.2, right).[6]

[5]At the time of download, there were 19 root-level concepts (level-0), and six layers of descendants branching out from them, representing a total of around 65,000 concepts. Level-0 concepts include very aggregate scientific areas such as Computer Science, Medicine, Physics, Biology, Chemistry, etc. Artificial intelligence and Machine Learning are two separate level-1 concepts, with Computer Science being their root-level.

[6]We are aware that the boundary between fundamental and applied AI can often be blurred. Indeed, practical applications can reveal new challenges and opportunities, which in turn drive further advances in fundamental AI research. Also, some articles may straddle the line between fundamental and applied AI, incorporating both new methods and practical implementations. However, we believe that OpenAlex's concepts offer a convenient way to organize and analyze AI/ML literature, allowing us to separate papers that propose new models or techniques from those that use (and thus should cite) existing methodologies.

Variables. For each paper in our sample, we built a set of metrics to reflect the impact and novelty of the contribution. Impact is measured by the conventional (weighted) citation count in the 3-year window following the publication year ('*Nb. Citations*'). However, not all citations are equal; some indicate that the cited work is used or extended in the new publication, some may be less important, as they for example discuss the cited work in the context of related literature. Therefore, we also considered the number of influential citations ('*Influential Citations*') as defined by Valenzuela et al. [2015]. Novelty is operationalized using the semantic distance between the focal paper and the prior art that is closest in scientific content ('*Novelty*'). Formally, it is computed as one minus the maximum pairwise cosine similarity between the focal paper and all prior papers published in the preceding 5 years [Arts et al., 2023]. Put it simply, this metric represents how distinct a new paper is from existing literature, with higher values indicating greater novelty and a larger departure from previous knowledge.

We used the additional metadata available in OpenAlex to build a wide range of features. The first set of variables concern the 113 transformer papers, namely: '*Public Involvement (Transformer)*' is binary variable that takes the value 1 when at least one public institution was involved in the development of the transformer architecture, and 0 otherwise; '*Company Share (Transformer)*' represents the share of authors with an industry affiliation; '*Category (Transformer)*' is a categorical variable which represents the main field of application for which the technology was developed; and '*Nb. Fields Application (Transformer)*' indicates the number of fields in which the transformer was applied.

The other variables were built for the $\sim 32,000$ publications citing transformers, namely: '*Company Collaboration*' is a binary variable that takes value 1 when a private company is involved in the research, and 0 otherwise; '*H-Index*' refers to the average H-index of the team authors and it is a measure of their productivity and impact; '*Team Size*' is the size of the research team; '*Academic Age*' refers to the average number of years since the authors' first publication; '*International Collaboration*' is a binary variable equal to 1 when the research results from an international collaboration, and 0 otherwise; '*AI/ML Experience*' represents the fraction of previous AI/ML publications for each author, averaged over the entire team; and, finally, '*Nb. References*' is the number of references cited in the paper.

Table 2.2 shows the basic descriptive statistics of the variables used in this study.

Table 2.2: Summary statistics

Variable	Mean	Std	Min	Median	Max
Nb. Citations	5.74	36.21	0	1	2505
Influential Citations	2.97	35.88	0	0	3771
Novelty	0.03	0.11	0	0	0.74
Public Involvement (Transformer)	0.06	0.18	0	0	1
Company Share (Transformer)	0.73	0.24	0	0.83	1
Company Collaboration	0.11	0.31	0	0	1
H-index	10.09	10.31	0	8	147
Team Size	4.63	2.66	0	4	25
International Collaboration	0.12	0.32	0	0	1
AI/ML Experience	0.49	0.50	0	0	1
Academic Age	4.69	4.50	0	3	11
Nb. Fields Application (Transformer)	17.07	1.82	1	18	18
Nb. References	49.40	37.98	0	43	1579

Notes: This table presents the summary statistics for 31,538 papers citing transformers over the period 2018–2022.

Empirical strategy. We aim to estimate the causal impact of the sudden popularity of transformers on various scientific outcomes, including the quality and novelty of scientific discoveries. We relied on the unexpected discovery of transformers in 2017 as an exogenous shock in the scientific landscape, marked by the publication of Vaswani et al.'s "Attention is All You Need", which revolutionized the field of NLP and beyond, as discussed in Section 2. The superior performance of transformers could not have been anticipated by the AI community, let alone by domain scientists, and triggered adoption across scientific domains since 2017/18, as documented in the next section. This setting provides us with a quasi-experimental design and allows us to draw (as far as possible) causal inference about the observed changes in knowledge creation triggered by the advent of transformer-based technology.

To isolate the effect of the shock, we combined Propensity Score Matching (PSM) with a Difference-in-Differences (DiD) estimator. PSM is important here for controlling selection bias among technology adopters. Indeed, selection bias may occur because scientists who choose to adopt transformers might differ systematically from those who do not. Moreover, matching allowed us to fulfill the so-called parallel trend assumption for the DiD estimation; this assumption requires that, in the absence of the treatment (transformer adoption), the average outcomes for the treatment and control groups would have followed parallel paths over time.

We collected all publications of authors who published in the same year and in the same journal as the authors of the papers using transformers: 27,929 authors and 944,707 publications. We performed PSM based on several pre-treatment characteristics, namely: the number of papers published before 2017, the number of citations received, academic age, H-index, AI experience, and the most frequent level-0 concept. These variables ensured that matched pairs were comparable in terms of productivity, impact, experience, and research focus.

Once the matching procedure was completed, the DiD estimator at the paper-author-time level was obtained using the following regression framework:

$$Y_{jit} = \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{Treatment}_{ji} + \beta_3 (\text{Post}_t \times \text{Treatment}_{ji}) + \lambda_t + \delta_j + \epsilon_{it} \quad (2.1)$$

where Y_{jit} is the outcome variable for paper j by researcher i at time t (i.e., number of citations, novelty score); β_0 is the intercept; '$Post_t$' is a binary variable that equals 1 in the post-treatment period and 0 otherwise; '$Treatment_{ji}$' is a binary variable that equals 1 for the treatment group (adopters of transformers) and 0 for the control group (matched non-adopters); '$Post_t \times Treatment_{ji}$' is the interaction term between the post-treatment period and the treatment group; λ_t and δ_j represent publication year and domain (level-0) fixed effects; and ϵ_{jit} is the stochastic error term. The coefficient of interest is β_3 which captures the impact of transformer.

2.4 Results

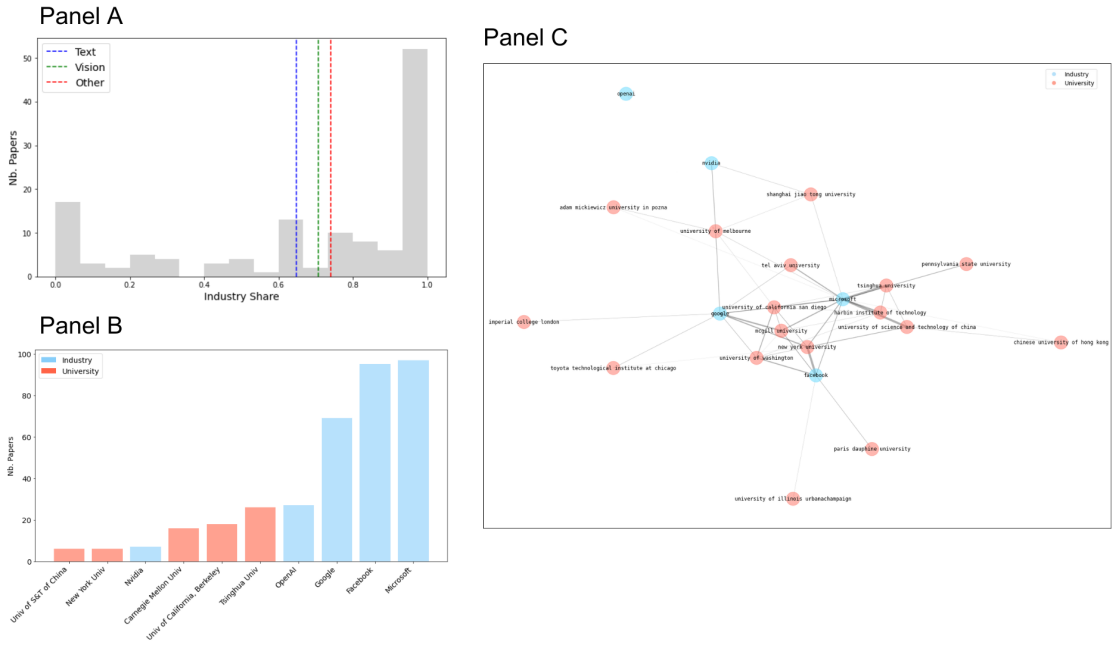
2.4.1 Development and adoption of transformers

Who developed transformer models? Which universities have collaborated with which companies? We focus first on the 113 transformer papers and consider the variable '*Company Share (Transformer)*', which represents the proportion of authors with an industry affiliation.[7]

As shown in Figure 2.3 (panel A), most of the transformer models were the result of university-industry partnerships (63 papers; 48 percent), 51 models (39 percent)

[7]To clarify, suppose there is a paper with 3 authors, 2 from academia and 1 from industry; in this case, the Company Share is 0.33 (1/3). In principle, an author can be affiliated with both university and industry. So, if we consider a paper with 2 authors, one from academia and the other with a double academic/industry affiliation, the Company Share is 0.5. However, in our sample, only 2 out of 115 unique authors have dual university-industry affiliations.

Figure 2.3: Trends in the development of transformer technology

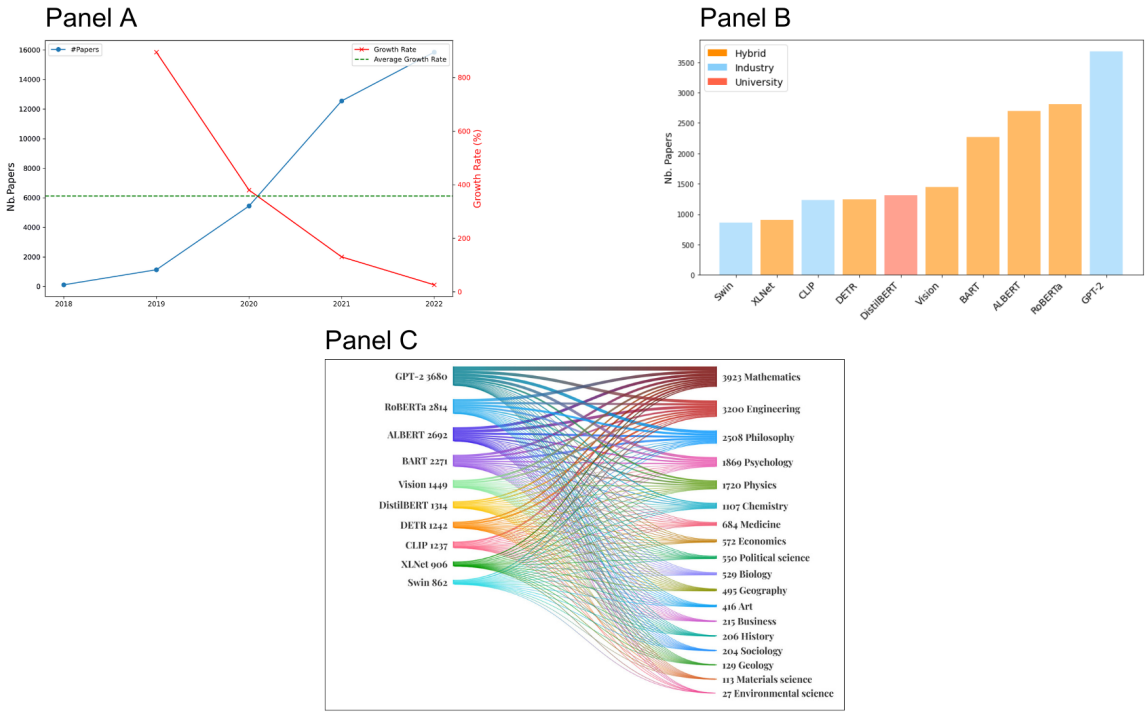


Notes: Panel A reports the number of transformers across different company shares. Panel B shows the leading companies (red) or institutions (blue) that have made significant contributions to this technology. Panel C shows the collaboration network among the top 5 companies involved in transformer development.

were developed by industry alone and 13 (12 percent) by university alone. The mean of company share exceeds 0.60 in all major fields of application for which the original architecture was designed. Figure 2.3 (panel B) displays the top 10 actors, which are defined as the companies and universities with the highest number of authors among the 113 transformer papers. Consistent with our expectations, Microsoft, Facebook, and Google lead the ranking, followed by OpenAI, Tsinghua University and UC Berkeley. As shown in Figure 2.3 (panel C), we identified strong ties in the collaborative network between Microsoft and several Chinese institutions, such as Tsinghua University and the University of Science and Technology of China; Facebook and New York University; and Google and McGill University. Moreover, OpenAI has no collaboration at all.

Which architectures are most adopted by scientists? In which domains do these architectures find applications? The number of papers citing transformers grows steeply from 2018 to 2022, confirming the influence of this technology in many application domains, as discussed in the previous section. The average growth rate during the

Figure 2.4: Trends in the adoption of transformer technology



Notes: Panel A shows the number (blue) and growth rate (red) of papers citing transformers. Panel B shows the most cited transformers broken down by category: university, industry and hybrid. Panel C shows the application domains of the most influential transformers.

period is about 400 percent (Figure 2.4–Panel A).

We classified the 113 transformers into three groups: those developed solely by industry; those developed solely by university; and those co-developed by both university and industry. Figure 2.4–Panel B shows that the most widely used transformer architecture is the Generative Pre-trained Transformer 2 (GPT-2), the large language model released by OpenAI in 2019 and the second in their fundamental series of GPT models. Almost 3,700 papers in our sample, or about 10%, have cited GPT-2. In Figure 2.4, we also see that 6 of the 10 most popular transformers are jointly developed by industry and academia. Table 2.3 provides further details on the most influential transformers. As shown in Figure 2.4–Panel C, the application domains for transformers are highly heterogeneous, encompassing fields from mathematics and engineering to chemistry, medicine, and material science. Transformers have been adopted in virtually all sciences, confirming their versatility and general-purpose nature.

Table 2.3: Most cited transformer-based models

Model	Year	Main innovation	Reference paper
Swin	2021	Introduced hierarchical vision transformer with shifted windows, improving efficiency and scalability in vision tasks.	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
XLNet	2019	Combined Transformer-XL and pre-training to capture longer context.	XLNet: Generalized Autoregressive Pretraining for Language Understanding
CLIP	2021	Unified vision and language representations using contrastive pre-training on a large dataset of images and text.	Learning Transferable Visual Models From Natural Language Supervision
DETR	2020	Proposed an end-to-end object detection model using transformers.	End-to-End Object Detection with Transformers
DistilBERT	2019	Distilled BERT to retain performance while being faster and smaller.	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter
Vision	2020	Applied transformers directly to image patches for image classification tasks.	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
BART	2019	Combined bidirectional and autoregressive transformers for improved sequence-to-sequence tasks.	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension
ALBERT	2019	Increased efficiency with factorized embedding parameterization and cross-layer sharing.	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
RoBERTa	2019	Optimized BERT's approach by modifying key hyperparameters and training methods.	RoBERTa: A Robustly Optimized BERT Pretraining Approach
GPT-2	2019	Extended text generation capabilities with larger model size and complexity.	Language Models are Unsupervised Multitask Learners

Notes: This table provides a brief description of the 10 most influential transformer models.
Source: own elaboration.

2.4.2 The impact of transformers on knowledge production

How does the adoption of transformer technology influence the impact and novelty in science? Table 2.4 presents the results of the Difference-in-Differences (DiD) models, evaluating the impact of transformers on the number of citations (columns 1–2), influential citations (columns 3–4), and novelty (columns 5–6). Overall, the positive and significant coefficients of the interaction terms (see also Eq. 2.1) indicate that, following the advent of transformers, both the impact and novelty of the papers by authors using transformers have increased.

Considering the model specifications that account for field and year fixed-effects, we can draw some conclusions regarding the magnitude of the effects. For instance, consider Model (2): the coefficient of 0.013 indicates an increase in the expected count of citations by a factor of $\exp(0.013) \approx 1.013$ (or 1.3% more citations) compared to the baseline after 2017. In Model (4), the coefficient of 0.024 shows that the effect of the treatment results in an additional 2.4% increase in influential citations. The effect on novelty is milder; indeed, looking at Model (6), we conclude that novelty is marginally increased in the post-transformer period by 0.3%.

Taken together, our findings suggest that using transformers is associated with a significant increase in the number of citations, and this effect is further amplified in the post-transformers period. Furthermore, authors using transformers tend to have more influential citations, and this effect is stronger after 2017. And the adoption of transformers also positively impacts novelty, though there is an overall slight decline in novelty over time.

2.4.3 "Partnerships is all you need"

A different effect depending on the technology used? In order to investigate whether the impact of transformer technology varies depending on whether it was developed by a private company, a public institution, or through hybrid collaboration, we replicated the DiD models separately for each of these three categories. Figure 2.5 shows that the effect of treatment on citation count, influential citations, and novelty is positive and significant for all categories. However, for impact (Panels A and B), the effect is much stronger when scientists used transformers co-developed by university and industry. We do not find the same effect on novelty (Panel C).

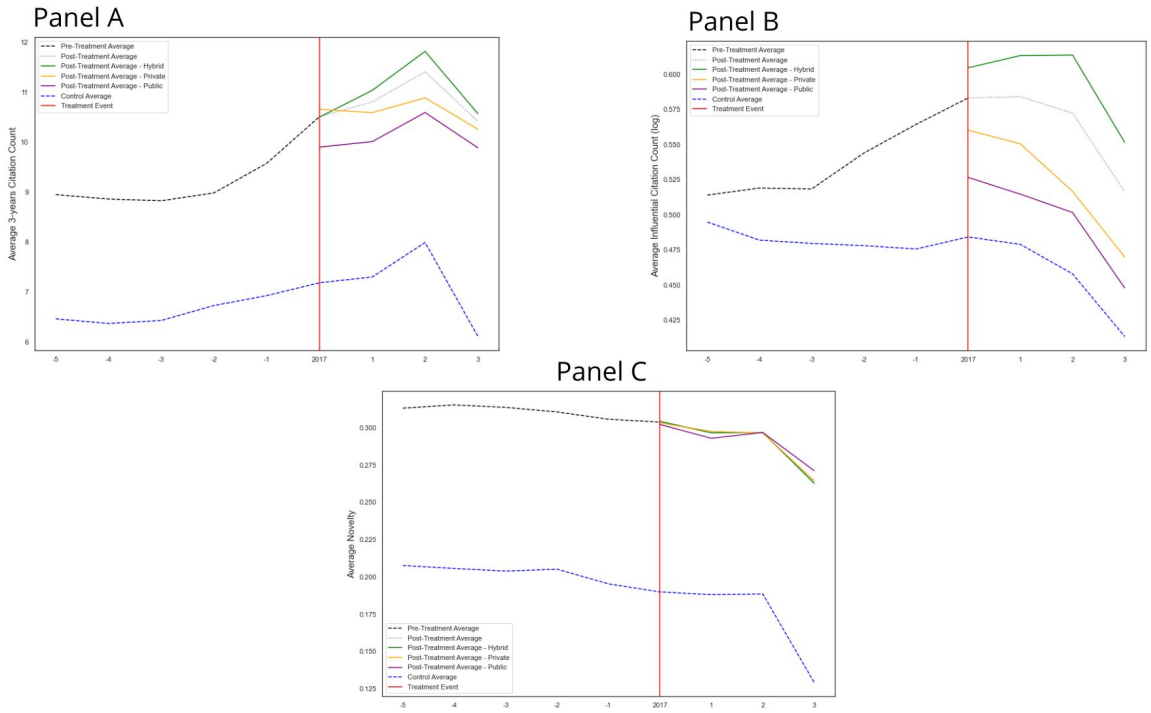
To further corroborate the above findings, we carried out a complementary exercise in a standard econometric setting. Specifically, we restricted the sample to

Table 2.4: DiD estimates

	Nb. Citations		*Influential Citation (log)*		*Novelty*	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.370***	0.370***	0.07***	0.04***	0.113***	0.107***
	(0.003)	(0.003)	(0.001)	(0.004)	(0.0001)	(0.0003)
Post	0.047***	0.014*	0.015***	0.008**	0.003***	−0.017***
	(0.006)	(0.008)	(0.008)	(0.002)	(0.001)	(0.001)
Post × Treatment	0.009***	0.013*	0.024***	0.024***	0.001***	0.003***
	(0.009)	(0.009)	(0.003)	(0.003)	(0.001)	(0.001)
Field		Yes		Yes		Yes
Year		Yes		Yes		Yes
Log Likelihood	−6,503,122	−6,499,635				
AIK	13,006,253	12,999,333				
Adjusted R^2			0.002	0.021	0.065	0.086
F Statistic			1,614.860***	1,743.530***	55,574.600***	7,473.822***
# Observations	2,381,543	2,381,543	2,381,543	2,381,543	2,381,543	2,381,543

Notes: The DiD model for evaluating the treatment effect on the indicators: the number of citations (estimated via negative binomial regression – Columns 1–2), influential citations (estimated via OLS regression – Columns 3–4), and novelty (estimated via OLS regression – Columns 5–6). The asterisks ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Figure 2.5: DiD estimates for different categories of transformers



Notes: These plots represent the impact of transformer technology on the number of citations (Panel A), influential citations (Panel B) and novelty (Panel C). The colours represent the category corresponding to the type of developers.

papers citing transformers (∼32,000 articles, see Figure 3.2) and modeled the main outcome variables as a function of '*Public Involvement (Transformer)*', '*Company Share (Transformer)*', and its square. This last term allows us to explore whether there is a curvilinear relationship of the extent of industry involvement in technology development on the outcomes of interest. Indeed, estimates shown in Table 2.5 suggest that such an effect is taking place: transformers jointly developed by academia and industry increase the scientific impact – evidenced by an inverted U-shape – of researchers using them. This trend is consistent across scientific disciplines (not shown here), implying that the most impactful "methods of invention" are those achieved through joint efforts between academia and industry. Note also that a larger public involvement seems to diminish the impact.

The estimates for the other covariates broadly align with our expectations: collaborations with companies and team experience in AI have positive effects. Other factors such as average H-index, team size, international collaboration, academic age, and number of references are also positively associated with impact measures. Different types of transformer technology yield varied impacts on citations, highly influential citations, and novelty, indicating the importance of technological focus for a paper of reception and impact.

2.5 Discussion

This article contributes to the fast-growing literature on the diffusion and impact of artificial intelligence in science, on the one hand, and the role of the private sector in AI/ML research, on the other. While a growing body of research has focused on assessing the degree of penetration of AI technology in the sciences, fewer attempts have been devoted to empirically investigating how AI can actually impact scientific outcomes. Among these, some scholars have placed emphasis on specific techniques (i.e., deep learning) in specific domains (i.e., health sciences), showing that AI can have a major impact on the number of citations but less so on recombinatorial novelty [Bianchini et al., 2022, Thu et al., 2022]. Another recent study has investigated the impact of AlphaFold on structural biologists' research output, finding no significant results when considering the number of publications and a positive impact on the number of citations [Yu, 2024]. Building on these contributions, we focused on a more recent and pervasive technology – *transformers* – and sought to capture as closely as possible the *causal* implications of its adoption. Given the observed positive impact

Table 2.5: The role of university-industry collaboration in transformer development

	Nb. Citations		*Influential Citation (log)*		*Novelty*	
	(1)	(2)	(3)	(4)	(5)	(6)
Public involvement (Transformer)	−0.150***		−0.088***		-0.0003	
	(0.044)		(0.020)		(0.002)	
Company Share (Transformer)		0.571***		0.190***		−0.014*
		(0.134)		(0.061)		(0.007)
Company Share Square (Transformer)		−0.309***		−0.092*		0.007
		(0.118)		(0.054)		(0.006)
Company Collaboration	0.633***	0.630***	0.366***	0.365***	0.009***	0.009***
	(0.033)	(0.033)	(0.016)	(0.016)	(0.002)	(0.002)
H-index (log)	0.079***	0.079***	0.053***	0.052***	0.019***	0.019***
	(0.008)	(0.008)	(0.004)	(0.004)	(0.0004)	(0.0004)
Team Size (log)	0.704***	0.703***	0.276***	0.276***	−0.002	−0.002
	(0.025)	(0.025)	(0.011)	(0.011)	(0.001)	(0.001)
International Collaboration	0.146***	0.151***	0.028*	0.029*	0.015***	0.015***
	(0.032)	(0.032)	(0.015)	(0.015)	(0.002)	(0.002)
AI/ML Experience	0.249***	0.250***	0.027*	0.027*	-0.0005	-0.006
	(0.031)	(0.031)	(0.014)	(0.011)	(0.006)	(0.006)
Academic Age (log)	0.066***	0.065***	0.015***	0.015***	0.00004	0.00003
	(0.004)	(0.004)	(0.001)	(0.001)	(0.0002)	(0.0002)
Category Audio (Transformer)	0.066	0.036	−0.077***	−0.087***	−0.007**	−0.006*
	(0.063)	(0.064)	(0.028)	(0.029)	(0.003)	(0.003)
Category Multimodal (Transformer)	0.231***	0.218***	0.088***	0.077***	−0.002	−0.002
	(0.045)	(0.045)	(0.021)	(0.021)	(0.002)	(0.002)
Category Reinforcement Learning (Transformer)	0.080	0.245	0.041	0.106	0.001	−0.003
	(0.154)	(0.155)	(0.070)	(0.071)	(0.008)	(0.008)
Category Vision (Transformer)	−0.019	−0.040	−0.026**	−0.036***	−0.001	−0.0003
	(0.028)	(0.028)	(0.013)	(0.013)	(0.001)	(0.001)
Nb. Fields Application (Transformer) (log)	−0.152**	−0.145**	−0.076**	−0.052	0.007*	0.008**
	(0.074)	(0.072)	(0.033)	(0.032)	(0.004)	(0.004)
Nb. References (log)	0.983***	0.984***	0.232***	0.231***	−0.008***	−0.008***
	(0.018)	(0.018)	(0.008)	(0.008)	(0.001)	(0.001)
Log Likelihood	−68,868	−68,846				
AIK	137,806	137,764				
Adjusted R^2			0.158	0.158	0.400	0.400
F Statistic			180.093***	175.500***	637.668***	619.740***
# Observations	31,538	31,538	31,538	31,538	31,538	31,538

Notes: The statistical model for evaluating the use of public, private or hybrid on three indicators: the number of citations received by the publication (Columns 1–2), the influential citation (Column 3–4) and novelty (Columns 5–6) Coefficient estimates of time and topic fixed effects have been omitted from the table. The asterisks ***, ** and * indicate significance at the 1%, 5% and 10% levels, respectively.

on different dimensions of productivity, we believe that policy should prioritize the adoption of AI technologies in different scientific fields; and this means focusing on 'institutional', 'social' and 'individual' factors to facilitate widespread integration (see the discussion in Radhakrishnan and Chattopadhyay [2020], Bianchini et al. [2023c]).

Also, many concerns have been levelled regarding the so-called "privatization of AI research". In this regard, Jurowetzki et al. [2023] has documented a significant transition of premier AI talent to industry roles over two decades, especially to tech giants. This "brain drain", they argue, has a snowball effect: it diminishes the qual-

ity of academic research and limits the scope of novel and risky projects that can be undertaken in the public sector due to the skills gap. Similarly, Ahmed et al. [2023] has provided evidence of a growing divergence in AI knowledge production between non-elite universities and large technology firms, partly attributable to the increasing divide in access to compute. Eastwood [2023] complements these studies by focusing on the political landscape and discussing the regulatory frameworks that could mitigate the risks associated with private sector dominance – see also OECD [2023b], which concludes that "*[g]overnments should support an extensive programme to build knowledge bases essential to AI in science, a need that will not be met by the private sector*" (p.11). While we agree that an equitable distribution of resources and opportunities in AI research is needed, our results suggest that the private sector remains a crucial player in developing AI technologies that can impact science. Yet, public-private partnerships appear to be a *condicio sine qua non* for success. Thus, an obvious implication is that governments should create incentives (e.g., grants, collaborative platforms) to encourage joint projects and resource sharing, while monitoring transparency in AI research and (of course) safeguarding public interests.

Future work could extend the analysis presented here in several ways. First, from a methodological perspective, the impact in different countries could be assessed by considering a national or regional context in order to evaluate how cultural, political, and economic factors influence university-industry dynamics in AI research. Second, further research could investigate the ethical dimensions of algorithms developed in public-private partnerships, assessing whether they differ from those developed independently of the two spheres. Third, AI technology is evolving at such a high pace that the inclusion of the most recent architectures could provide additional robustness to our findings. So far, we can safely conclude that artificial intelligence is pushing the frontiers of science further and further. Yet much work remains to be done to make sure that everyone can benefit from it.

2.6 Appendix

Table 2.6: All transformer model per category

Text	Vision	Multimodal	Audio	Reinforcement learning
ALBERT	BEiT	CLIP	Hubert	Decision
BART	Conditional	CLIPSeg	M-CTC-T	Trajectory
BARThez	ConvNeXT	Data2Vec	SEW	
BARTpho	CvT	FLAVA	SEW-D	
BERT	Deformable	GroupViT	SpeechToTextTransformer	
BERTweet	DeiT	LayoutLMv2	SpeechToTextTransformer2	
BigBird-Pegasus	DETR	LayoutLMv3	UniSpeech	
BigBird-RoBERTa	DiT	LayoutXLM	Wav2Vec2	
Blenderbot	DPT	LXMERT	Wav2Vec2-Conformer	
BlenderbotSmall	ImageGPT	OWL-ViT	Wav2Vec2Phoneme	
BLOOM	LeViT	TrOCR	WavLM	
BORT	MaskFormer	ViLT	Whisper	
ByT5	MobileViT	VisualBERT	XLS-R	
CamemBERT	SegFormer		XLSR-Wav2Vec2	
CANINE	Swin			
ConvBERT	Table			
CPM	VAN			
CTRL	VideoMAE			
DeBERTa	Vision			
DeBERTa-v2	ViTMAE			
DialoGPT	ViTMSN			
DistilBERT	YOLOS			
DPR				
ELECTRA				
EncoderDecoder				
ERNIE				
ESM				
FLAN-T5				
FlauBERT				
FNet				
Funnel				
GPT				
GPT-2				
I-BERT				
Jukebox				
LayoutLM				
LED				
Longformer				
LongT5				
LUKE				
M2M100				
MarianMT				
MarkupLM				
mBART				
mBART-50				
Megatron-BERT				
Megatron-GPT2				
mLUKE				
MobileBERT				
MPNet				
MT5				
Nezha				
Nyströmformer				
OPT				
Pegasus				
PEGASUS-X				
PhoBERT				
PLBart				
ProphetNet				
QDQBert				
RAG				
REALM				
Reformer				
RemBERT				
RoBERTa				
RoCBert				
RoFormer				
Splinter				
SqueezeBERT				
SwitchTransformers				
T5				
T5v1.1				
TAPAS				
Transformer-XL				
XLM				
XLM-ProphetNet				
XLM-RoBERTa				
XLM-RoBERTa-XL				
XLNet				

Table 2.7: Most used transformer per each field

Model Name	Fields of Study	Nb. Publications	Share of Nb. Publications over all fields	Share of Nb. Publications overall publications
GPT-2	Mathematics	787	11.74%	2.23%
GPT-2	Philosophy	579	13.53%	1.64%
GPT-2	Engineering	565	9.80%	1.60%
GPT-2	Psychology	526	18.11%	1.49%
GPT-2	Physics	317	10.67%	0.89%
ESM	Biology	259	22.87%	0.73%
GPT-2	Chemistry	195	9.63%	0.55%
Vision	Medicine	141	13.22%	0.40%
RoBERTa	Economics	122	12.42%	0.34%
RoBERTa	Political science	118	13.02%	0.33%
GPT-2	Art	95	12.97%	0.26%
Vision	Geography	76	9.20%	0.21%
GPT-2	History	58	17.31%	0.16%
GPT-2	Sociology	43	12.79%	0.12%
RoBERTa	Business	41	13.05%	0.11%
Vision	Geology	32	13.79%	0.09%
ALBERT	Materials science	21	10.19%	0.05%
Swin	Environmental science	7	12.96%	0.01%

Table 2.8: Occurrences of top 5 disciplines across different transformer categories

Discipline	Audio	Text	Vision	Multimodal	Reinforcement Learning
Mathematics	188	4714	1336	417	44
Philosophy	168	3478	421	200	10
Engineering	124	3833	1568	212	27
Psychology	95	2409	–	–	20
Physics	77	2081	628	172	11
Chemistry	–	–	–	185	–
Medicine	–	–	383	–	–

Chapter 3

Public sentiments on the fourth industrial revolution: An unsolicited public opinion poll from Twitter

Summary of the chapter

This chapter explores public perceptions on the Fourth Industrial Revolution (4IR) through an analysis of social media discourse across six European countries. Using sentiment analysis and machine learning techniques on a dataset of tweets and media articles, we assess how the public reacts to the integration of technologies such as artificial intelligence, robotics, and blockchain into society. The results highlight a significant polarization of opinions, with a shift from neutral to more definitive stances either embracing or resisting technological impacts. Positive sentiments are often associated with technological enhancements in quality of life and economic opportunities, whereas concerns focus on issues of privacy, data security, and ethical implications. This polarization underscores the need for policymakers to engage proactively with the public to address fears and harness the benefits of 4IR technologies. The findings also advocate for digital literacy and public awareness programs to mitigate misinformation and foster an informed public discourse on future technological integration. This study contributes to the ongoing debate on aligning technological advances with societal values and needs, emphasizing the role of informed public opinion in shaping effective policy.

3.1 Introduction

There is an increasing call for data policy and governance to be aligned with societal values and needs, and worthy of public trust, such that it is necessary to understand people's perception and experience in relation to data and data-driven technologies. This interaction takes different forms, including public discourse on regulatory policies [Douglas, 2012], consumer preferences impacting upon product development [Hekkert et al., 2007], and grassroots movements advocating ethical considerations in technology use [Jasanoff, 2005]. In this chapter, we delve into the Fourth Industrial Revolution (4IR henceforth) [Schwab, 2017], providing a first large-scale study of public opinion on its associated technologies. In particular, we refer to artificial intelligence (AI), robotics, blockchain, cloud computing, the Internet of Things (Iot) and virtual reality, which are reshaping societal processes and systems [Geels, 2002, Orben and Przybylski, 2019]. However, 4IR is not only a technological phenomenon; it is deeply human and societal in nature [Yun and Liu, 2019]. The introduction of these advanced technologies in everyday life can disrupt existing social structures with a corresponding threat in terms of inequalities and the need for new governance models [Rainie and Anderson, 2017]. Society is not a mere passive recipient of these shifts [Sartori and Bocca, 2022], but it actively plays a role in shaping and directing the evolution of technology [Nelson and Sampat, 2001, Ostrom, 2009]. These transformations can be observed in various fields, from labor market dynamics due to automation [Autor, 2015] and to changes in communication patterns as a result of social media [Van Dijck, 2013].

The initial decades after the implementation of such new technological systems have shown a clear difference between the economic and social aspects of technological change [Perez, 2003]. For example, concerns about data privacy have led to significant changes in how personal data is managed and regulated [Zuboff, 2023]. The widespread use of AI in decision-making processes raises ethical concerns on privacy, consent, and accountability of automated systems [Cath, 2018]. Moreover, 4IR shapes social interactions and cultural norms for digital connectivity enhances the boundaries of communities and changes the way people communicate and interact [Holm et al., 2023]. 4IR points out that the introduction and integration of new technologies not only bring about economic transitions, but also significant transformations in social structures and functions [Schwab, 2017]. With its participation in this dynamic process, society influences not only the direction but also the pace of technological advancements [Hughes et al., 1987], with the possibility of hindering

the adoption of certain technologies. For instance, the widespread social demand for sustainable energy solutions has accelerated progress in renewable energy technologies [Jacobsson and Lauber, 2006], while resistance from society can slow down the development of technologies like GMOs as well [Paarlberg, 2000].

The fundamental problem concerning media management derives from a deep cultural rift between the world of science and the world of news and commentary. History has shown that when scientists run to the press with sensational claims that haven't been properly checked, the outcome is very damaging to the credibility of science itself, not to mention the reputations of the scientists involved. Therefore, the role of society goes beyond a simple neat choice of accepting or rejecting technological innovations - it actively shapes its trajectory and impact [Pinch and Bijker, 1984]. Such a mutual relationship between society and technology suggests that understanding technological progress requires a comprehensive approach that not only focuses on the economic and technological dimension but also on social, cultural, and ethical dimensions [Latour, 2007].

The democratization of digital technologies is a first example of the way 4IR has made advanced technologies more accessible. Though these technologies may be enough expensive to be available to a narrow community of institutions and corporations only, the progressive improvements in their components and architecture allowed for a sustained decrease in sale price across time, making them widely available to the majority of the population [Ceruzzi, 2012]. For instance, smartphones, which have advanced computing capabilities, have become widely accessible and have had a significant impact on social dynamics [West, 2012]. The spread of smartphones has largely increased the access to information, allowing people from different socioeconomic backgrounds to join the digital world. The economic theory about the diffusion of innovation contributes to explaining the shift from exclusivity to ubiquity suggesting that technological advancements become more accessible and affordable over time, reaching a wider audience [Rogers et al., 1962]. Furthermore, the rise of social media and digital platforms has created new forms of social engagement and expression but has also introduced challenges related to misinformation and digital well-being [Twenge, 2017]. Nevertheless, we should recognise that the digital divide is still a challenge. While many technologies have become more accessible allowing people from different socioeconomic backgrounds to join the digital world, disparities in access still exist, influenced by factors such as income, geography, and education [Van Dijck, 2013]. In this context, managing the socio-economic

considerations brought about by 4IR is crucial to ensure that its benefits are widely distributed and that potential harms are mitigated. This requires a collaborative approach involving policymakers, industry leaders, and public institutions to develop strategies that promote inclusive growth and safeguard ethical standards [Brynjolfsson and McAfee, 2014, Min et al., 2019].

Social networks have become crucial in shaping public opinion, transforming communication and information dissemination. The extensive use of platforms like Facebook, Twitter, and Instagram has revolutionized how people access and engage with information, creating new dynamics in the formation of public opinion [Allcott and Gentzkow, 2017]. These networks enable rapid information sharing, allowing news and ideas to spread quickly to large and diverse audiences. Consequently, they have become influential tools in political campaigns, social movements, and public discourse [Bakshy et al., 2015]. For example, the rise of hashtag activism and online communities exemplifies how social media can bring attention to societal issues and influence public opinion on a global scale [Jackson et al., 2020].

It is important to underline that the content *algorithm* of these platforms plays a significant role in influencing what users see and engage with. It can potentially raise echo chambers and filter bubbles that reinforce existing beliefs and viewpoints [Pariser, 2011], leading polarisation in public opinion. In such scenarios, users are less likely to be exposed to different perspectives and challenging viewpoints [Sunstein, 2018].

However, these platforms also face challenges such as the spread of misinformation and fake news, which can significantly distort public perceptions and decision-making [Lazer et al., 2018]. The ease with which misleading information can be spread on social networks calls for greater accountability and regulation. This issue is essential to ensure the integrity of public discourse and to prevent the negative consequences of selective exposure. While social networks have democratised the means of influencing public opinion, their impact requires careful consideration and management. Effective strategies are needed to ensure the quality and diversity of public discourse, and to counter the formation of echo chambers and the spread of misinformation [Gillespie, 2018].

Currently, the advent of ChatGPT together with the enormous progress in the field of AI have led researchers to investigate the economic impacts of AI-based technologies [Agrawal et al., 2019b, Furman and Seamans, 2019, Cockburn et al., 2018], and their integration in organizational structures [Brynjolfsson and McAfee, 2014].

The use of AI has led to significant contributions in several disciplines including healthcare, finance, and the like (see Introduction).

However, the way in which AI evolved before the advent of COVID and ChatGPT is still a largely unexplored issue in the literature. Whilst there are several papers that do explore AI dimension in literature [Horowitz, 2016, Awad et al., 2018, Brundage et al., 2020, Merenkov et al., 2021, Kelley et al., 2021, Zhang et al., 2021, Liehner et al., 2023] they deal mostly with surveys and do not consider the potential of social media - e.g., Twitter - to be a key factor in the analysis of public opinion with respect to other technologies. Analysing the influence of the media in shaping public opinion prior to these events can reveal the extent to which media narratives influence public perceptions of this technologies [Maxwell et al., 1972].

Given the complexity of this narrative surrounding the 4IR, two research questions emerge that warrant further investigation. The first question focuses on the evolution in time of public opinions about 4IR. Specifically, *do people's attitudes towards technology become more positive or negative as they are exposed to the advances and implications of the 4IR?* The objective is to quantify and track social sentiment towards the transformative potential of digital technologies. Additionally, this works aims at examining the extent to which the public discourse reflects optimism or concerns on the risks associated with 4IR. The second question *what is the nature of interactions between users with different viewpoints on 4IR?* Thus weexplores whether users with similar viewpoints tend to form polarized communities or engage with open discourse and debate with those holding contrasting beliefs.
This study contributes to the ongoing debate on aligning technological advances with societal values and needs, emphasizing the role of informed public opinion in shaping effective policy. Additionally, it presents a first large-scale study of public opinion on 4IR technologies. Answering these questions contributes to better policy-making in several ways. Firstly, by analysing how public opinions about 4IR technologies evolve over time, policymakers can identify patterns and shifts in sentiment. This insight allows for the anticipation of public concerns and misconceptions before they become widespread. Secondly, during periods of rapid technological change or crisis, for example during the introduction of new technology, understanding public opinion and interaction patterns helps in developing clear and effective communication plans to quickly address and correct any misinformation. Additionally, the recognition of unique misinformation patterns associated with different technology types allows for the implementation of more targeted countermeasures.

3.2 Background literature

3.2.1 Narratives

The narratives surrounding AI and the technologies of the Fourth Industrial Revolution (4IR) have a significant impact on society perceptions and understanding. A number of scholars have conducted in-depth research into the representation of AI in various forms of media, including scientific and popular publications, as well as in fictional contexts. Their findings indicate that this representation tends to oscillate between two extremes: optimism and pessimism. This oscillation is believed to reflect deeply-rooted beliefs, hopes, and fears related to technological advancements [Fast and Horvitz, 2017, Cave and Dihal, 2019, Cave et al., 2020]. In addiction Cave and Dihal [2019] have identified four main narratives that interpret these feelings (Tab.3.1):

- *Immortality-Dehumanization* explores the medical field, in which AI is used in research, suggesting a utopic vision of human immortality, contrasted with dystopic concerns of dehumanization and the loss of human values.
- *Freedom-Obsolescence* in which the former symbolizes the liberation from mundane tasks, promising a future free from physical and mental strain, while the latter is associated with the risks of job losses caused by abrupt technological shifts.
- *Gratification-Alienation* celebrates the potential of AI to fulfill any human desire, offering gratification in the several dimensions of life. However, it is counterbalanced by the risk of alienation, in which technology threatens human interaction.
- *Dominance-Uprising* addresses the role of AI in military applications, oscillating between the need of dominance and security, and the fear about machine uprising and loss of human control.

Table 3.1: AI narratives

Field	Hopes	Fear	Debate
Health	Immortality	Dehumanization	Man conquering immortality while on the other side humans lose their essence, ditching values and emotions.
Employment	Freedom	Job replacement/Obsolescence	Humans will be liberated from tedious or tiring tasks, be they physical or cognitive. The opposite representation is the risk linked to this technical turning point.
Sociology	Gratification	Alienation	AI and robots fulfill every human desire, but on the other hand, the opposite scenario predicts that individuals will only interact with technologies rather than with other people.
Surveillance	Security	Uprising	The optimistic scenario predicts that new tools will enable nations and communities to ensure security for all, while on the other hand there is the iconic narrative of sci-fi where AI will take over humans.

Notes: Own elaboration based on Cave and Dihal [2019]

These narratives not only reflect but also shape social engagement with technology, which as a practice, reveals the dynamics of production and usage [Suchman et al., 2017]. Nevertheless, these narratives often deviate from AI current technical capabilities [Floridi and Chiriatti, 2020, Musa Giuliano, 2020]. This discrepancy is often attributed to the thought capabilities of AI [Neff and Nagy, 2018], which leads to some mismatch in user expectations, e.g., the Tay chatbot incident [Nagy and Neff, 2015, Zemčík, 2021].[1]

Furthermore, human-like perceptions of technology fuelled by the need for social interaction and the push for technological acceptance in robot research [Katz et al., 2015, Salles et al., 2020, Zemčík, 2021], contribute to the construction of these narratives. These factors highlight that technology extends into the social realm through interactions and beliefs.

Alongside, narratives about the 4IR are intertwined with societal progress. Fast and Horvitz [2017] argue that technological advancements under 4IR will fundamentally shape societal evolution, driven by the promise of intelligent machines, improvements in healthcare, and enhanced well-being. Yet, concerns raised by many scholars focus on potential threats, including Orwellian surveillance, job displacement, and

[1]Tay chatbot was launched on Twitter in 2016 as an experiment in *conversational understanding*. However, it was quickly corrupted by users that filled it with racist and offensive remarks, leading the bot to ex inappropriate and inflammatory statements. Therefore, Microsoft shut it down less than 24 hours after its launch.

further ethical challenges [Perkowitz, 2007, Frey and Osborne, 2017, Obozintsev, 2018, Jobin et al., 2019]. Regarding other technologies, such as Virtual Reality (VR), the narratives often focus on the possibilities of *Enhanced Experience* and the risks of *Escapism*. VR provides immersive experiences that enhance learning, entertainment, and social interaction. This process fuels an optimistic perspective beyond physical limitations and enables access to further experiences. Social isolation and escapism may none the less threaten optimistic scenarios and calls in the right balance between virtual and real-world interactions. For what concerns to Blockchain, the literature highlights the contrast between *Decentralization and Trust* and *Complexity and Misuse*. Blockchain technology is emphasised for its ability to decentralize power structures and improve transparency. It enhances trust in transactions without the need of central authorities, as observed in sectors like finance, supply chain, and digital identity. Despite this potential, the complexity of the technology and its association with illegal activities, as well as concerns about energy consumption, presents a counter-narrative [Khan and Salah, 2018]. Seemingly the narrative on the Internet of Things (IoT), turns around *Connectivity and Efficiency* versus *Privacy and Security Risks*. IoT and its interconnected network of devices promise to enhance efficiency and convenience in daily life [Atzori et al., 2010]. However, this increased connectivity also brings significant concerns regarding privacy and data security [Weinberg et al., 2015].

3.2.2 Echo chambers, polarization and misinformation

The advent of the digital era, characterized by the rapid expansion of the internet and social media, has fundamentally altered the landscape of information dissemination and consumption. Despite offering unparalleled access to diverse perspectives, this transformation poses significant challenges, including the creation of echo chambers, the spread of misinformation, and increased polarization. These challenges threaten the integrity of public discourse and the cohesion of social fabric.

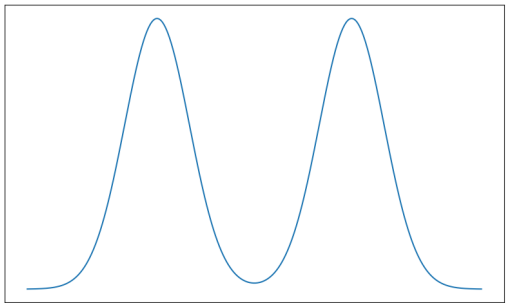
Echo chambers refer to the situation in which individuals are predominantly exposed to opinions and information that reinforce pre-existing beliefs [Del Vicario et al., 2016, Quattrociocchi et al., 2016]. On the one hand, this selective exposure, often exacerbated by algorithmic filtering, facilitates the reinforcement of existing viewpoints. On the other hand, polarisation results from the homogenization of

thought, leading to societal attitudes that increasingly diverge towards the ideological extremes. The presence of echo chambers contributes to a social divide and intensifies both polarization and its deleterious effects on democratic discourse. Furthermore, the circulation of misinformation within these isolated communities can deepen public polarization and distort the collective comprehension of crucial issues [Lazer et al., 2018].

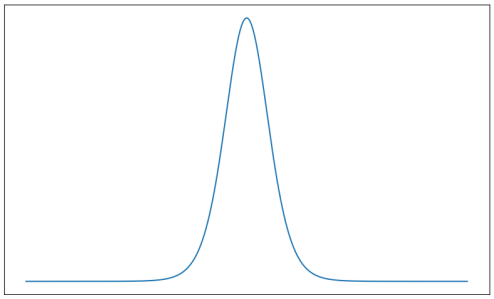
The phenomenon of echo chambers has been identified as a significant contributor to social polarization. These environments are characterized by the amplification of existing beliefs and the minimization of exposure to conflicting viewpoints, which collectively foster a false consensus [Sunstein, 2018]. The critical examination of digital-platforms impact on public opinion and discourse is imperative, given the role of social media algorithms in perpetuating these echo chambers [Nyhan and Reifler, 2010, Pariser, 2011, Lewandowsky et al., 2017].

Misinformation further fuels social polarization by skewing the information landscape and reinforcing pre-existing biases (Fig.3.1). The propagation of false information through social media platforms exacerbates this issue and undermines the integrity of public discourse and the democratic process [Allcott and Gentzkow, 2017]. The swift spread of misinformation within echo chambers not only fixes biased beliefs but also diminishes trust in credible information sources [Bakshy et al., 2015, Wineburg and McGrew, 2017, Lazer et al., 2018, Vosoughi et al., 2018]. The misinformation exposure is a complex interplay of technological, social, psychological, and economic factors that contribute to its proliferation. Social media platforms, with their vast reach and rapid dissemination capabilities, act as catalysts for the spread of false information, driven by algorithms that prioritize engagement over accuracy [Del Vicario et al., 2016, Wu et al., 2019]. Cognitive biases, such as the confirmation bias, play a significant role by leading individuals to favor information that confirms their pre-existing beliefs, thereby intensifying polarization [Ecker et al., 2011, Lewandowsky et al., 2017].

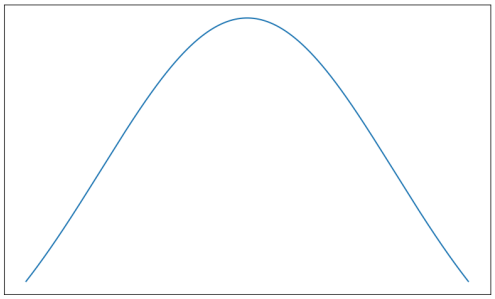
Figure 3.1: Different aspects of public opinion dynamics



(a) Polarization



(b) Consensus



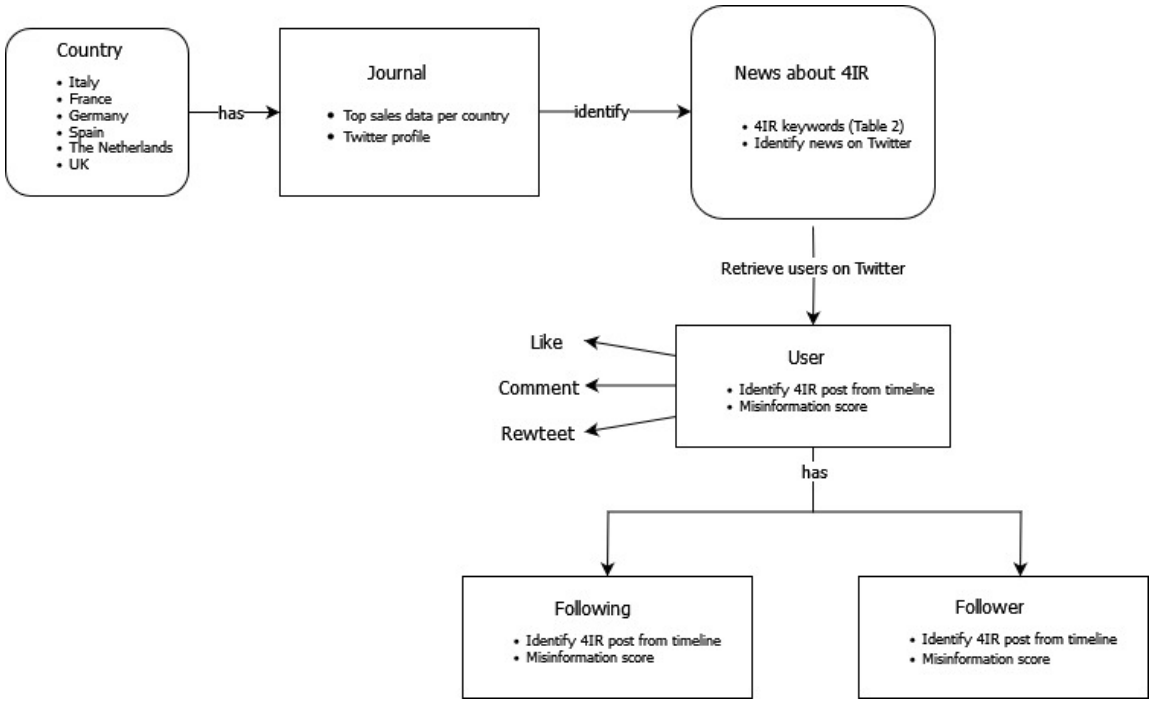
(c) Dissent

Notes: In panel (a) population is divided into two dominant groups with opposing views on a specific issue. The peaks indicate the concentration of individuals within each opinion group, while the trough indicates a lack of moderate stances. This highlights the clear divide and potential for increased social tensions; in panel (b) the "lock-in" effect in public opinion occurs when a single viewpoint has become overwhelmingly predominant. This marginalizes alternative perspectives and demonstrates the societal or cultural homogeneity on a specific issue; in paneò (c) the dissent shows a spectrum of views where the majority holds a central opinion, while a range of dissenting views exists on either side. This indicates a diverse and engaged public discourse

3.3 Data and methods

We focused on the analysis of the discourse surrounding 4IR across six European countries: France, Germany, Italy, Spain, the Netherlands, and United Kingdom. We adopt a multi-step approach to gather an original dataset (Fig.3.2).

Figure 3.2: Data pipeline-time period considered from 01/01/2006 to 31/12/2019



3.3.1 Data sources

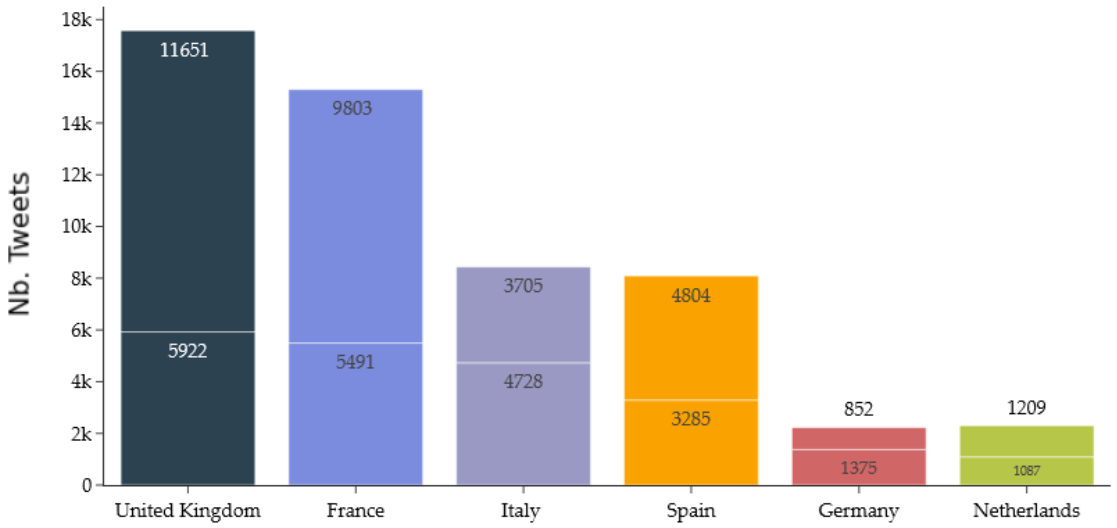
The data collection process begins with the identification of the most widely circulated newspapers in each country according to the number of copies sold (Tab.3.2). Afterwards, we detect their official Twitter profiles and their corresponding tweets which contain some keywords belonging to 4IR (Tab.3.2). To guarantee linguistic precision and cultural appropriateness, we perform translations of keywords. For example, in the case of Italy, we search for both *Artificial Intelligence* and *Intelligenza Artificiale*.

Table 3.2: Country-specific 4IR keywords and associated newspapers

Country	**4IR Keywords**	**Newspaper**
UK	artificial intelligence; robot; blockchain; cloud computing; IoT; virtual reality	DailyMailUK; guardiannews; EveningStandard; thetimes; MetroUK; MailOnline; guardian; TheSun; DailyMirror
France	intelligence artificielle; robot; blockchain; cloud computing; IoT; virtual reality; réalité virtuelle; Internet des objets	humanite_fr; Mediapart; LaCroix; libe; lopinion_fr; le_Parisien; lemondefr; Le_Figaro; LesEchos
Spain	inteligencia artificial; robot; blockchain; cloud computing; IoT; virtual reality; realidad virtual; Internet de las cosas	ElMundoEspan; elcorreo_com; lavozdegalicia; diariovasco; elperiodico; abc_es; larazon_es; el_país; LaVanguardia
Germany	künstliche Intelligenz; robot; blockchain; cloud computing; IoT; virtual reality; virtuelle realität; internet der dinge	Ndaktuell; tazgezwitscher; Tagesspiegel; BILD; SZ; faznet; welt; handelsblatt
Netherlands	kunstmatige intelligentie; robot; blockchain; cloud computing; IoT; virtual reality; virtuele werkelijkheid; internet der dingen	Delimburger; DeGelderlander; trouw; De_Stentor; nrc; Telegraaf; volkskrant; Adnl
Italy	intelligenza artificiale; robot; blockchain; cloud computing; IoT; virtual reality; realtà virtuale; internet delle cose	Ilgiornale; LaVeritaWeb; Avvenire_Nei; Libero_official; LaStampa; fattoquotidiano; repubblica; Corriere; Solo24ore

To create a sample of users who show interest in AI and related technologies, we observe interactions – likes, retweets, and comments – with tweets from the selected newspapers. We consider users who engage with them as if they have a potential

Figure 3.3: Number of news and tweets collected per country

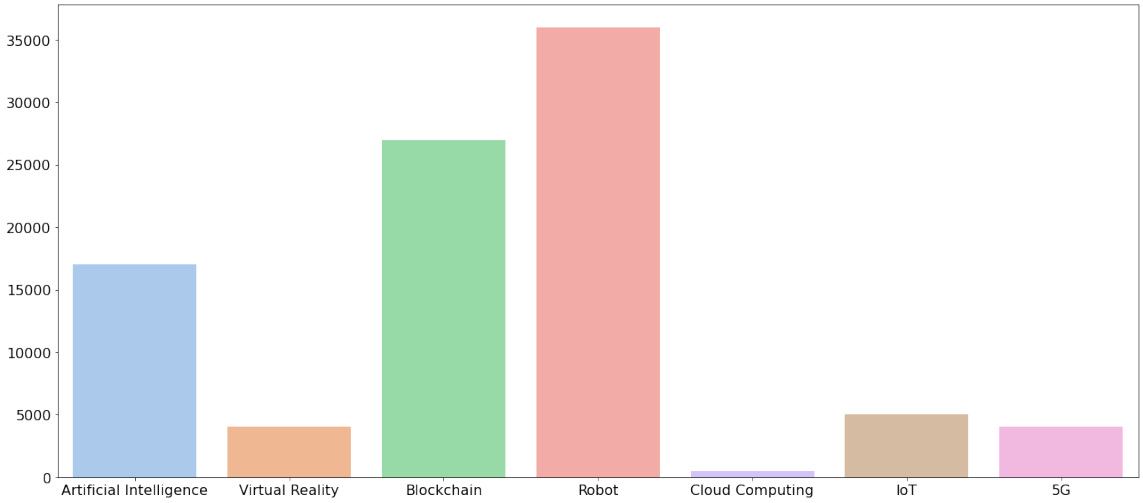


Notes: The lower number in each bar represent the number of news articles collected, while at the top there is the number of tweets by the users.

personal interest in the technologies given their interaction with the newspapers. For each user, we collect their tweet timeline from January 2006 to December 2019 using Twitter API and *twarc*[2]. This time frame was selected to mitigate the potential influence of the COVID-19 pandemic on the data. Furthermore, we collect data on the followers and followed accounts and apply a similar process to gather tweets from their timelines. The final dataset includes approximately 25,000 users and 90,000 tweets (Fig.3.3). Each tweet is identified by a unique ID and includes information about author, text, date, as well as details about the location, number of likes, and retweets.

[2]Twarc is a command line tool and Python library for collecting and archiving Twitter JSON data via the Twitter API. It handles Twitter API's rate limits and can be used to collect tweets, users, trends, and hydrate tweet IDs.

Figure 3.4: Number of tweets retrieved per keyword



Notes: Frequency of specific technology-related keywords. The data reflects the number of mentions for each technology, highlighting the evolving interest in various technological fields among Twitter users.

Building on the data collection framework described above, we then proceeded to analyze the contents of the gathered tweets. As first approach we focused on identifying the occurrence of the key technology-related terms within the tweets. As illustrated in Fig.3.4, the prevalence of discussions on technologies such as *Robots*, *Blockchain*, and AI mirrors the findings of other studies, which highlight the increasing penetration of these technologies in various sectors and their perceived impact [Ford, 2015, Swan, 2015]. The moderate mentions of cloud computing and virtual reality align with the observations by Greenhalgh et al. [2017], who suggest that while these technologies are well-established, they may not provoke the same level of continuous public intrigue as more disruptive technologies. On the other hand, the relatively lower frequency for emerging technologies like IoT and $5G$ towards the later part of the analyzed period can be understood through the lens of diffusion of innovations theory, which posits that newer technologies typically undergo a phase of gradual adoption marked by lesser public discourse initially [Rogers et al., 1962]. This trend underscores the necessity to continuously monitor technological discourse over time to capture shifting public and professional interests as new technologies mature and penetrate different market segments.

Regarding intra-country differences, the results in Tab.3.3 suggests pattern variances in the discussion frequency of technological terms among European countries. AI is more debated in France and Italy compared to other countries such as United

Table 3.3: Share of technology terms by country

Keyword	UK	France	Italy	Spain	Germany	Netherlands
Artificial Intelligence	14.49	27.38	25.25	19.14	13.84	19.94
Virtual Reality	7.15	0.19	2.48	3.46	2.28	11.98
Blockchain	49.11	21.21	15.68	19.32	49.73	24.86
Robot	59.91	34.68	34.98	26.90	22.09	78.76
Cloud Computing	0.92	0.31	0.41	0.28	0.27	0.23
IoT	11.83	3.34	3.81	3.20	4.99	2.90
5G	6.57	3.70	5.40	4.61	5.99	7.06

Kingdom and Germany, suggesting a greater focus or investment in AI technologies in these countries. The Netherlands exhibits increased discussion rates on topics such as VR and robotics, which may indicate stronger industrial applications or governmental support in these fields. Germany and the UK both exhibit a high interest in blockchain, which might reflect a robust engagement with cryptocurrency and blockchain technologies. On the other hand, cloud computing and IoT show low percentages across all countries, indicating that these technologies are still in the early stages of adoption or discussion saturation.

3.3.2 Text analysis

We use sentiment analysis techniques to understand public sentiment. Specifically, we employ the XLM-T Roberta model [Barbieri et al., 2021] which is available on Hugging Face and represents a transformer model trained on a dataset including over 15 millions tweets in 10+ languages.[3]

Sentiment analysis, also known as opinion mining, is a key task in Natural Language Processing (NLP) that involves the identification and categorisation of sentiments expressed in text that relate to specific topics, products, or services. By using language models, it is possible to determine the type of sentiment within the text,

[3]Hugging Face is a platform for machine learning and data science that simplifies building, deploying and training of machine learning models. It is often referred to as the 'GitHub of machine learning' due to its ability to enable developers to share and discover machine learning models. The platform offers infrastructures for deploying and running AI in live applications, along with tools to decrease model training time, resource consumption, and environmental impact of AI development.

which can be classified as positive, negative, or neutral. For instance, a sentence such as "I love you and I like you" expresses a positive sentiment.

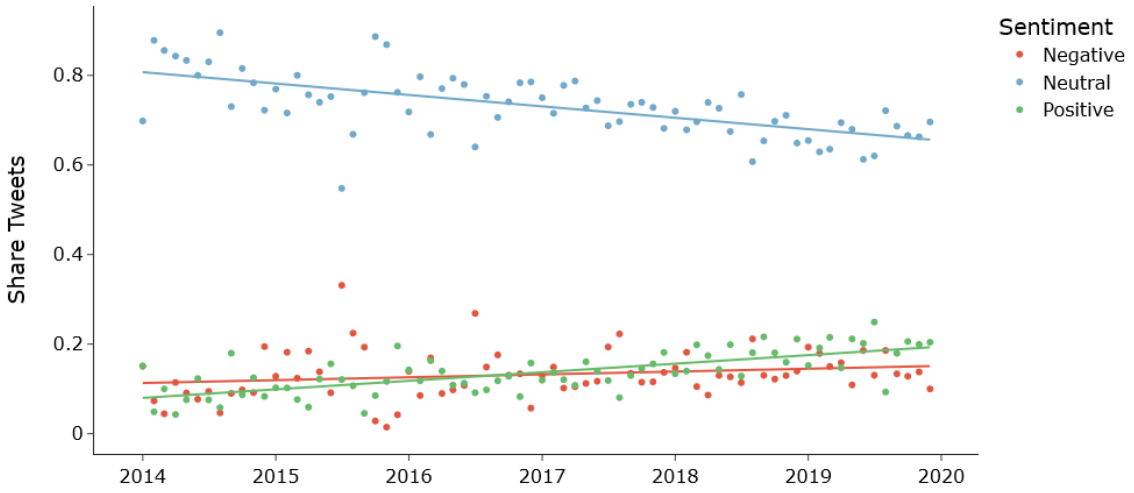
Sentiment analysis is the process of analysing and interpreting express opinions, allowing for the extraction of information from unstructured textual data. It has a wide range of applications across various sectors and functions. In the business sector, it serves as an important mechanism for gathering business intelligence, helping companies capture customer feedback on their offerings for product development, define marketing strategies, and revise customer service operations [Liu et al., 2022]. Sentiment analysis is proving to be crucial to market research, enabling companies to gain a deeper understanding of market trends and consumer preferences by analysing social media posts, reviews and forums, allowing them to tailor services and products to consumer needs and market dynamics, and to manage a brand and its reputation in real time [Pang et al., 2008]. The use of sentiment analysis is not limited to commercial applications. It is also employed in the political sphere to gauge public sentiment towards policies, debates, and election messages. This provides political parties and candidates with valuable insights into their campaign strategy [Tumasjan et al., 2010].

To determine the presence of echo chambers within our dataset, we analyzed the sentiment distribution of tweets shared by users and their followers and followings. The intensity of each sentiment – negative, neutral, and positive – was averaged within the followers and followers for each user. The application of sentiment analysis to detect echo chambers is based on the premise that echo chambers typically exhibit homogeneous sentiment expressions, as members reinforce each other's viewpoints [Garimella et al., 2018]. Using sentiment analysis to examine the tweets of users and their followers/followings allows researchers to detect patterns of agreement or disagreement, which are indicative of the presence or absence of echo chambers. Averaging the sentiment scores among a user's followers and followings to assess consensus and the reinforcement of beliefs is a methodological choice supported by literature on social media dynamics. It helps in understanding the collective sentiment within a user's network, which is crucial for identifying echo chambers where prevalent sentiments can suggest a uniformity in attitudes and beliefs [Sunstein, 2001, Del Vicario et al., 2016]. The assumption here is that high average sentiment scores (either positive or negative) within a network signal agreement and potentially an echo chamber environment. We decided to set as threshold 0.6 to balance significant sentiment indicative of agreement and maintaining robustness across different

topics. The decision to set a threshold of 0.6 for sentiment scores to classify the presence of an echo chamber is a critical step that requires justification. This threshold balance sensitivity (the ability to detect actual echo chambers) and specificity (the ability to exclude non-echo chamber cases). The choice of 0.6 as a threshold implies a significant skew towards a specific sentiment. This approach aligns with the work by Cinelli et al. [2021], who suggest that clear demarcations in sentiment can help identify highly polarized communities, akin to echo chambers. When the sentiment score is above 0.6, it indicates the likelihood of a user as part of an echo chamber, characterized by a high level of agreement and reinforcement of existing beliefs. The use of a sentiment score threshold to infer these characteristics is therefore a rational extension of this definition, aiming to carry out the detection of such environments through quantitative measures of sentiment agreement.

For the purpose of the narrative identification, the DeBERTa algorithm was employed on zero-shot classification, a machine learning technique that enables a model to accurately classify data into categories that were not present during training. DeBERTa was selected over other models due to its disentangled attention mechanism, which distinguishes between the relative positions of words from their absolute positions in a sentence, thereby enhancing its ability to understand complex language patterns. This capability enables the model to apply generalisation effectively from observed categories to unobserved categories through the utilisation of semantic relationships between categories [Lampert et al., 2009]. Moreover, DeBERTa enhances this capability with a robust pre-training on a diverse dataset, which provides it with a broad linguistic understanding necessary for handling the novel and complex sentence structures encountered in differents scenarios. Additionally, the performance of DeBERTa on various NLP benchmarks indicates its capability on feature extraction and contextual understanding capabilities. Typically, rich feature representations, such as embeddings, are employed to capture underlying similarities between different classes. For instance, a model trained on images about animals and their corresponding labels could correctly classify an unobserved image of a “zebra” if it has learned the concept of animals and similar features from observed categories such as “horse”.

Figure 3.5: Trends in newspaper tweet sentiments



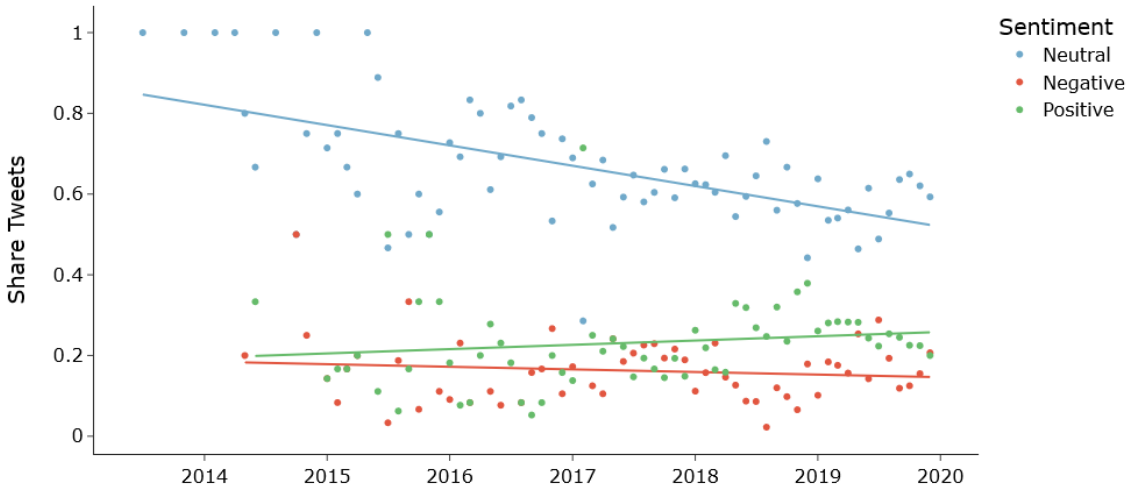
Notes: The figure shows a decrease in neutral sentiment tweets and a gradual increase in both positive and negative sentiment tweets, indicating a slightly growing polarization in public discourse over time.

3.4 Results

The investigation into public-opinion dynamics surrounding 4IR technologies, conducted prior to the widespread adoption of ChatGPT and the onset of the COVID-19 pandemic, revealed a set of insights into societal perceptions and discursive patterns. We analyzed a comprehensive dataset of tweets and news articles from six European countries (Italy, France, Germany, United Kingdom, Netherlands, and Spain), quantifying public sentiment and identifying prevalent themes and narratives that shape societal engagement with 4IR technologies.

We apply the sentiment analysis both on the tweet shared by the newspaper and the user, discovering a similar pattern(Fig.3.5 and 3.6). Over time, there is a tendency towards an increased revelance of negative and positive sentiments with respect to some neutrality. Moreover, we compute the sentiment analysis on the follower and the following for each user.

Figure 3.6: Trends in user tweet sentiments



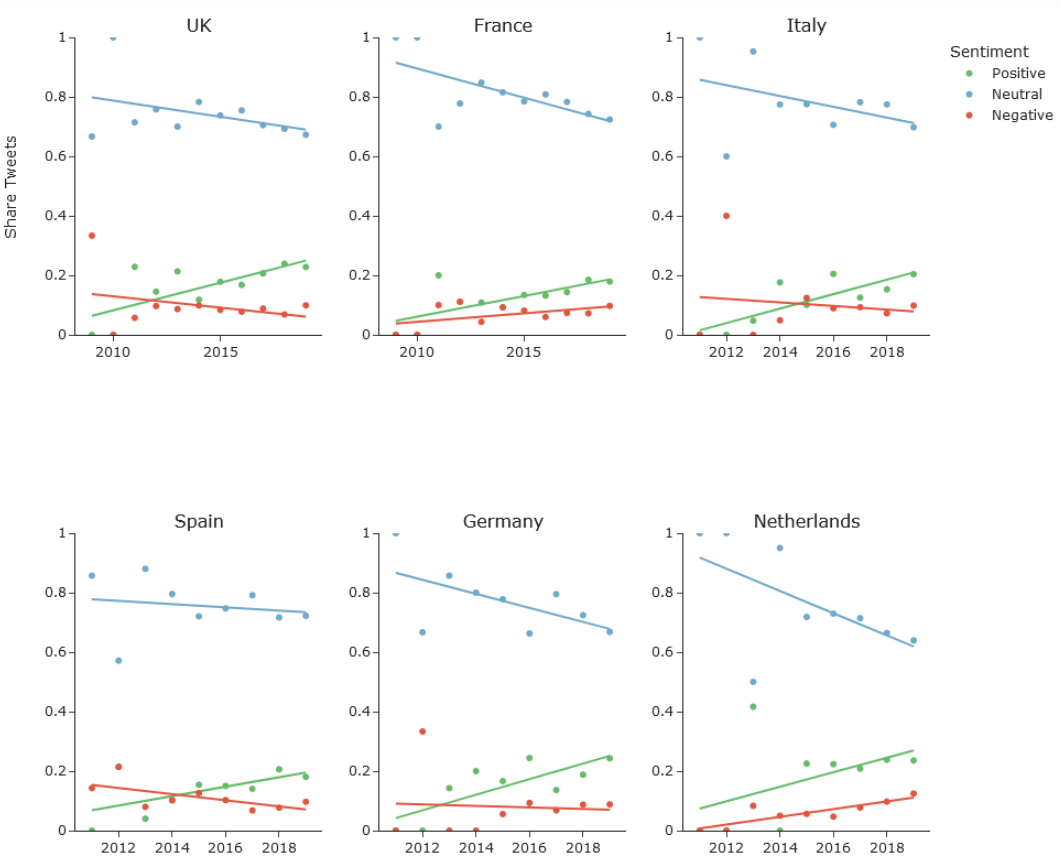
Notes: The figure shows a decrease in neutral sentiment tweets and a gradual increase in both positive and negative sentiment tweets, indicating a slightly polarization in public discourse over time.

The data indicate a decrease in neutrality in both media coverage and user responses. Specifically, the number of neutral tweets and articles has decreased over time, suggesting that more individuals and media outlets are taking a stand as discussions around 4IR intensify. This shift signifies greater public awareness and engagement with emerging technologies, reflecting a more polarized and active debate.

With applying zero-shot classification using DeBERTa [Laurer et al., 2023] we categorise tweets according to key themes partially considering Tab.3.1, such as *employment* (14%), *environment* (20%), *privacy* (3%), *health* (7%), and *other* (56%) partly following the narratives in Tab.3.1.

Furthermore, our analysis revealed a slight slope towards positivity in the discourse about new technologies. The data show an increase in the frequency of positive tweets and articles, reflecting an optimistic outlook on the potential benefits these technologies can bring. This positive trend is particularly evident in the fields of health and employment, where 4IR technologies, such as AI and robotics, are perceived as tools that can improve quality of life and create new job opportunities. We also conducted specific country analysis, which reveals distinct trends in sentiment toward new technologies across various nations, highlighting the complex landscape of public opinion. Figure 3.7 illustrates that while there is an overall decrease in

Figure 3.7: Trends in user tweets sentiments by country



neutral sentiment across several countries like the UK, Germany, and the Netherlands, indicating a possibly cautious or ambivalent attitude toward new technologies, the trends in positive and negative sentiments show more variability. For instance, countries like Spain and the Netherlands exhibit a rising trend in positive sentiment, aligning with a generally optimistic view on the potential of 4IR technologies. On the other hand, the negative sentiment remains relatively low and stable across most countries, suggesting that while enthusiasm varies, there is not a significant rise in skepticism or opposition.

Building on the results of our sentiment analysis, Tab.3.4 further deepens our understanding of the specific issues that dominate discussions about new technologies in different countries. This topic-categorisation shows a diversified interest that varies significantly between regions. For instance, the share of discussion on *employment*, particularly in country such as the Netherlands and France, could suggests a strong interest in how new technologies are reshaping labour markets. This reflects the positive sentiment towards 4IR technologies observed in these countries, indicat-

Table 3.4: Share of topic per country

Topic	UK	France	Italy	Spain	Germany	Netherlands
Employment	10.26	14.80	13.98	13.43	10.35	16.60
Environment	17.44	14.43	13.13	15.15	15.72	14.94
Health	5.67	7.96	7.83	5.09	3.16	7.06
Other	64.90	60.72	63.45	64.50	68.95	59.24
Privacy	1.76	1.56	1.61	1.83	1.82	2.16

ing optimism about the potential for job creation and economic growth. Moreover, in countries like France and the Netherlands, the importance of this issue is higher than in Germany. This may reflect the integration of technology issues into public health discussions, particularly in the context of recent global health challenges. The relatively low engagement in *privacy* could indicate a need for increased awareness and education on privacy issues as technology becomes more pervasive. These findings complement the sentiment trends by revealing not only the emotional tone of discussions, but also the substantive concerns and interests of the public. A comprehensive analysis of the trends by country for both keywords and topics is detailed in the Appendix.

3.4.1 Echo chamber identification

Our findings indicate that slightly more than 6% of users may be situated within an echo chamber, with minimal variation observed across different topics (Tab.3.5). This suggests a moderate level of topic-dependent engagement within echo chambers, with privacy showing the highest propensity and health the lowest.

In addition, the work of Mosleh and Rand [2022] here was used to associate each user with an elite misinformation-exposure score based on the elite misinformation-exposure score that users follow on Twitter. For instance, by following individuals such as Trump, who are known to disseminate false information, a user is likely to receive a high misinformation-exposure score.[4] This score is negatively correlated with the quality of the news disseminated and positively correlated with conservative ideology. Although misinformation levels are generally low, as indicated by the results

[4]see `https://misinfoexpose.com/`

Table 3.5: Topic analysis with sentiment variance and misinformation mean

Topic	Sentiment Variance	Misinformation mean	Share in echo chamber (%)
Employment	0.29	0.25	5.82
Environment	0.30	0.24	7.08
Health	0.26	0.24	5.22
Other	0.26	0.25	6.14
Privacy	0.30	0.27	7.20

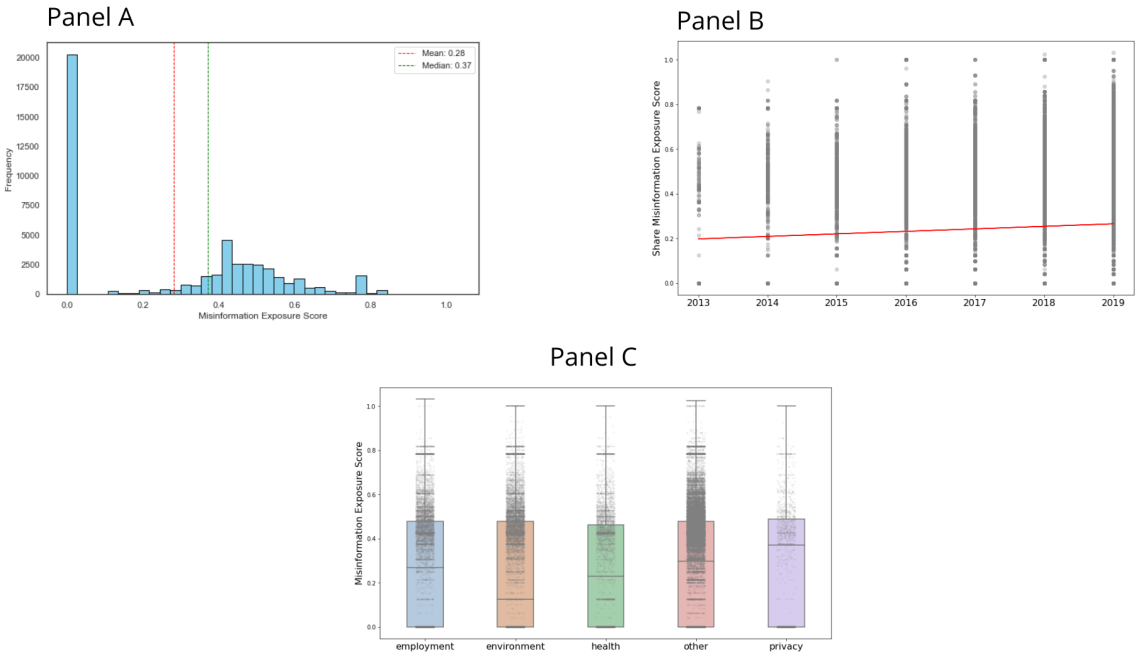
in Tab.3.5 and Figure3.9 (Panel C), privacy has a higher average misinformation score, suggesting that this topic is more affected to misinformation than the others. In Fig.3.9 (Panel B) is illustrate how the gradual increase in the average misinformation exposure score may reflect various factors, including a greater prevalence of fake news on social media and greater polarisation in online discussions. The presence of high scores in each year indicates that misinformation is a persistent problem, but the moderate growth suggests that the dynamics behind its spread may be multiple.

Our analysis highlighted that misinformation and polarization are significant issues in the public discourse on 4IR technologies. The spread of misinformation is facilitated by the presence of echo chambers, where false information can be easily shared and accepted without verification. Polarization is further exacerbated by this dynamic, creating a growing divide between groups with different opinions. The diversity of opinions is another key element that emerged from our analysis. Despite the trend towards increased polarization, there remains a significant variety of viewpoints in public discourse. This plurality of voices supports inclusive and critical debates about the implications of 4IR technologies, highlighting the importance of considering all perspectives in decision-making processes.

At the country-level as shown in Tab.3.6 reveals marked differences in how misinformation and echo chambers influence public opinion across various nations. For instance, in countries with robust digital literacy programs and stringent media regulations, misinformation spread appears to be more contained, and echo chambers less prevalent. This contrasts with countries where digital education is lacking and media regulations are lenient, where misinformation tends to flourish and echo chambers solidify, deepening societal divides.

These disparities not only reflect the effectiveness of national policies but also underscore the varying cultural attitudes towards technology and information con-

Figure 3.8: Misinformation exposure score

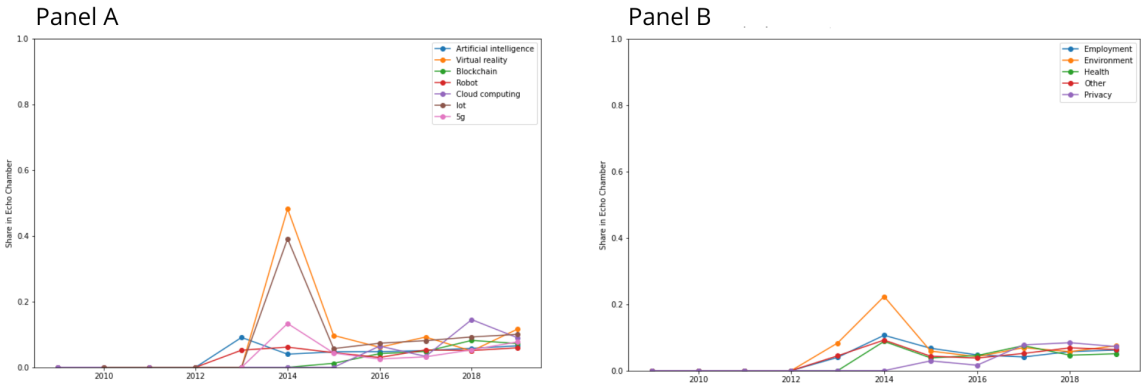


Notes: Panel A represents the histogram of the misinformation exposure score. Panel B the share of misinformation exposure scores. This suggests that exposure to misinformation has gradually risen over this period, highlighting a growing challenge in fighting misinformation in public discourse. Panel C illustrates the distribution of misinformation exposure scores across five topics: employment, environment, health, other, and privacy. The median scores are similar across topics, with a moderate spread in the interquartile ranges. This indicates a consistent exposure to misinformation across these key areas, with no single topic showing significantly higher or lower levels of misinformation exposure

sumption. For example, countries that prioritize education in media literacy and critical thinking skills show a higher resilience to misinformation and a more diverse and healthy public discourse. This is evident in nations like Germany and the Netherlands, where the public debates around 4IR technologies are characterized by a higher degree of skepticism and critical engagement, despite the challenges of polarization and echo chambers.

Furthermore, the degree of technological advancement and the prevalence of technology in everyday life also play crucial roles in shaping the discourse. In technologically advanced countries, there is a tendency for more nuanced discussions about the benefits and risks of 4IR technologies. Conversely, in countries where technology penetration is lower, discussions are often more polarized, with a pronounced divide between pro-technology advocates and those wary of the rapid changes brought about by 4IR technologies.

Figure 3.9: Trends in echo chamber partcipation



Notes: Panel A shows the share of echo chamber participation over time for various technology-related keywords including AI, VR, Blockchain, Robots, Cloud Computing, IoT, and 5G. The sharp peak in 2014 suggests a significant moment of concentrated discussion, possibly linked to pivotal technological developments or debates. Panel B represents the fluctuation in echo chamber shares for discussions on key social issues such as employment, environment, health, privacy, and others over time. The graph highlights a notable spike around 2014, indicating a year of possibly polarized discussions across these topics

3.5 Discussion

Our study highlights a complex set of attitudes towards emerging technologies that policymakers could consider as they shape the future of 4IR technologies regulations. These attitudes reflect the intricate interplay between technological advancements and societal needs and concerns.

The observed reduction in neutrality in both media coverage and public responses indicates a growing polarization in the discourse surrounding 4IR technologies.

The increased polarization can be leveraged to engage more deeply with the public, ensuring that the deployment of 4IR technologies aligns with societal values and needs. For instance, understanding the roots of public skepticism can guide the development of targeted educational campaigns and transparent information sharing. These efforts can foster public trust and support for 4IR initiatives.

The general appreciation for AI when it is used in ways that clearly benefit society, such as improving health and science, underscores the positive externalities associated with technological innovation. This positive view aligns with findings from studies, which noted strong public support for AI applications that enhance societal welfare [Zhang and Dafoe, 2020, Birkstedt et al., 2023]. The recognition of the

Table 3.6: Topic analysis with sentiment variance, misinformation mean, and echo chamber share by country

Country	Sentiment variance	Misinformation mean	Share in echo chamber (%)
		Employment	
UK	0.31	0.30	6.61
France	0.26	0.20	4.23
Italy	0.29	0.26	4.99
Spain	0.28	0.20	5.89
Germany	0.28	0.24	9.30
Netherlands	0.37	0.31	6.50
		Environment	
UK	0.31	0.27	8.20
France	0.27	0.20	4.59
Italy	0.25	0.24	6.60
Spain	0.29	0.16	5.04
Germany	0.33	0.24	9.41
Netherlands	0.35	0.26	8.89
		Health	
UK	0.28	0.28	5.65
France	0.22	0.20	4.52
Italy	0.33	0.22	4.86
Spain	0.25	0.19	4.63
Germany	0.33	0.30	7.61
Netherlands	0.31	0.31	7.06
		Other	
UK	0.28	0.29	7.33
France	0.23	0.21	3.89
Italy	0.23	0.26	6.45
Spain	0.25	0.20	5.14
Germany	0.27	0.27	6.54
Netherlands	0.28	0.24	6.48
		Privacy	
UK	0.32	0.31	7.73
France	0.25	0.21	7.14
Italy	0.35	0.32	3.95
Spain	0.24	0.20	6.19
Germany	0.28	0.30	5.66
Netherlands	0.24	0.26	11.54

AI potential to drive significant improvements in healthcare and scientific research highlights the importance of innovation policies that support and promote beneficial applications. Public support for AI in health and science suggests a broad recognition

of the technology's role in solving complex problems and improving quality of life, which is a key driver of technological adoption and diffusion as articulated in the theory of diffusion of innovations [Rogers et al., 1962].

However, significant concerns arise when AI systems make critical decisions affecting individual lives, such as determining eligibility for welfare benefits. These concerns echo broader issues of accountability and transparency in automated decision-making processes [Butcher and Beridze, 2019]. Public wariness of delegating critical decision-making to automated systems without human oversight reflects the broader economic concern of asymmetric information and the potential for technology to exacerbate inequalities if not properly managed. As AI systems take on more significant roles in governance and administration, the potential for unintended consequences increases, necessitating robust safeguards and accountability measures. The fear of automated decision-making systems potentially mishandling personal data or making biased decisions illustrates the need for transparency and explainability in AI, which are crucial for maintaining public trust [Pasquale, 2015].

Moreover, the strong public call to protect basic rights like privacy highlights the need for regulatory frameworks that safeguard individual freedoms while promoting technological innovation [Brown and Marsden, 2023]. Privacy concerns are paramount in the digital age, where data is a critical resource driving innovation. The economic trade-offs between data utility and privacy must be carefully managed to ensure that advancements in AI do not come at the cost of fundamental rights [Acquisti et al., 2015]. Public demand for stringent privacy protections underscores the importance of developing AI systems that are secure and respect user confidentiality, aligning with the principles of data protection regulations such as GDPR. The nuances in public opinion are evident. People support AI that simplifies tasks and enhances accessibility, recognizing the potential benefits for the greater good. However, they also worry about over-reliance on technology at the expense of human judgment, especially in areas that significantly impact personal and professional lives. This dual sentiment underscores the economic principle of balancing efficiency gains from technology with the maintenance of human-centric values and the potential costs associated with technological disruptions. This concern about the balance between technology and human interaction is echoed in the broader discourse on the social impacts of automation and AI [Brynjolfsson and McAfee, 2014]. Public involvement can take various forms, such as consultations, surveys, and participatory governance models. These approaches help bridge the gap between technological

experts and the public, fostering a collaborative environment where diverse perspectives contribute to more robust and socially acceptable technological solutions.

Regarding regulation, the public desires rules that can address the complex issues AI presents. There is skepticism about leaving AI regulation solely in the hands of the private sector, with a preference for robust oversight to ensure fairness and transparency. This perspective is supported by research indicating that public trust in governance is crucial for the successful implementation of AI technologies [Butcher and Beridze, 2019]. Effective regulation can help mitigate the risks of market failures, such as monopolistic practices and the misuse of AI, ensuring that technological benefits are widely shared. Regulatory frameworks need to be adaptive and forward-looking to keep pace with rapid technological changes, ensuring that they do not stifle innovation while protecting public interests [Birkstedt et al., 2023].

Lastly, there is a strong desire for more public involvement in AI decision-making. People want their voices heard, especially on matters that directly impact their daily lives, supporting the advocacy for participatory approaches in tech policy [Fung, 2006]. This aligns with the economic theory of democratic governance in innovation, which posits that inclusive and participatory policy-making processes can lead to more equitable and effective outcomes [Papadopoulos and Warin, 2007]. By involving the public in decision-making, policymakers can ensure that AI technologies are developed and deployed in ways that reflect societal values and priorities. Participatory governance models help bridge the gap between technological experts and the public, fostering a collaborative environment where diverse perspectives contribute to more robust and socially acceptable technological solutions.

Our findings suggest that data governance policies must align with societal values and needs to earn and maintain public trust. This requires a collaborative approach involving policymakers, industry leaders, and public institutions to develop strategies that promote inclusive growth and uphold ethical standards. There is a strong public demand for robust and independent regulations to address the complex ethical and social issues posed by 4IR technologies. Regulation should not be left entirely to the private sector; public oversight is necessary to ensure fairness and transparency. Educating the public about the potential benefits and risks of emerging technologies is crucial to mitigate concerns and increase acceptance. Raising awareness of the mechanisms of misinformation and echo chambers can help reduce polarization and improve the quality of public discourse. It is important to involve the public in decisions regarding the adoption and regulation of new technologies, ensuring that

their voices are heard, especially on issues directly impacting their daily lives. Continuous monitoring of public discourse and social perceptions over time is essential to adapt policies and strategies in response to changing opinions and concerns. The dynamics of public discourse on social media should be carefully examined to better understand how they influence societal perceptions and behaviors.

Our research focuses on the time period before the worldwide release of ChatGPT in December 2022 [Marr, 2023]. This decision is based on the timing of our data acquisition, which occurred before. The exploitation of the Twitter API also took place prior to Elon Musk's acquisition of the platform in October 2022. This event has significantly altered the conditions of data access and the amount of retrievable information [Conger and Hirsch, 2022]. While we acknowledge the importance of considering the perspectives that reflect the post-ChatGPT era in the analysis of public opinion, this study aims at establishing a baseline framework. Moreover, this benchmark aims at suggesting avenues for further research on the period that follows the introduction of ChatGPT.

3.6 Appendix

Figure 3.10: Sentiment over time per keywords

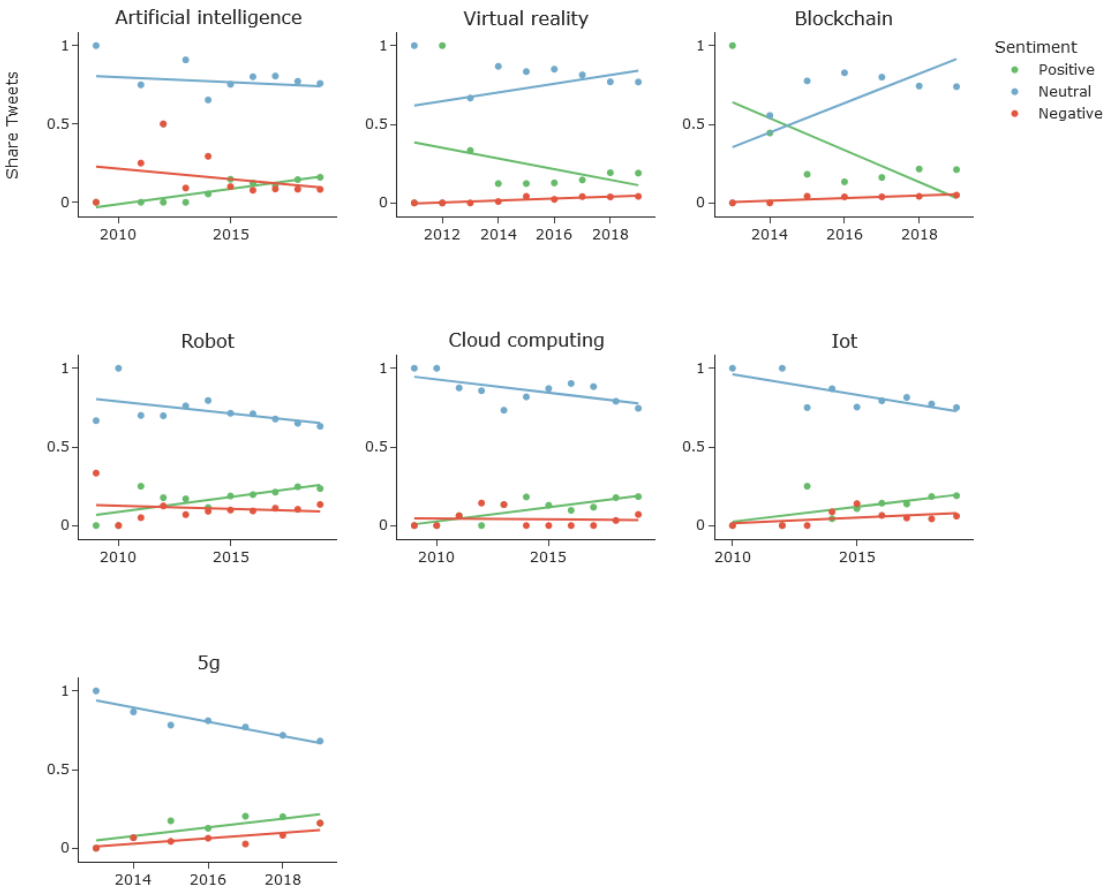


Figure 3.11: Sentiment overtime per topics

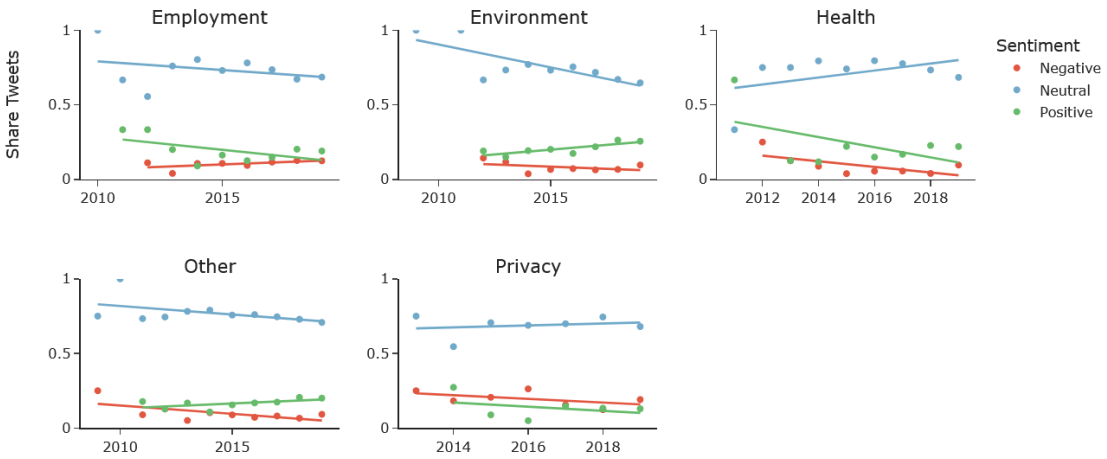


Figure 3.12: Country sentiment overtime AI

Figure 3.13: Country sentiment overtime VR

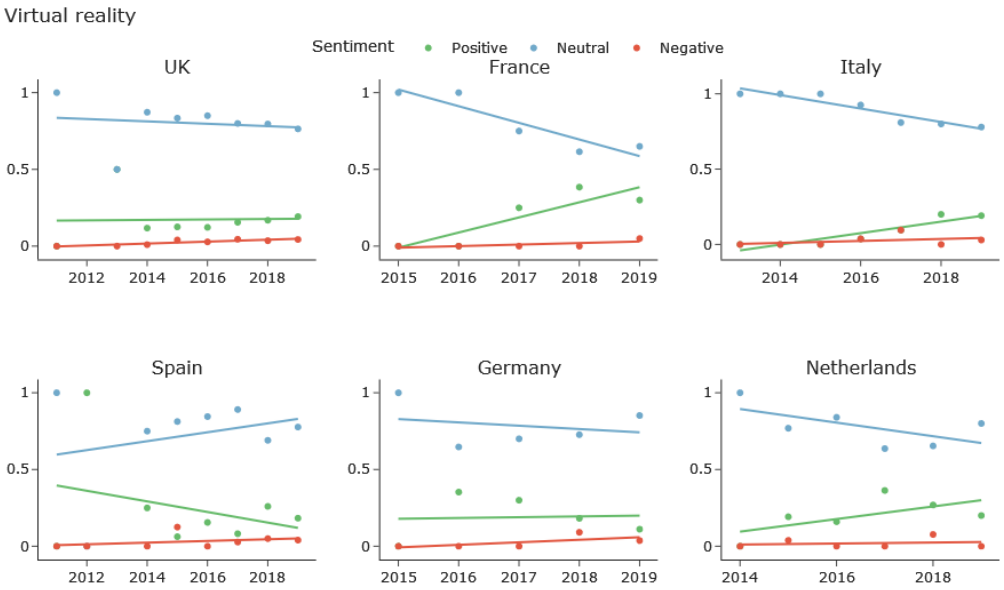


Figure 3.14: Country sentiment overtime Blockchain

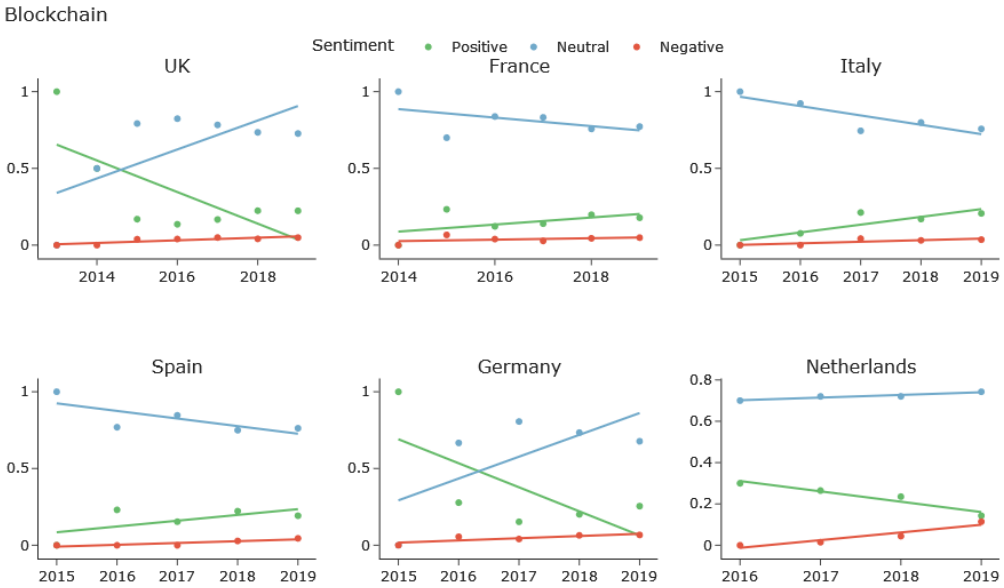


Figure 3.15: Country sentiment overtime Robot

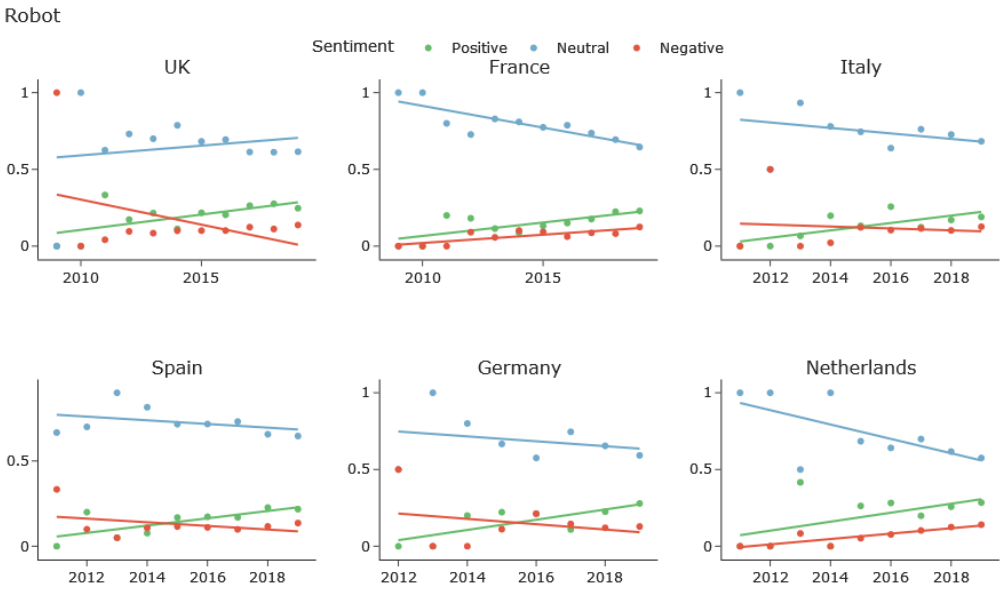


Figure 3.16: Country sentiment overtime Cloud Computing

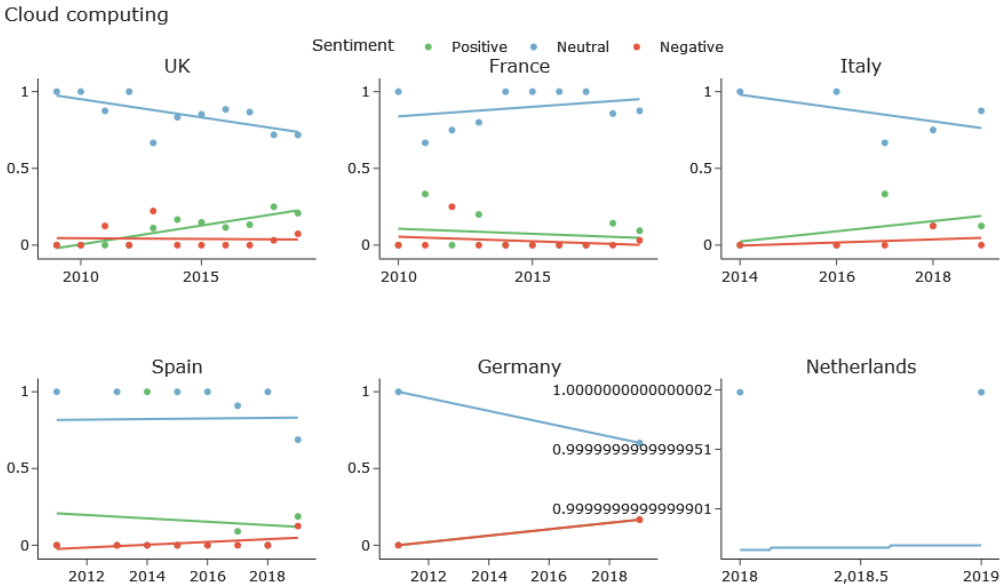


Figure 3.17: Country sentiment overtime IoT

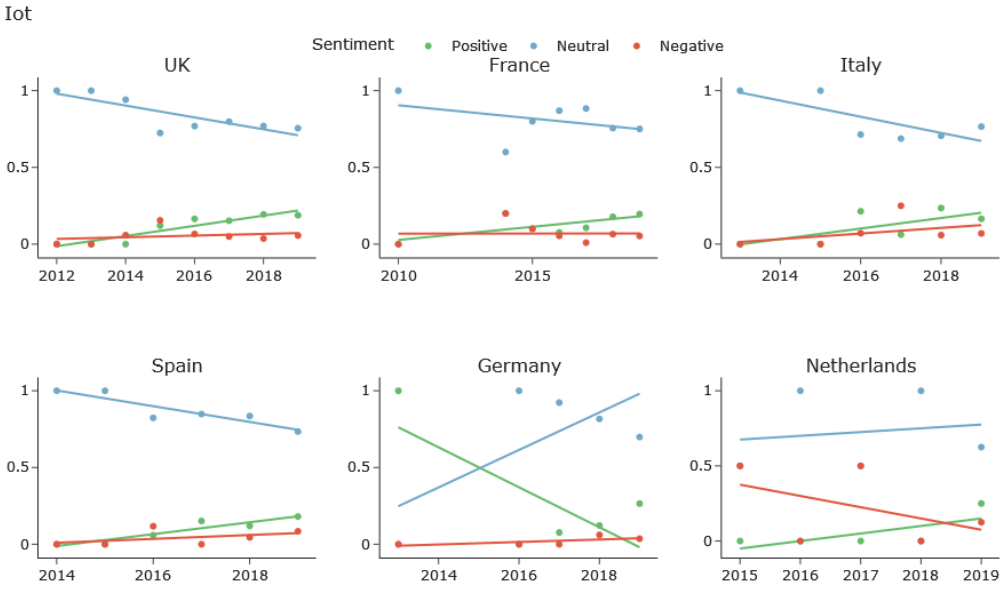


Figure 3.18: Country sentiment overtime 5g

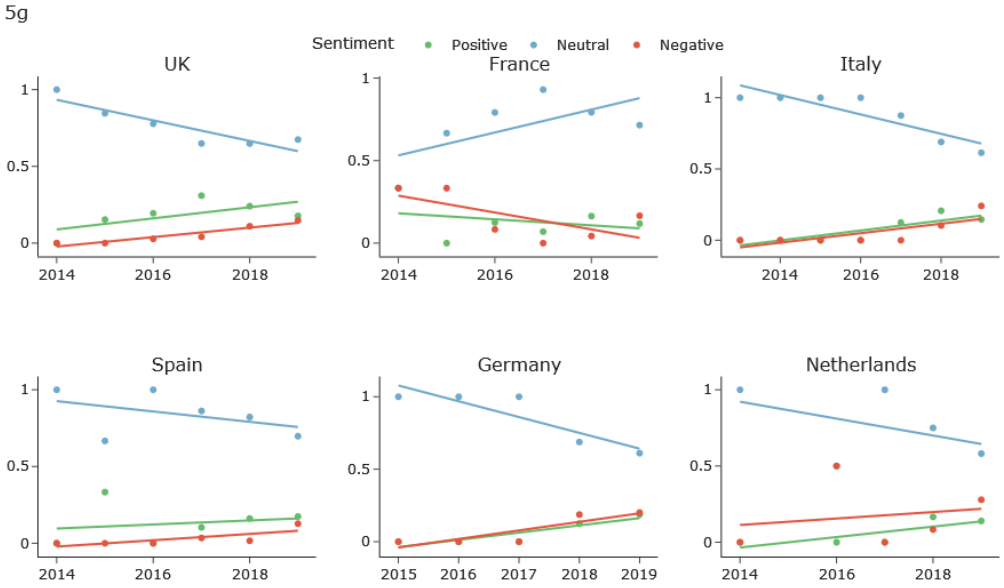


Figure 3.19: Country sentiment overtime employment

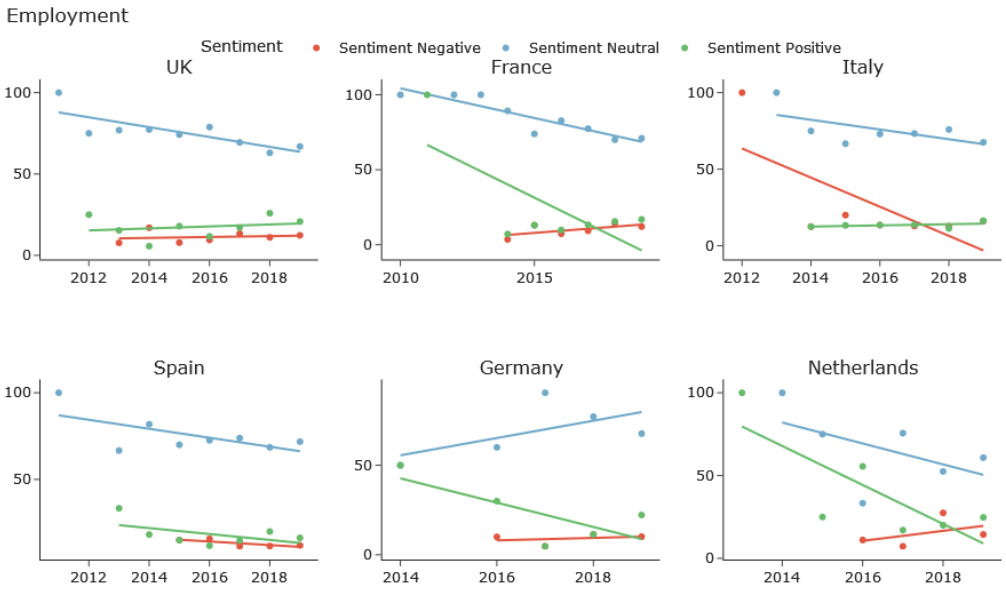


Figure 3.20: Country sentiment overtime Environment

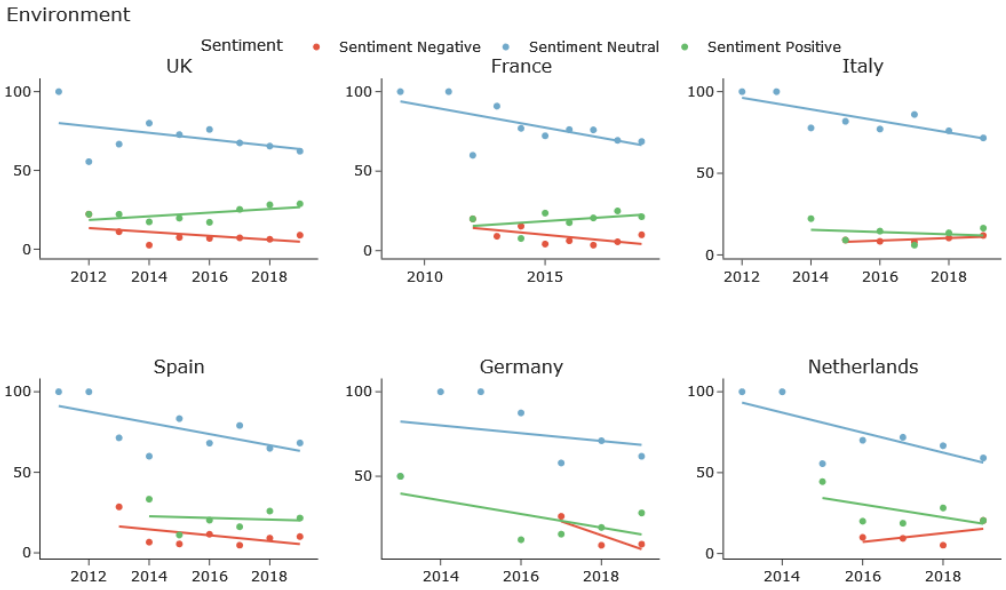


Figure 3.21: Country sentiment overtime Health

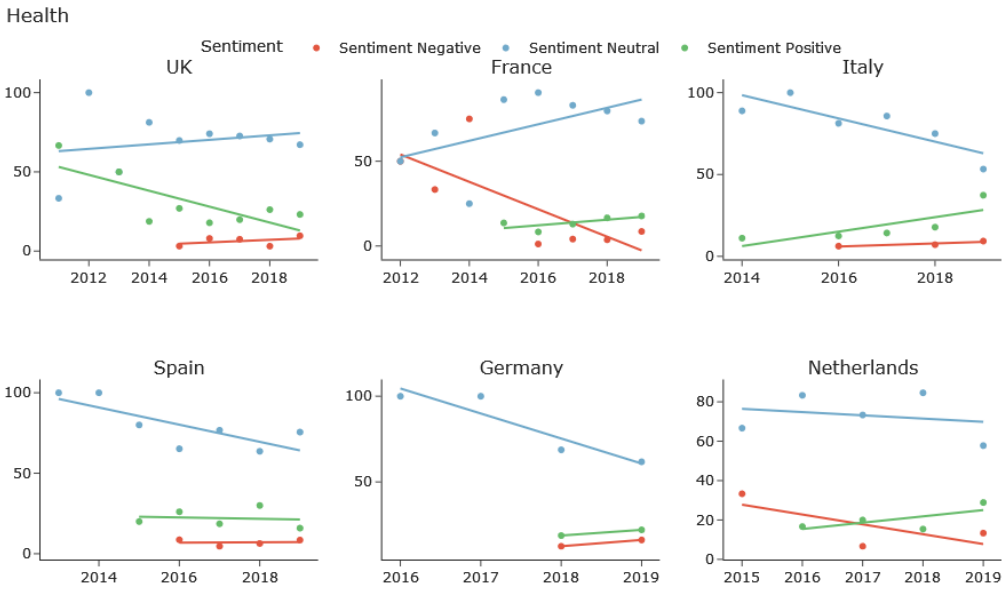
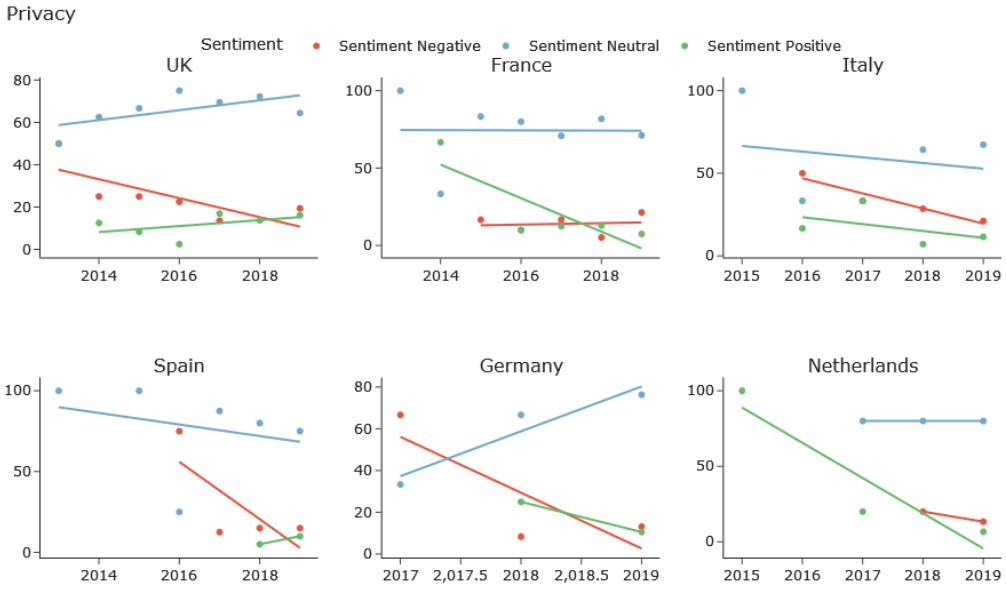


Figure 3.22: Country sentiment overtime Other

Figure 3.23: Country sentiment overtime Privacy



General conclusion

In this thesis we have explored the multifaceted role of Artificial Intelligence (AI) in addressing societal challenges. By examining the intersections of AI with scientific research, industry developments and public perceptions, this thesis highlights the profound impact that AI technology have on society. Each chapter of the thesis explores distinct but interrelated areas where AI is acting not only as a technological innovator, but also as a transformative force within society.

In chapter one, we explored interdisciplinarity, particularly highlighted during the COVID-19 pandemic, where epidemiologists and medical researchers sought external collaborations, notably with AI. Our study examines factors for successful AI-COVID-19 research collaborations, using data from CORD-19, Semantic Scholar, and Altmetric, and evaluating interdisciplinarity through team and epistemological diversity. Our findings reveal that both diversities positively impact citation counts, media attention, and interdisciplinary outreach. However, involvement of AI experts negatively impacts research influence, indicating collaboration challenges. Epistemological diversity is more significant for impact beyond academia. This chapter aims to enhance understanding of computational technologies in addressing societal challenges.

In the second chapter, the focus shifted to the relationship between academia and industry in the development and application of AI, particularly through the lens of transformative technologies such as the transformer models in deep learning. The exploration of these partnerships highlighted their central role in driving scientific and technological progress. Moreover, the chapter argued for the necessity of these collaborations to achieve breakthroughs that could not be achieved by either sector alone.

The third chapter presented an analysis of the societal dimensions of AI, recognising the public as active participants in the technological discourse. Through sentiment analysis of multi-country datasets, this chapter illustrated the polarised public

perception of AI and other 4IR technologies. It showed that public engagement with AI is deeply intertwined with cultural and economic contexts that shape attitudes towards technology.The chapter highlighted the need for informed public discourse and robust digital literacy programmes to navigate the complexities of integrating AI into everyday life.

Throughout the thesis, AI was seen as a potential General Purpose Technology (GPT), similar to previous innovations such as electricity and the internet, which have transformed all sectors of the economy and society. The discussion highlighted the role of AI in driving not only technological, but also economic and social change. The dual potential of AI to both disrupt and enhance different aspects of human life calls for a proactive approach to governance and policy-making in order to harness the benefits of AI while mitigating the risks associated with it.

This work lays the foundation for understanding the expansive role of AI within the current techno-economic paradigm, but it also opens up several avenues for future research.

A key area of future development in chapter one is the expansion of AI applications to address broader global challenges, particularly those related Sustainable Development Goals (SDGs). For instance, the potential of AI to improve resource efficiency, to predict environmental changes and to optimise the use of energy makes it a powerful tool in the fight against climate change. Future studies could explore specific AI applications in these areas, and assess their effectiveness and scalability in different regions and contexts.

In addition, qualitative analysis through interviews with members of interdisciplinary teams could provide deeper insights into the synergies that contribute (or not) to the success of AI-driven projects. Understanding these dynamics can help to design better collaborative frameworks that effectively leverage diverse expertise, thereby increasing the overall impact of AI on global challenges.

In the field of transformer technology, as explored in chapter two, the rapid evolution of these models provides an opportunity for ongoing analysis. Future research could include an in-depth analysis of the most relevant large language models (LLMs) due to the swift pace at which the field advances. Periodic assessments of newer transformer models should also be conducted to verify that the observed impacts on scientific output and innovation are consistent over time. This ongoing evaluation will help to track the progress of these technologies and adjust academic and industrial strategies to take advantage of the most effective AI tools as they

evolve.

The exploration of societal impacts undertaken in chapter three can be extended by focusing specifically on the influence of LLMs such as ChatGPT. Given their significant impact compared to other technologies, a dedicated study of LLMs could provide a deeper understanding of their role in shaping public discourse, decision-making, and ethical considerations in the use of AI. Such research is necessary to guide the development of regulatory frameworks and ethical guidelines tailored to the unique challenges posed by these advanced AI systems.

Finally, as AI technologies continue to permeate different sectors, the integration of these systems with human-centred approaches will be essential. This involves not only technological integration, but also aligning AI developments with human values and ethical standards. Future research should focus on creating adaptive AI systems that work in harmony with human needs and societal norms, ensuring that technology enhances rather than detracts from human well-being.

As AI technologies continue to evolve and permeate various aspects of human life, it is imperative to maintain a vigilant and proactive approach. This means not only advancing technological capabilities, but also fostering an inclusive discourse that incorporates ethical considerations and promotes equitable access to technology. We can ensure that AI serves as a true catalyst for positive societal change by bridging the gap between innovation and sustainability.

Conclusion générale

Dans cette thèse, nous avons exploré le rôle multifacette de l'intelligence artificielle (IA) dans la résolution des défis sociétaux. En examinant les intersections de l'IA avec la recherche scientifique, les développements industriels et les perceptions publiques, cette thèse met en lumière l'impact profond que la technologie IA a sur la société. Chaque chapitre de la thèse explore des domaines distincts mais interdépendants où l'IA agit non seulement en tant qu'innovateur technologique, mais aussi en tant que force transformatrice au sein de la société. Au premier chapitre, nous avons exploré l'interdisciplinarité, particulièrement mise en évidence pendant la pandémie de COVID-19, où les épidémiologistes et les chercheurs médicaux ont recherché des collaborations externes, notamment avec l'IA. Notre étude examine les facteurs de réussite des collaborations entre l'IA et la recherche sur la COVID-19, en utilisant des données de CORD-19, Semantic Scholar et Altmetric, et en évaluant l'interdisciplinarité à travers la diversité des équipes et épistémologique. Nos résultats révèlent que ces deux formes de diversité ont un impact positif sur le nombre de citations, l'attention des médias et le rayonnement interdisciplinaire. Cependant, la participation d'experts en IA a un impact négatif sur l'influence de la recherche, indiquant des défis de collaboration. La diversité épistémologique est plus significative pour l'impact au-delà du milieu académique. Ce chapitre vise à améliorer la compréhension des technologies computationnelles dans l'adressage des défis sociétaux.

Dans le deuxième chapitre, l'accent a été mis sur la relation entre le monde académique et l'industrie dans le développement et l'application de l'IA, particulièrement à travers le prisme des technologies transformatrices telles que les modèles de transformateurs en apprentissage profond. L'exploration de ces partenariats a mis en évidence leur rôle central dans la promotion du progrès scientifique et technologique. De plus, le chapitre a plaidé pour la nécessité de ces collaborations afin de réaliser des percées qui ne pourraient être atteintes par aucun des deux secteurs

seuls Le troisième chapitre a présenté une analyse des dimensions sociétales de l'IA, reconnaissant le public comme participant actif dans le discours technologique. À travers l'analyse des sentiments de jeux de données multi-pays, ce chapitre a illustré la perception publique polarisée de l'IA et d'autres technologies de la quatrième révolution industrielle (4IR). Il a montré que l'engagement du public avec l'IA est profondément entrelacé avec les contextes culturels et économiques qui façonnent les attitudes envers la technologie. Le chapitre a souligné la nécessité d'un discours public éclairé et de programmes robustes de littératie numérique pour naviguer dans les complexités de l'intégration de l'IA dans la vie quotidienne.

Tout au long de la thèse, l'IA a été considérée comme une potentielle General Purpose Technology (GPT), similaire à des innovations antérieures telles que l'électricité et l'internet, qui ont transformé tous les secteurs de l'économie et de la société. La discussion a mis en évidence le rôle de l'IA non seulement dans la propulsion des changements technologiques, mais aussi économiques et sociaux. Le double potentiel de l'IA à la fois de perturber et d'améliorer différents aspects de la vie humaine appelle à une approche proactive en matière de gouvernance et de formulation de politiques afin de tirer parti des avantages de l'IA tout en atténuant les risques associés.

Ce travail pose les bases pour comprendre le rôle expansif de l'IA dans le paradigme techno-économique actuel, mais ouvre également plusieurs pistes pour des recherches futures. Un domaine clé de développement futur présenté dans le Chapitre 1 est l'expansion des applications de l'IA pour répondre à des défis mondiaux plus larges, notamment ceux liés aux Objectifs de Développement Durable (ODD). Par exemple, le potentiel de l'IA pour améliorer l'efficacité des ressources, prévoir les changements environnementaux et optimiser l'utilisation de l'énergie en fait un outil puissant dans la lutte contre le changement climatique. Les études futures pourraient explorer des applications spécifiques de l'IA dans ces domaines et évaluer leur efficacité et leur scalabilité dans différentes régions et contextes. De plus, une analyse qualitative par le biais d'entretiens avec des membres d'équipes interdisciplinaires pourrait fournir des perspectives plus approfondies sur les synergies qui contribuent (ou non) au succès des projets pilotés par l'IA. Comprendre ces dynamiques peut aider à concevoir de meilleurs cadres collaboratifs qui tirent efficacement parti des expertises diverses, augmentant ainsi l'impact global de l'IA sur les défis mondiaux. Dans le domaine de la technologie des transformateurs, comme exploré dans le Chapitre 2, l'évolution rapide de ces modèles offre une opportunité pour une anal-

yse continue. Les recherches futures pourraient inclure une analyse approfondie des modèles LLMs les plus pertinents en raison de la rapidité avec laquelle le domaine progresse. Des évaluations périodiques des nouveaux modèles de transformateurs devraient également être menées pour vérifier que les impacts observés sur la production scientifique et l'innovation sont constants dans le temps. Cette évaluation continue aidera à suivre les progrès de ces technologies et à ajuster les stratégies académiques et industrielles pour tirer parti des outils IA les plus efficaces à mesure qu'ils évoluent. Enfin, alors que les technologies d'IA continuent de s'implanter dans différents secteurs, l'intégration de ces systèmes avec des approches centrées sur l'humain sera essentielle. Cela implique non seulement une intégration technologique, mais aussi l'alignement des développements de l'IA avec les valeurs humaines et les normes éthiques. Les recherches futures devraient se concentrer sur la création de systèmes d'IA adaptatifs qui fonctionnent en harmonie avec les besoins humains et les normes sociétales, en s'assurant que la technologie améliore plutôt qu'elle ne diminue le bien-être humain. Alors que les technologies d'IA continuent d'évoluer et de s'infuser dans divers aspects de la vie humaine, il est impératif de maintenir une approche vigilante et proactive. Cela signifie non seulement faire progresser les capacités technologiques, mais aussi favoriser un discours inclusif qui intègre les considérations éthiques et promeut un accès équitable à la technologie. Nous pouvons garantir que l'IA agit comme un véritable catalyseur pour un changement sociétal positif en comblant le fossé entre innovation et durabilité.

Bibliography

D. Acemoglu and P. Restrepo. Artificial intelligence, automation and work. *National Bureau of Economic Research*, 2019a.

D. Acemoglu and P. Restrepo. Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives*, 33(2):3–30, May 2019b. ISSN 0895-3309. doi: 10.1257/jep.33.2.3. URL https://pubs.aeaweb.org/doi/10.1257/jep.33.2.3.

D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo. Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1):S293–S340, Apr. 2022. ISSN 0734-306X, 1537-5307. doi: 10.1086/718327. URL https://www.journals.uchicago.edu/doi/10.1086/718327.

A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham. Review of Big Data Analytics, Artificial Intelligence and Nature-Inspired Computing Models towards Accurate Detection of COVID-19 Pandemic Cases and Contact Tracing. *International Journal of Environmental Research and Public Health*, 17(15): 5330, jul 2020. ISSN 1660-4601. doi: 10.3390/ijerph17155330. URL https://www.mdpi.com/1660-4601/17/15/5330.

P. Aghion and P. Howitt. A model of growth through creative destruction, 1990.

P. Aghion, C. Antonin, and S. Bunel. Artificial intelligence, growth and employment: The role of policy. *Economie et Statistique/Economics and Statistics*, (510-511-512):150–164, 2019.

A. Agrawal, J. McHale, and A. Oettl. Finding needles in haystacks: Artificial intelligence and recombinant growth. In *The economics of artificial intelligence: An agenda*, pages 149–174. University of Chicago Press, 2018.

A. Agrawal, J. Gans, and A. Goldfarb. The economics of artificial intelligence. NBER, 2019a.

A. Agrawal, J. S. Gans, and A. Goldfarb. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6, 2019b.

A. Agrawal, J. S. Gans, and A. Goldfarb. Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*, 33(2):31–50, May 2019c. ISSN 0895-3309. doi: 10.1257/jep.33.2.31. URL https://pubs.aeaweb.org/doi/10.1257/jep.33.2.31.

N. Ahmed and M. Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.

N. Ahmed, M. Wahed, and N. C. Thompson. The growing influence of industry in ai research. *Science*, 379(6635):884–886, 2023. doi: 10.1126/science.ade2420. URL https://www.science.org/doi/abs/10.1126/science.ade2420.

A. S. Ahuja, V. P. Reddy, and O. Marques. Artificial intelligence and covid-19: A multidisciplinary approach. *Integrative medicine research*, 9(3), 2020.

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

Amazon. What is agi?- artificial general intelligence explained - aws, 2023. URL https://aws.amazon.com/what-is/artificial-general-intelligence.

D. Arranz, S. Bianchini, V. Di Girolamo, and J. Ravet. *Trends in the use of AI in science : a bibliometric analysis*. Publications Office of the European Union, 2023. doi: https://data.europa.eu/doi/10.2777/418191.

S. Arts, N. Melluso, and R. Veugelers. Beyond citations: Measuring novel scientific ideas and their impact in publication text. *arXiv e-prints*, pages arXiv–2309, 2023.

B. Attard-Frost, A. De los Ríos, and D. R. Walters. The ethics of ai business practices: A review of 47 ai ethics guidelines. *AI and Ethics*, 3(2):389–406, 2023.

L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

D. Autor, D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen. The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics*, 135(2):645–709, May 2020. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjaa004. URL `https://academic.oup.com/qje/article/135/2/645/5721266`.

D. H. Autor. Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30, 2015.

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

D. Baidoo-Anu and L. O. Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

F. Barbieri, L. E. Anke, and J. Camacho-Collados. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*, 2021.

S. Baruffaldi, B. van Beuzekom, H. Dernis, D. Harhoff, N. Rao, D. Rosenfeld, and M. Squicciarini. Identifying and measuring developments in artificial intelligence: Making the impossible possible. 2020.

T. Besiroglu, N. Emery-Xu, and N. Thompson. Economic impacts of ai-augmented r&d. *arXiv preprint arXiv:2212.08198*, 2022.

S. Bianchini, M. Müller, and P. Pelletier. Artificial intelligence in science: An emerging general method of invention. *Research Policy*, 51(10):104604, 2022.

S. Bianchini, P. Bottero, M. Colagrossi, G. Damioli, C. Ghisetti, K. Michoud, et al. Mapping the scientific base for sdgs and digital technologies, 2023a.

S. Bianchini, G. Damioli, and C. Ghisetti. The environmental effects of the "twin" green and digital transition in european regions. *Environmental and Resource Economics*, 84(4):877–918, 2023b.

S. Bianchini, M. Müller, and P. Pelletier. Drivers and barriers of ai adoption and use in scientific research. *arXiv preprint arXiv:2312.09843*, 2023c.

A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023.

T. Birkstedt, M. Minkkinen, A. Tandon, and M. Mäntymäki. Ai governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

A. Bohr and K. Memarzadeh. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, pages 25–60. Elsevier, 2020.

F. Bordot. Artificial intelligence, robots and unemployment: Evidence from oecd countries. *Journal of Innovation Economics & Management*, (1):117–138, 2022.

A. Borsato and A. Lorentz. The kaldor–verdoorn law at the age of robots and ai. *Research Policy*, 52(10):104873, 2023.

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

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

I. Brown and C. T. Marsden. *Regulating code: Good governance and better regulation in the information age*. MIT Press, 2023.

M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

E. Brynjolfsson and A. McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company, 2014.

E. Brynjolfsson and A. McAfee. The business of artificial intelligence. *Harvard Business Review*, 2017.

E. Brynjolfsson, T. Mitchell, and D. Rock. What Can Machines Learn and What Does It Mean for Occupations and the Economy? *AEA Papers and Proceedings*, 108:43–47, May 2018. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20181019. URL `https://pubs.aeaweb.org/doi/10.1257/pandp.20181019`.

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023. URL `https://www.nber.org/papers/w31161`. No. w31161.

J. Bughin et al. *Artificial Intelligence: The Next Digital Frontier?* McKinsey Global Institute, 2018.

J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz. Mapping the landscape of artificial intelligence applications against covid-19. *Journal of Artificial Intelligence Research*, 69:807–845, 2020.

J. Butcher and I. Beridze. What is the state of artificial intelligence governance globally? *The RUSI Journal*, 164(5-6):88–96, 2019.

M. Campbell, A. J. Hoane Jr, and F.-h. Hsu. Deep blue. *Artificial intelligence*, 134 (1-2):57–83, 2002.

D. Cardon, J.-P. Cointet, A. Mazières, and L. Carey-Libbrecht. Neurons spike back. *Réseaux*, 211(5):173–220, 2018.

C. Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180080, 2018.

S. Cave and K. Dihal. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2):74–78, 2019.

S. Cave, K. Dihal, and S. Dillon. *AI narratives: A history of imaginative thinking about intelligent machines*. Oxford University Press, 2020.

P. E. Ceruzzi. *Computing: a concise history*. MIT press, 2012.

M. Chahrour, S. Assi, M. Bejjani, A. A. Nasrallah, H. Salhab, M. Fares, H. H. Khachfe, H. A. Salhab, and M. Y. Fares. A bibliometric analysis of covid-19 research activity: a call for increased output. *Cureus*, 12(3), 2020.

L. Chen et al. International competitiveness and the fourth industrial revolution. *Entrepreneurial Business and Economics Review*, 5(4):111–133, 2017.

J. Chubb, P. Cowling, and D. Reed. Speeding up to keep up: exploring the use of ai in the research process. *AI & Society*, 37:1439–1457, 2021. doi: 10.1007/s00146-021-01259-0.

M. Chui, E. Hazan, R. Roberts, A. Singla, and K. Smaje. The economic potential of generative ai. *Report*, 2023.

M. Chui et al. The power of artificial intelligence: An executive's guide. *McKinsey Quarterly*, 2018.

M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

I. M. Cockburn, R. Henderson, and S. Stern. The impact of artificial intelligence on innovation: An exploratory analysis. In *The economics of artificial intelligence: An agenda*, pages 115–146. University of Chicago Press, 2018.

E. Commission. Regulatory framework proposal on artificial intelligence, 2021. URL `https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`.

K. Conger and L. Hirsch. Elon musk completes 44 billion deal to own twitter. *The New York Times*, 2022.

J. Cowls. 'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good. *Philosophy & Technology*, 34(S1):1–5, Nov. 2021. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-021-00466-3. URL `https://link.springer.com/10.1007/s13347-021-00466-3`.

A. De Mauro, M. Greco, and M. Grimaldi. What is big data? a consensual definition and a review of key research topics. In *AIP conference proceedings*, volume 1644, pages 97–104. American Institute of Physics, 2015.

A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

G. Domini, M. Grazzi, D. Moschella, and T. Treibich. Threats and opportunities in the digital era: Automation spikes and employment dynamics. *Research Policy*, 50(7):104137, 2021.

G. Dosi. Technological paradigms and technological trajectories. 1982.

D. G. Douglas. *The Social Construction of Technological Systems, anniversary edition: New Directions in the Sociology and History of Technology*. MIT press, 2012.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.

B. Eastwood. Study: Industry now dominates AI research — MIT Sloan — mitsloan.mit.edu. `https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research`, 2023. [Accessed 01-Jul-2023].

U. K. Ecker, S. Lewandowsky, B. Swire, and D. Chang. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic bulletin & review*, 18:570–578, 2011.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

E. Fast and E. Horvitz. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

L. Floridi, J. Cowls, T. C. King, and M. Taddeo. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 26(3):1771–1796, June 2020. ISSN 1353-3452, 1471-5546. doi: 10.1007/s11948-020-00213-5. URL `http://link.springer.com/10.1007/s11948-020-00213-5`.

M. Fontana, M. Iori, F. Montobbio, and R. Sinatra. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, 49(7):104063, 2020.

M. Fontana, M. Iori, V. L. Sciabolazza, and D. Souza. The interdisciplinarity dilemma: Public versus private interests. *Research Policy*, 51(7):104553, 2022.

M. Ford. *The Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, 2015.

S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

M. R. Frank, D. Wang, M. Cebrian, and I. Rahwan. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2):79–85, 2019.

C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 2017. doi: 10.1016/j.techfore.2016.08.019.

C. V. Fry, X. Cai, Y. Zhang, and C. S. Wagner. Consolidation in a crisis: Patterns of international collaboration in early covid-19 research. *PloS one*, 15(7):e0236307, 2020.

A. Fung. Varieties of participation in complex governance. *Public administration review*, 66:66–75, 2006.

J. Furman and R. Seamans. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191, 2019.

F. Galindo-Rueda. How are science, technology and innovation going digital? the statistical evidence. *The Digitalisation of Science, Technology and Innovation*, page 51, 2020, OECD.

F. Gargiulo, M. Castaldo, T. Venturini, and P. Frasca. Distribution of labor, productivity and innovation in collaborative science. *Applied Network Science*, 7(1): 19, 2022.

F. Gargiulo, S. Fontaine, M. Dubois, and P. Tubaro. A meso-scale cartography of the ai ecosystem. *Quantitative Science Studies*, 4(3):574–593, 2023.

K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922, 2018.

G. M. Garrido, J. Sedlmeir, Ö. Uludağ, I. S. Alaoui, A. Luckow, and F. Matthes. Revealing the landscape of privacy-enhancing technologies in the context of data markets for the iot: A systematic literature review. *Journal of Network and Computer Applications*, 207:103465, 2022.

F. W. Geels. Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. *Research Policy*, 2002. doi: 10.1016/s0048-7333(02)00062-8.

A. Gefen, L. Saint-Raymond, and T. Venturini. Ai for digital humanities and computational social sciences. *Reflections on Artificial Intelligence for Humanity*, pages 191–202, 2021.

T. Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

A. Gillioz, J. Casas, E. Mugellini, and O. Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

H.-H. Goh. Artificial intelligence in achieving sustainable development goals. *arXiv preprint arXiv:2107.13966*, 2021.

GoogleDeepMind. Using ai to fight climate change, 2023. URL `https://deepmind.google/discover/blog/using-ai-to-fight-climate-change/`.

M. A. Goralski and T. K. Tan. Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1):100330, Mar. 2020. ISSN 14728117. doi: 10.1016/j.ijme.2019.100330. URL `https://linkinghub.elsevier.com/retrieve/pii/S1472811719300138`.

R. Gozalo-Brizuela and E. C. Garrido-Merchan. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*, 2023.

T. Greenhalgh, J. Wherton, C. Papoutsi, J. Lynch, G. Hughes, S. Hinder, N. Fahy, R. Procter, S. Shaw, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*, 19(11):e8775, 2017.

Z. Griliches. Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4):501–522, 1957. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1905380`.

S. Guenat, P. Purnell, Z. G. Davies, M. Nawrath, L. C. Stringer, G. R. Babu, M. Balasubramanian, E. E. F. Ballantyne, B. K. Bylappa, B. Chen, P. De Jager, A. Del Prete, A. Di Nuovo, C. O. Ehi-Eromosele, M. Eskandari Torbaghan, K. L. Evans, M. Fraundorfer, W. Haouas, J. U. Izunobi, J. C. Jauregui-Correa, B. Y. Kaddouh, S. Lewycka, A. C. MacIntosh, C. Mady, C. Maple, W. N. Mhiret, R. K. Mohammed-Amin, O. C. Olawole, T. Oluseyi, C. Orfila, A. Ossola, M. Pfeifer, T. Pridmore, M. L. Rijal, C. C. Rega-Brodsky, I. D. Robertson, C. D. F. Rogers, C. Rougé, M. B. Rumaney, M. K. Seeletso, M. Z. Shaqura, L. M. Suresh, M. N. Sweeting, N. Taylor Buck, M. U. Ukwuru, T. Verbeek, H. Voss, Z. Wadud, X. Wang, N. Winn, and M. Dallimer. Meeting sustainable development goals via robotics and autonomous systems. *Nature Communications*, 13 (1):3559, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31150-5. URL `https://www.nature.com/articles/s41467-022-31150-5`.

X. Han, Y.-T. Wang, J.-L. Feng, C. Deng, Z.-H. Chen, Y.-A. Huang, H. Su, L. Hu, and P.-W. Hu. A survey of transformer-based multimodal pre-trained modals. *Neurocomputing*, 515:89–106, 2023.

M. P. Hekkert, R. A. Suurs, S. O. Negro, S. Kuhlmann, and R. E. Smits. Functions of innovation systems: A new approach for analysing technological change. *Technological forecasting and social change*, 74(4):413–432, 2007.

E. Helpman. *General purpose technologies and economic growth*. MIT press, 1998.

J. R. Holm, D. S. Hain, R. Jurowetzki, and E. Lorenz. Innovation dynamics in the age of artificial intelligence: introduction to the special issue, 2023.

L. Hong and S. E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

M. C. Horowitz. Public opinion and the politics of the killer robots debate. *Research & Politics*, 3(1):2053168015627183, 2016.

S. Hu, L. Shen, Y. Zhang, Y. Chen, and D. Tao. On transforming reinforcement learning by transformer: The development trajectory. *ArXiv*, abs/2212.14164, 2022.

T. P. Hughes et al. The evolution of large technological systems. *The social construction of technological systems: New directions in the sociology and history of technology*, 82:51–82, 1987.

K. Huutoniemi, J. T. Klein, H. Bruun, and J. Hukkinen. Analyzing interdisciplinarity: Typology and indicators. *Research policy*, 39(1):79–88, 2010.

S. J. Jackson, M. Bailey, and B. F. Welles. *# HashtagActivism: Networks of race and gender justice*. Mit Press, 2020.

S. Jacobsson and V. Lauber. The politics and policy of energy system transformation—explaining the german diffusion of renewable energy technology. *Energy policy*, 34(3):256–276, 2006.

S. Jasanoff. *Technologies of humility: Citizen participation in governing science*. Springer, 2005.

F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

B. Jindra and M. Leusin. *The development of digital sustainability technologies by top R&D investors*. Publications Office of the European Union, 2022.

A. Jo. The promise and peril of generative ai. *Nature*, 614(1):214–216, 2023.

A. Jobin, M. Ienca, and E. Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

B. F. Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

S. Jordan, C. Fontaine, and R. Hendricks-Sturrup. Selecting privacy-enhancing technologies for managing health data use. *Frontiers in Public Health*, 10:814163, 2022.

R. Jurowetzki, S. Bianchini, D. S. Hain, and K. Wirtz. The private sector is hoarding ai researchers: Drivers and consequences. MIMEO, 2023.

J. E. Katz, D. Halpern, and E. T. Crocker. In the company of robots: views of acceptability of robots in social settings. In *Social robots from a human perspective*, pages 25–38. Springer, 2015.

P. G. Kelley, Y. Yang, C. Heldreth, C. Moessner, A. Sedley, A. Kramm, D. T. Newman, and A. Woodruff. Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 627–637, 2021.

M. Khan, M. T. Mehran, Z. U. Haq, Z. Ullah, S. R. Naqvi, M. Ihsan, and H. Abbass. Applications of artificial intelligence in covid-19 pandemic: A comprehensive review. *Expert systems with applications*, 185:115695, 2021.

M. A. Khan and K. Salah. Iot security: Review, blockchain solutions, and open challenges. *Future generation computer systems*, 82:395–411, 2018.

S. Khin and T. C. Ho. Digital technology, digital capability and organizational performance: A mediating role of digital innovation. *International Journal of Innovation Science*, 11(2):177–195, 2018.

J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos. Deep learning, deep change? mapping the development of the artificial intelligence general purpose technology. *arXiv preprint arXiv:1808.06355*, 2018.

J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos. A narrowing of ai research? *arXiv preprint arXiv:2009.10385*, 2020.

M. Koehler and H. Sauermann. Algorithmic management in scientific research. *Available at SSRN 4497871*, 2023.

M. Koehler and H. Sauermann. Algorithmic management in scientific research. *Research Policy*, 53(4):104985, 2024.

M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4 (12):761–769, 2022.

T. S. Kuhn. The structure of scientific revolutions. *The Un of Chicago Press*, 2:90, 1962.

R. Kurzweil. The singularity is near. In *Ethics and emerging technologies*, pages 393–406. Springer, 2005.

M. La Quatra and L. Cagliero. Transformer-based highlights extraction from scientific papers. *Knowledge-Based Systems*, 252:109382, 2022.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

J. Lanier. *Dawn of the new everything: Encounters with reality and virtual reality*. Henry Holt and Company, 2017.

V. Larivière, Y. Gingras, C. R. Sugimoto, and A. Tsou. Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 66(7):1323–1332, 2015.

B. Latour. *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford, 2007.

M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers. Building Efficient Universal Classifiers with Natural Language Inference, Dec. 2023. URL `http://arxiv.org/abs/2312.17543`. arXiv:2312.17543 [cs].

D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

E. Leahey. From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual review of sociology*, 42:81–100, 2016.

E. Leahey, C. M. Beckman, and T. L. Stanko. Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly*, 62(1):105–139, 2017.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Y.-N. Lee, J. P. Walsh, and J. Wang. Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3):684–697, 2015.

M. E. Leusin. The development of al in multinational enterprises-effects upon technological trajectories and innovation performance. 2022.

J. M. Levitt and M. Thelwall. Is multidisciplinary research more highly cited? a macrolevel study. *Journal of the American Society for Information Science and Technology*, 59(12):1973–1984, 2008.

S. Lewandowsky, U. K. Ecker, and J. Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369, 2017.

L. Leydesdorff and I. Rafols. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1):87–100, 2011.

G. L. Liehner, H. Biermann, A. Hick, P. Brauner, and M. Ziefle. Perceptions, attitudes and trust towards artificial intelligence—an assessment of the public opinion. *Artificial Intelligence and Social Computing*, 72(72), 2023.

T. Lin, Y. Wang, X. Liu, and X. Qiu. A survey of transformers. *AI Open*, 2022.

R. G. Lipsey, K. I. Carlaw, and C. T. Bekar. *Economic transformations: general purpose technologies and long-term economic growth*. Oup Oxford, 2005.

X. Liu, J. Zhao, R. Liu, and K. Liu. Event history analysis of the duration of online public opinions regarding major health emergencies. *Frontiers in Psychology*, 13: 954559, 2022.

M. Luengo-Oroz, K. Hoffmann Pham, J. Bullock, R. Kirkpatrick, A. Luccioni, S. Rubel, C. Wachholz, M. Chakchouk, P. Biggs, T. Nguyen, et al. Artificial intelligence cooperation to support the global response to covid-19. *Nature Machine Intelligence*, 2(6):295–297, 2020.

G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

G. Marcus and E. Davis. *Rebooting AI: Building artificial intelligence we can trust*. Vintage, 2019.

B. Marr. A short history of chatgpt: How we got to where we are today. *Forbes. Accessed on May*, 23:2023, 2023.

N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, and R. Perrault. *The AI Index 2023 Annual Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA*. 2023.

J. Mateos-Garcia and J. Klinger. *Is there a narrowing of AI research?* OECD Publishing, 2023. URL `https://doi.org/10.1787/77709ef0-en`.

M. Maxwell, S. Donald, et al. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187, 1972.

J. McCarthy. Lisp programmers manual, handwritten draft, 1959.

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12, 2006.

K. McElheran, J. F. Li, E. Brynjolfsson, Z. Kroff, E. Dinlersoz, L. S. Foster, and N. Zolas. Ai adoption in america: Who, what, and where. Technical report, National Bureau of Economic Research, 2023.

A. V. Merenkov, R. Campa, and N. Dronishinets. Public opinion on artificial intelligence development. *KnE Social Sciences*, pages 565–574, 2021.

Y.-K. Min, S.-G. Lee, and Y. Aoshima. A comparative study on industrial spillover effects among korea, china, the usa, germany and japan. *Industrial Management & Data Systems*, 2019. doi: 10.1108/imds-05-2018-0215.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, jan 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa913. URL `https://academic.oup.com/nar/article/49/D1/D412/5943818`.

B. Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, 2019.

B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. *The ethics of algorithms: Mapping the debate*. Big data & society, 2016.

J. Mondolo. The composite link between technological change and employment: A survey of the literature. *Journal of Economic Surveys*, 36(4):1027–1068, 2022.

J. Moore. AI for Not Bad. *Frontiers in Big Data*, 2:32, Sept. 2019. ISSN 2624-909X. doi: 10.3389/fdata.2019.00032. URL `https://www.frontiersin.org/article/10.3389/fdata.2019.00032/full`.

R. Morrar, H. Arman, and S. Mousa. The fourth industrial revolution (industry 4.0): A social innovation perspective. *Technology innovation management review*, 7(11):12–20, 2017.

M. Mosleh and D. G. Rand. Measuring exposure to misinformation from political elites on twitter. *nature communications*, 13(1):7144, 2022.

B. Murdoch. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1):1–5, 2021.

R. Musa Giuliano. Echoes of myth and magic in the language of artificial intelligence. *AI & society*, 35(4):1009–1024, 2020.

P. Nagy and G. Neff. Imagined affordance: Reconstructing a keyword for communication theory. *Social Media+ Society*, 1(2):2056305115603385, 2015.

W. Naudé. Artificial intelligence vs covid-19: limitations, constraints and pitfalls. *AI & society*, 35(3):761–765, 2020.

W. Naudé. Artificial intelligence: neither utopian nor apocalyptic impacts soon. *Economics of Innovation and new technology*, 30(1):1–23, 2021.

G. Neff and P. Nagy. Agency in the digital age: Using symbiotic agency to explain human-technology interaction. 2018.

R. R. Nelson and B. N. Sampat. Making sense of institutions as a factor shaping economic performance. *Journal of economic behavior & organization*, 44(1):31–54, 2001.

A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

K. Y. Ngiam and W. Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

N. J. Nilsson. *The quest for artificial intelligence*. Cambridge University Press, 2009.

R. V. Noorden and J. M. Perkel. AI and science: what 1,600 researchers think. *Nature*, 621(7980):672–675, September 2023. doi: 10.1038/d41586-023-02980-. URL `https://ideas.repec.org/a/nat/nature/v621y2023i7980d10.1038_d41586-023-02980-0.html`.

B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

L. Obozintsev. *From Skynet to Siri: An exploration of the nature and effects of media coverage of artificial intelligence*. University of Delaware, 2018.

OECD. Ai language models: Technological, socio-economic and policy considerations. Technical Report No. 352, OECD Digital Economy Papers, 2023a. URL `https://doi.org/10.1787/13d38f92-en`.

OECD. Artificial intelligence in science: Challenges, opportunities and the future of research. *OECD Publishing, Paris*, 2023b.

K. Okamura. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1), 2019.

C. O'Neil. Weapons of math destruction: How big data increases inequality and threatens democracy. Website, 2016.

A. Orben and A. K. Przybylski. The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 2019. doi: 10.1038/s41562-018-0506-1.

E. Ostrom. *Understanding institutional diversity*. Princeton university press, 2009.

R. Paarlberg. The global food fight. *Foreign Aff.*, 79:24, 2000.

B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

Y. Papadopoulos and P. Warin. Are innovative, participatory and deliberative procedures in policy making democratic and effective? *European journal of political research*, 46(4):445–472, 2007.

E. Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

J. J. Park, J. Tiefenbach, and A. K. Demetriades. The role of artificial intelligence in surgical simulation. *Frontiers in Medical Technology*, 4:1076755, 2022.

F. Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

C. Perez. Technological revolutions and financial capital: The dynamics of bubbles and golden ages. *Foreign Affairs*, 2003. doi: 10.2307/20033522.

C. Perez. Technological revolutions, paradigm shifts and socio-institutional change. *Globalization, economic development and inequality: An alternative perspective*, pages 217–242, 2004.

C. Perez. Technological revolutions and techno-economic paradigms. *Cambridge Journal of Economics*, 34(1):185–202, Jan. 2010. ISSN 0309-166X, 1464-3545. doi: 10.1093/cje/bep051. URL `https://academic.oup.com/cje/article-lookup/doi/10.1093/cje/bep051`.

S. Perkowitz. Hollywood science. In *Hollywood Science*. Columbia University Press, 2007.

T. Philbeck and N. Davis. The fourth industrial revolution. *Journal of International Affairs*, 72(1):17–22, 2018.

F. Piccialli, V. S. Di Cola, F. Giampaolo, and S. Cuomo. The role of artificial intelligence in fighting the covid-19 pandemic. *Information Systems Frontiers*, 23 (6):1467–1497, 2021.

T. J. Pinch and W. E. Bijker. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, 14(3):399–441, 1984.

A. Porter, A. Cohen, J. David Roessner, and M. Perreault. Measuring researcher interdisciplinarity. *Scientometrics*, 72(1):117–147, 2007.

J. Priem, H. Piwowar, and R. Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

W. Quattrociocchi, A. Scala, and C. R. Sunstein. Echo chambers on facebook. *Available at SSRN 2795110*, 2016.

J. Radhakrishnan and M. Chattopadhyay. Determinants and barriers of artificial intelligence adoption–a literature review. In *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, Proceedings, Part I*, pages 89–99. Springer, 2020.

L. Rainie and J. Anderson. The future of jobs and jobs training. 2017.

S. Ramli, M. Mustafa, A. Hussain, and D. Wahab. Histogram of intensity feature extraction for automatic plastic bottle recycling system using machine vision. *American Journal of Environmental Sciences*, 4:583–588, 2008. doi: 10.3844/ajessp.2008.583.588.

C. Rammer, D. Czarnitzki, and G. P. Fernández. Artificial intelligence and industrial innovation: Evidence from firm-level data. *ZEW-Centre for European Economic Research Discussion Paper*, (21-036), 2021.

N. Rane. Transformers for medical image analysis: Applications, challenges, and future scope. *Challenges, and Future Scope*, November 2023. Available at SSRN: `https://ssrn.com/abstract=4622241` or `http://dx.doi.org/10.2139/ssrn.4622241`.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, and N. Carvalhais. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204, 2019. doi: 10.1038/s41586-019-0912-1.

M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

E. M. Rogers, A. Singhal, and M. M. Quinlan. Diffusion of innovations. In *An integrated approach to communication theory and research*, pages 432–448. Routledge, 1962.

P. M. Romer. Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102, 1990.

N. Rosenberg. Factors affecting the diffusion of technology. *Explorations in economic history*, 10(1):3, 1972.

N. Rosenberg. Technological interdependence in the american economy. *Technology and Culture*, 20(1):25–50, 1979.

M. Roser. Ai timelines: What do experts in artificial intelligence expect for the future?, Dec 2023. URL `https://ourworldindata.org/ai-timelines`.

P. Ross and K. Maynard. Towards a 4th industrial revolution. *Intelligent Buildings International*, 13(3):159–161, 2021. doi: 10.1080/17508975.2021.1873625. URL https://doi.org/10.1080/17508975.2021.1873625.

D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al. Learning internal representations by error propagation, 1985.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2016.

N. Salari, A. Hosseinian-Far, R. Jalali, A. Vaisi-Raygani, S. Rasoulpoor, M. Mohammadi, S. Rasoulpoor, and B. Khaledi-Paveh. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. *Globalization and Health*, 16(1):57, dec 2020. ISSN 1744-8603. doi: 10.1186/s12992-020-00589-w. URL https://globalizationandhealth.biomedcentral.com/articles/10.1186/s12992-020-00589-w.

A. Salles, K. Evers, and M. Farisco. Anthropomorphism in ai. ajob neuroscience 11 (2), 88-95, 2020.

L. Sartori and G. Bocca. Minding the gap (s): public perceptions of ai and socio-technical imaginaries. *AI & SOCIETY*, pages 1–16, 2022.

M. Savona. The value of data: Towards a framework to redistribute it. 2019.

M. A. Schilling and E. Green. Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. *Research Policy*, 40(10): 1321–1331, 2011.

K. Schwab. *The fourth industrial revolution*. Currency, 2017.

N. Schwalbe and B. Wahl. Artificial intelligence and the future of global health. *The Lancet*, 395(10236):1579–1586, 2020.

F. Series. Harnessing artificial intelligence for the earth. In *World Economic Forum System Initiative on Shaping the Future of Environment and Natural Resource Security in partnership with PwC and the Stanford Woods Institute for the Environment*, 2018.

J. Sevilla, L. Heim, A. C. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. Compute trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. URL https://api.semanticscholar.org/CorpusID:246822642.

R. M. Solow. Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics*, 39(3):312, Aug. 1957. ISSN 00346535. doi: 10.2307/1926047. URL https://www.jstor.org/stable/1926047?origin=crossref.

H. Song and S. Roh. Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, 13, 2021. doi: 10.1029/2021ms002609.

T. W. Steele and J. C. Stier. The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science*, 51(5):476–484, 2000.

P. Stephan. *How economics shapes science*. Harvard University Press, 2012.

A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society interface*, 4(15):707–719, 2007.

L. Suchman, J. Blomberg, J. E. Orr, and R. Trigg. Reconstructing technologies as social practice. In *The Anthropology of Organisations*, pages 431–447. Routledge, 2017.

C. Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton university press, 2018.

C. R. Sunstein. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press Princeton, NJ, 2001.

M. Swan. *Blockchain: Blueprint for a new economy*. " O'Reilly Media, Inc.", 2015.

A. Taylor and H. R. Greve. Superman or the fantastic four? knowledge combination and experience in innovative teams. *Academy of management journal*, 49(4):723–740, 2006.

L. Taylor. The ethics of big data as a public good: which public? Whose good? 2016.

R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, et al. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

M. K. Thu, S. Beppu, M. Yarime, and S. Shibayama. Role of machine and organizational structure in science. *Plos one*, 17(8):e0272280, 2022.

P. Trammell and A. Korinek. Economic growth under transformative ai. Technical report, National Bureau of Economic Research, 2023. URL `https://www.nber.org/papers/w31815`. No. w31815.

J. Truby. Governing artificial intelligence to benefit the un sustainable development goals. *Sustainable Development*, 28(4):946–959, 2020.

A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 178–185, 2010.

A. M. Turing. Computing machinery and intelligence. *Creative Computing*, 6(1): 44–53, 1950.

J. M. Twenge. *iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy–and completely unprepared for adulthood–and what that means for the rest of us*. Simon and Schuster, 2017.

TWI2050. The digital revolution and sustainable development: Opportunities and challenges. report prepared by the world in 2050 initiative. 2019.

S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

UN-General-Assembly. *Transforming our world: The 2030 agenda for sustainable development*. UN, 2015.

B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

M. Valenzuela, V. Ha, and O. Etzioni. Identifying meaningful citations. In *AAAI workshop: Scholarly big data*, volume 15, page 13, 2015.

J. Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.

R. Van Noorden and J. M. Perkel. Ai and science: what 1,600 researchers think. *Nature*, 621(7980):672–675, 2023.

S. Vannuccini and E. Prytkova. Artificial intelligence's new clothes? a system technology perspective. *Journal of Information Technology*, page 02683962231197824, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani. Bertology meets biology: interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. Fuso Nerini. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11 (1):233, Jan. 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14108-y. URL `https://www.nature.com/articles/s41467-019-14108-y`.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

C. J. Wang, C. Y. Ng, and R. H. Brook. Response to COVID-19 in Taiwan. *JAMA*, 323(14):1341, apr 2020a. ISSN 0098-7484. doi: 10.1001/jama.2020.3151. URL `https://jamanetwork.com/journals/jama/fullarticle/2762689`.

J. Wang, B. Thijs, and W. Glänzel. Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PloS one*, 10(5):e0127298, 2015.

L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020b.

Q. Wang and J. W. Schneider. Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, 1(1):239–263, 2020.

X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, pages 1–36, 2023.

A. Webb. *The big nine: How the tech titans and their thinking machines could warp humanity*. Hachette UK, 2019.

B. D. Weinberg, G. R. Milne, Y. G. Andonova, and F. M. Hajjat. Internet of things: Convenience vs. privacy and secrecy. *Business Horizons*, 58(6):615–624, 2015.

D. M. West. *Digital schools: How technology can transform education*. Brookings Institution Press, 2012.

S. Wineburg and S. McGrew. Lateral reading: Reading less and learning more when evaluating digital information. 2017.

F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, pages 1–9, 2023.

M. Wooldridge. *A brief history of artificial intelligence: what it is, where we are, and where we are going*. Flatiron Books, 2021.

A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

L. Wu, F. Morstatter, K. M. Carley, and H. Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90, 2019.

S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, E. Albu, B. Arshi, V. Bellou, M. M. J. Bonten, D. L. Dahly, J. A. Damen, T. P. A. Debray, V. M. T. de Jong, M. De Vos, P. Dhiman, J. Ensor, S. Gao, M. C. Haller, M. O. Harhay, L. Henckaerts, P. Heus, J. Hoogland, M. Hudda, K. Jenniskens, M. Kammer, N. Kreuzberger, A. Lohmann, B. Levis, K. Luijken, J. Ma, G. P. Martin, D. J. McLernon, C. L. A. Navarro, J. B. Reitsma, J. C. Sergeant, C. Shi,

N. Skoetz, L. J. M. Smits, K. I. E. Snell, M. Sperrin, R. Spijker, E. W. Steyerberg, T. Takada, I. Tzoulaki, S. M. J. van Kuijk, B. C. T. van Bussel, I. C. C. van der Horst, K. Reeve, F. S. van Royen, J. Y. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K. G. M. Moons, and M. van Smeden. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, page m1328, apr 2020. ISSN 1756-1833. doi: 10.1136/bmj.m1328. URL `https://www.bmj.com/lookup/doi/10.1136/bmj.m1328`.

Y. Xu, L. Chen, M. Fang, Y. Wang, and C. Zhang. Deep reinforcement learning with transformers for text adventure games. *2020 IEEE Conference on Games (CoG)*, pages 65–72, 2020.

Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4):100179, 2021.

G.-Z. Yang, B. J. Nelson, R. R. Murphy, H. Choset, H. Christensen, S. H. Collins, P. Dario, K. Goldberg, K. Ikuta, N. Jacobstein, et al. Combating covid-19—the role of robotics in managing public health and infectious diseases, 2020.

A. Yegros-Yegros, I. Rafols, and P. D'este. Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PloS one*, 10(8):e0135095, 2015.

Z. Yu. The impacts of prediction ai on scientists: Evidence from alphafold. *Available at SSRN 4711334*, 2024.

J. J. Yun and Z. Liu. Micro- and macro-dynamics of open innovation with a quadruple-helix model. *Sustainability*, 2019. doi: 10.3390/su11123301.

T. Zemčík. Failure of chatbot tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & SOCIETY*, 36: 361–367, 2021.

B. Zhang and A. Dafoe. Us public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 187–193, 2020.

B. Zhang, M. Anderljung, L. Kahn, N. Dreksler, M. C. Horowitz, and A. Dafoe. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research*, 71:591–666, 2021.

Y. Zhang, C. Liu, M. Liu, T. Liu, H. Lin, C. B. Huang, and L. Ning. Attention is all you need: utilizing attention in ai-enabled drug discovery. *Briefings in Bioinformatics*, 25(1), 2024.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

S. Zuboff. The age of surveillance capitalism. In *Social Theory Re-Wired*, pages 203–213. Routledge, 2023.

List of Figures

List of Tables

Diletta Abbonato

The Role of Artificial Intelligence for Societal Challenges

RÉSUMÉ

Cette thèse examine le rôle de l'intelligence artificielle (IA) dans la résolution des problèmes sociétaux, en se concentrant sur son impact sur la recherche scientifique, le développement industriel et les perceptions du public. Chapitre 1 explore les résultats scientifiques des collaborations interdisciplinaires entre les médecins et les spécialistes de l'IA pendant la pandémie COVID-19. Chapitre 2 traite de l'impact de Transformers sur la science, en mettant l'accent sur le codéveloppement de la technologie de l'IA entre les universités et l'industrie. Chapitre 3 explore les perceptions du public sur les principales technologies de la quatrième révolution industrielle (4IR). La thèse considère l'IA comme une technologie transformatrice, appelant à une gouvernance proactive afin d'optimiser ses avantages et d'atténuer ses risques.

Mots clefs: Intelligence Artificielle; 4ème Révolution Industrielle; Equipes Scientifiques; Impact Social; Impact Scientifique

RÉSUMÉ EN ANGLAIS

This thesis examines the role of artificial intelligence (AI) in addressing societal challenges, focusing on its impact on scientific research, industrial development, and public perceptions. Chapter 1 explores the scientific outcome of interdisciplinary collaborations between physicians and AI specialist during the COVID-19 pandemic. Chapter 2 discusses the impact of Transformers on science, with a focus on the co-development of AI technology between universities and industry. Chapter 3 explores public perceptions on the main technologies of fourth industrial revolution (4IR). The thesis positions AI as a transformative technology, calling for proactive governance to optimize its benefits and mitigate its risks.

Keywords: Artificial Intelligence; 4th Industrial Revolution; Scientific Teams; Social Impact; Scientific Impact