Data-Biased Innovation: Directed Technological Change and the Future of Artificial Intelligence[*]

Carl Benedikt Frey[1,3], Giorgio Presidente[2,3], and Pia Andres[4]

[1]Oxford Martin School, University of Oxford
[2]Institute for European Policy-Making, Bocconi University
[3]Oxford Internet Institute, University of Oxford
[4]Centre for Economic Performance, London School of Economics

December 12, 2024

Abstract

This paper examines how privacy regulation has shaped the trajectory of artificial intelligence (AI) innovation across jurisdictions. We construct a novel taxonomy of AI technologies based on their data intensity and analyze patent applications from 57 countries across 76 industries from 2010 to 2021. Our descriptive analysis reveals three key patterns: a substantial shift from data-saving to data-intensive AI methods over the 2010s, increasing market concentration in innovation among established firms, and pronounced geographic heterogeneity in both innovation output and technological focus around the world. Exploiting variation in firms' exposure to the European Union's General Data Protection Regulation (GDPR), we find that exposed applicants significantly altered their technological trajectories. Applicants with greater exposure to EU markets increased their data-saving patents while decreasing data-intensive ones relative to the pre-GDPR period. This effect is driven primarily by EU-based firms. Additionally, the GDPR appears to have reduced overall AI patenting in the EU while reinforcing the market dominance of established companies.

Keywords: artificial intelligence, privacy regulation, innovation, directed technical change, patents, data-biased technical change, data-saving, data-intensive
JEL classification: O31, O39, K29

[*] We are grateful to Citi for generous funding and Leopoldo Lazcano and Alex Nannig Larsen for excellent research assistance. Frey also acknowledges funding from the Dieter Schwarz Foundation. We also thank Michael Osborne, Chris Russell and Michael Wooldridge for their support in developing our AI taxonomy.

1 Introduction

Recent scholarship argues that the current path of artificial intelligence (AI) may produce various social, economic, and political harms (Acemoglu, 2023*a*,*b*; Frey, 2019; Kasy, 2022). Among other worries, the violation of privacy due to the accumulation of personal data, coupled with enhanced capabilities for behavioral manipulation (Acemoglu, 2023*b*) have led scholars to suggest that companies are building "the wrong kind of AI" (Acemoglu and Restrepo, 2020; Acemoglu and Johnson, 2023).

However, these concerns may manifest differently across regions. In the United States, where data privacy relies on a patchwork of sector-specific regulations rather than comprehensive federal protection, AI research has increasingly focused on data-intensive deep learning methods (Klinger et al., 2020). This institutional environment, together with deep learning's inherent returns to scale, has generated strong corporate incentives for systematic personal data accumulation. China's developmental trajectory demonstrates even stronger tendencies in this direction: state procurement policies have explicitly incentivized AI development for surveillance applications (Beraja et al., 2021; Beraja, Yang and Yuchtman, 2023), and these technologies are now being exported globally (Beraja, Kao, Yang and Yuchtman, 2023). The European Union, in contrast, has pursued a distinctive regulatory approach through its General Data Protection Regulation (GDPR), emphasizing individual privacy protection over data accumulation.

In this paper, we investigate how heterogeneous privacy regulations have impacted innovation output in AI and shaped the trajectory of AI development across jurisdictions. Doing so, we situate our analysis within the literature on directed technological change, which posits that when a factor of production becomes scarce or costly, technological innovation tends to shift toward reducing dependence on that factor (Acemoglu, 1998, 2002; Hanlon, 2015; Acemoglu et al., 2015; Hassler et al., 2021). Examples include high wages and labor shortages spurring the invention and adoption of labor-saving technologies (Habakkuk, 1962; Allen, 2009; Hornbeck and Naidu, 2014; Presidente, 2023), and oil price shocks or carbon taxes aiding the development of green technologies (Acemoglu et al., 2012; Hassler et al., 2021). Following this logic, we hypothesize that the increasing cost of storing and processing personal data resulting from the GDPR (Frey and Presidente, 2024) has prompted companies to invest more in data-saving methods relative to data-intensive ones, thereby shifting the composition of AI patenting.

For our empirical analysis, we develop a novel framework for analyzing AI technologies through the lens of their data intensity. *Deep learning methods* are the most data-hungry, requiring vast datasets to tune millions of parameters effectively. In contrast, *knowledge or rules-based systems* rely on structured rules and expert knowledge rather than extensive data, while *Bayesian methods* offer efficiency by incorporating prior knowledge. In addition, several techniques have emerged to reduce data requirements: *transfer learning* (zero-shot and few-shot learning included) repurposes knowledge across tasks, while *synthetic data* generation creates artificial training examples. We collectively classify these methods as “data-saving”, in contrast to deep learning approaches, which we refer to as “data-intensive”.

We next take our taxonomy to the data. Specifically, we identify relevant patents through keyword searches of patent titles and abstracts in PATSTAT, and analyze them at the patent family level to avoid counting the same invention multiple times across jurisdictions. Based on this dataset, we proceed to document a number of stylized facts about AI patenting around the world. First, we document a striking technological shift from data-saving AI to data-intensive deep learning over the 2010s: while the data-intensive AI stock grew by 52% percent per year over the decade, data-saving patenting recorded a relatively modest growth rate of 19% per year. Notably, however, data-saving patent activity remained below its 2004 level until 2013, but experienced an uptick following the GDPR's implementation in 2018.

Second, the global AI patent distribution reveals pronounced geographic heterogeneity in both volume and technological focus. While the United States and China showed comparable AI patent filings in 2014, China's activity surged following its "Made in China 2025" initiative. Chinese patenting activity is particularly notable in academia, with Chinese universities contributing 86% of global data-intensive AI patents filed by higher education institutions, while Chinese companies represent 37% of private sector data-intensive patents globally. In contrast, U.S. companies lead in data-saving AI with 45% of global corporate applications, while the EU shows relative strength in private sector data-saving technologies despite its overall modest innovation activity in AI.

With these patterns in mind, we next turn to systematically investigate the role of regulation in shaping AI innovation trajectories. By exploiting the timing of the introduction of the European Union's General Data Protection Regulation (GDPR), and taking advantage of applicants' varying exposure to it, we examine how data privacy has impacted technological choices in AI development. For this part of our empirical analysis, we construct a novel dataset of 6,801 patenting entities from 49 countries, spanning the period of the most recent AI wave from 2010 to 2021. To measure applicants' GDPR exposure, we use inter-industry linkages detailed in the OECD Inter-Country Input-Output (ICIO) Tables, following a strategy similar to Frey and Presidente (2024). Specifically, we calculate the proportion of output each 2-digit industry-country pair sells to European consumers and use this data to assess the extent of applicants' regulatory exposure. This approach is crucial as the GDPR pertains solely to personal data. By utilizing the breakdown of the ICIO Tables into goods sent to businesses for intermediate use versus final household consumption, we can exclude business-to-business transactions, which are less likely to be impacted by the regulation, and focus on transactions with end-consumers.

Based on this approach, we provide suggestive evidence that applicants exposed to the GDPR disproportionately reduced their patenting in data-intensive AI technologies, while increasing their patenting in data-saving approaches to AI. Our baseline analysis—including patents filed by firms, universities, governments, and individuals—reveals that GDPR-exposed applicants increased data-saving AI patents while reducing data-intensive ones by approximately 1% in absolute terms. This effect is particularly pronounced among EU-based applicants, who increased data-saving patents by 1.6 percentage points while reducing data-intensive patents by 1.5 percentage points after 2018. The primary drivers of this shift, we find, were firms based in the European Union. This implies that companies responded to the new regulation by shifting their technological efforts away from relying on increasingly costly data, limiting their dependence on it, in line with the larger literature on directed technological change (Acemoglu, 1998, 2002; Hanlon, 2015; Acemoglu et al., 2015; Hassler et al., 2021).

Our paper relates to three strands of research. First, we provide novel empirical evidence on directed technological change in artificial intelligence, demonstrating how regulatory constraints can systematically redirect innovation efforts. This way, we contribute to research on national innovation systems by revealing how data privacy regulation shapes distinct regional innovation trajectories, with implications for technological specialization across jurisdictions (Edler et al., 2023; Edquist, 2013; Fagerberg and Srholec, 2008; Nelson, 1993; Nelson and Nelson, 2002).

Second, we add to a growing body of work investigating the economic consequences of data privacy regulation (Aridor et al., 2020; Goldberg et al., 2024; Goldfarb and Tucker, 2012; Johnson et al., 2022; Johnson, forthcoming; Campbell et al., 2015), which has mostly focused on online outcomes, with some noteworthy exceptions (Frey and Presidente, 2024). Our paper is most closely related to a subset of studies focusing on the impact of privacy regulation on the adoption of technology in healthcare, showing that such regulations, which limit hospitals' ability to release health information, have led to a slower uptake of data-intensive medical technologies (Miller and Tucker, 2009, 2011, 2018). Conversely, Frey and Presidente (2024) show that the GDPR was accompanied by an uptake in patenting related to the development of compliant IT systems. In contrast to these studies, we examine the impact of privacy regulation on the *direction* of technological change in AI.

Finally, we contribute to an emerging literature focusing on the measurement of AI adoption and innovation (Babina et al., 2020; Bonney et al., 2024). For example, Cockburn et al. (2019) have used keyword search to identify patterns of patenting in the AI fields of symbolic systems, learning systems and robotics; while WIPO (2019) has developed another methodology, combining a set of keywords with computer-specific technological classification codes. Similarly, Giczy et al. (2022) trained a machine learning classifier to analyze patent texts and citations, aiming to pinpoint innovations across various domains such as knowledge processing, speech technology, hardware design, evolutionary computation, natural language processing (NLP), computer vision, and planning and control systems. What these approaches have in common is that they do not consider the data-intensity of different AI methodologies. We add to this literature by providing the first taxonomy for classifying AI patents according to their data-intensity, which we proceed to implement.

The remainder of this paper is structured as follows. In Section 2, we discuss our data sources and approaches to measurement. Section 3 discusses our results and probes our key findings in a series of robustness checks. Finally, in Section 4, we provide our conclusions and discuss avenues for future research.

2 Data and Measurement

In this section, we outline our taxonomy for classifying AI patents, as well as the data sources and methods employed in our analysis.

2.1 A Taxonomy of AI

We begin by creating a new dataset of AI patents, categorizing various AI methodologies. Unlike prior research, our aim is to classify these approaches based on their data intensity, employing an extensive set of keywords for this particular differentiation. To compile a comprehensive list keywords, we rely on the literature outlining different AI methodologies (Russell and Norvig, 2016; Goodfellow et al., 2016; Wooldridge, 2020; Marcus, 2018; Michalski et al., 2013), as well as the help of leading experts in the field of Computer Science.[1] The resulting classification scheme, detailed in Table 1, allows us to systematically analyze AI patents based on their data intensity. Below, we describe the key technologies captured in our taxonomy and their distinctive characteristics regarding data requirements.

Deep learning: Keywords like back-propagation, connectionist, deep learning, and neural network all represent the most data-hungry AI technologies. These methods simulate brain processes by adjusting connection weights between artificial neurons to minimize prediction errors (Russell and Norvig, 2016; Goodfellow et al., 2016; Mitchell, 2019). The word connectionist, for example, refers to how nodes (analogous to neurons in the brain) are connected within these networks and how these connections can be strengthened or weakened to process and store information, while back-propagation is an approach that adjusts network weights during training to enhance output accuracy. Moreover, the term

[1] We have benefited from wide-ranging conversations with AI researchers at the Oxford Martin School, the Oxford Internet Institute, the Oxford Department of Computer Science, as well as the Department of Engineering Science. We are especially indebted to Chris Russell, Michael Osborne and Michael Wooldridge for their generous advice.

deep learning denotes the depth of the layers in the network. The reason these approaches are so data-intensive is that the learning process involves adjusting the weights of the network's connections to minimize errors—-more data enables better modeling of complex patterns while preventing overfitting across the many parameters.

Transfer learning: Modern deep learning has developed several sophisticated approaches to reduce data requirements through knowledge transfer (Zhuang et al., 2020). Transfer learning significantly cuts down training data needs by repurposing models across related tasks. This family of techniques includes increasingly sophisticated variants: zero-shot learning enables models like GPT-3 to perform entirely new tasks without any task-specific training examples, while one-shot or few-shot learning allows face recognition systems to identify new individuals from just a few examples. These approaches achieve efficiency by leveraging knowledge accumulated from broader domains—similar to how humans apply existing knowledge to learn new but related skills.

Synthetic data: When real-world data is scarce or costly, synthetic data generation can provide an alternative. This approach uses algorithms, simulations, or generative models to create artificial training examples that mimic real-world data characteristics. In protein structure prediction, for example, DeepMind's AlphaFold incorporated its own high-confidence predictions as additional training data to improve performance (Jumper et al., 2021). Similarly, autonomous vehicle companies generate millions of simulated driving scenarios to train their systems, capturing rare events and dangerous situations that would be impractical or unsafe to collect in reality.

Knowledge-based systems: Expert systems, intelligent knowledge-based systems, knowledge representation, knowledge-based systems, ontologies, and rule-based systems operate using structured, predefined knowledge rather than learning from large datasets. At their core, these systems encode human expertise through explicit rules, logical frameworks, and formal relationships. Medical expert systems like MYCIN examplify this approach by using thousands of if-then rules to diagnose infections and recommend antibiotics, while legal expert systems apply structured rules to assess contract compliance or tax obligations. Moreover, ontologies, such as SNOMED CT in healthcare with its 350,000+ medical concepts, provide the fundamental structure for organizing domain knowledge. In short, what distinguishes these approaches from data-intensive AI is their reliance on explicit knowledge engineering rather than bottom-up machine learning.

Bayesian statistics: Bayesian methods stand out for their ability to perform effectively with limited data by formally incorporating prior knowledge into the learning process. For example, in drug discovery, Bayesian optimization can identify promising compounds after just 50-100 experiments. Similarly, a Bayesian spam filter can achieve reasonable performance with just hundreds of emails by leveraging prior probabilities of word frequencies. This efficiency stems from their mathematical framework that updates initial beliefs (priors) with new evidence, rather than learning patterns from scratch. For this reason, we collectively refer to Bayesian methods and knowledge-based systems as exploiting structure.

Data acquisition methods: Another set of approaches relate to data acquisition methods, which aim to reduce the need for large amounts of manually labeled data. Automatic image annotation, for example, generates labels/tags for images using existing algorithms, reducing manual effort but still requiring initial labeled data for training. Semi-supervised learning, on the other hand, efficiently combines a small amount of labeled data with a large amount of unlabeled data, while self-supervised learning requires no manual labels, instead creating supervisory signals from the data itself, such as predicting missing parts of images. Meanwhile, active learning achieves efficiency by having the model identify the most informative samples for human labeling, minimizing effort by selecting only the most valuable cases to label.

Reinforcement learning: Finally, while reinforcement learning is computationally data-hungry, requiring millions of interactions to learn effectively, it can potentially be privacy-preserving since it may learn through self-generated experiences and simulated environments rather than requiring sensitive personal data like the user records, medical histories, and behavioral data typically needed for supervised learning. For example, game-playing agents like AlphaGo achieved superhuman performance through self-play rather than studying human game records. Moreover, in industrial applications, reinforcement learning systems optimize manufacturing processes or energy management through direct interaction with system models, eliminating the need for historical operational data that might contain sensitive information.

Throughout the paper, we use the terms "deep learning" and "data-intensive" interchangeably. Our measure of data-saving AI, in contrast, encompasses knowledge-based systems, Bayesian statistics, transfer learning and synthetic data. However, we take no position on whether reinforcement learning and data acquisition methods primarily save or consume data. Although they can operate without personal data, many commercial applications heavily depend on it. Recommendation systems, social media feeds, and personalized advertising use reinforcement learning to process individual user interactions, clicks, viewing time, and purchase history. Similarly, modern data acquisition approaches like semi-supervised and active learning, though designed to reduce manual labeling, often require processing larger volumes of user-generated content and behavioral data. Given this ambiguity, we consider these categories separately in our empirical analysis.

2.2 Tagging AI Patents

We next take our taxonomy to the data. The main dataset used in this paper is built from EPO's PATSTAT Global database (2024 Spring Edition). PATSTAT Global contains global patent data, including information on patent classification codes, patent families, applicants and inventors, and full titles and abstracts. To construct our dataset, we first identify patents whose titles or abstracts contain any of the keywords listed in Table 1. We then collect information on the families of tagged patents. Because the same invention can be filed multiple times (both within the same patent office and in multiple jurisdictions), the patent family is the most appropriate unit of analysis when measuring new inventions. We collect all patent classification codes associated with each family, and retain only those which include the International Patent Classification group G06 ('Computing; Calculating; Counting'), which account for more than 80% of identified patents.[2] We use the smallest value of the variable $earliest\_filing\_year$ across all applications in the family as the year the invention was made.

[2] Specifically, 83.36% of all patent families tagged and 81.91% of those with an *earliest_filing_year* between 1980 and 2021.

Category	**Keywords**
Deep learning	back-propagation, connectionist, deep learning, neural network
Synthetic data	synthetic data, virtual sample generation
Transfer learning	few-shot learning, one-shot learning, transfer learning, zero-shot learning
Exploiting structure	bayesian, equivariance AND (machine learning OR algorithm OR computer science), expert system, invariance AND (machine learning OR algorithm OR computer science), knowledge representation, knowledge-based systems, ontology, rule-based systems
Data acquisition	active learning, automatic image annotation, self-supervised learning, semi-supervised learning
Reinforcement learning	reinforcement learning

Table 1: Taxonomy of AI categories

2.3 Applicant-level Data

We next draw on PATSTAT to construct our applicant-level dataset. The EPO provides information on inventors and applicants, along with their country codes—we collect this information for personal and institutional records associated with the patent families we tag. As we are interested in analyzing innovation across all applicant types, we retain records regardless of psn_sector classification. In other words, our dataset contains patents filed by individuals, universities, governments and firms.

In some cases, applicant names are harmonized by the EPO (such that small discrepancies in spelling do not lead to an entity being assigned two separate psn_id numbers). For our empirical analysis in Section 4, we drop all psn_id records which have $psn\_level == 0$ (indicating no harmonization), and retain those which have at least been harmonized via an automated procedure. For our descriptive analysis of cross-country trends in Section 3, however, we include the complete set of AI patents including non-harmonized records to maximize representativeness.

After selecting the applicants associated with our tagged patents, we collect information on all patent families (whether tagged as AI or not) associated with those applicants in PATSTAT. We use this information to proxy applicants' age based on the year they are first observed in PATSTAT; size is proxied by the cumulative stock of patent families—including non-AI ones—and the industries associated to the patents. The latter is based on PATSTAT's Table TLS229, which assigns a set of NACE2 codes to each patent application, alongside a variable $weight$. Specifically, the NACE2 codes are based on a mapping between the International Patent Classification (IPC) and the Statistical Classification of Economic Activities in the European Community (Rev 2). There can be multiple NACE2 codes per application, as there are usually multiple IPC codes associated with a patent. The $weight$ indicates the degree to which the application is associated with each particular industry, with the weights for all NACE2 codes summing up to 1 for each application.

We use the NACE2 codes associated with applications filed between 2000 and 2010 to construct applicant-level NACE2 lists and weights. We then sum up the weights associated with each combination of psn_id and $nace2\_code$ (at the 2-digit level) across all applications associated with the psn_id. We further divide this sum by the sum of all weights per psn_id, such that the weights associated with all of the applicant's NACE2 codes sum up to 1. We then use the NACE2 code with the highest weight as the institution's primary industry and the full set of NACE2 codes and weights to construct our exposure variable in Section 4.2.

Finally, we construct applicant-level patent family counts and family stocks for each type of patent family (e.g. reinforcement learning, deep learning, etc.). For each applicant and category, we calculate annual patent family counts and construct cumulative patent stocks, where for each applicant j and year t, $FamStock_{jt} = FamStock_{jt-1} * 0.85 + FamCount_{jt}$, starting from 1980. Patent families are discounted at an annual rate of 15% to account for the decay in their value over time (Hall et al., 2005). Our baseline analysis uses changes in the respective shares of each category in the overall family stock, effectively measuring changes in the composition of innovation by exposed firms.

As it can take up to three years for new patents to be added, we use data on patent families with an earliest filing year of 2021 or earlier when constructing applicant-level patent counts. Our final dataset covers the period 2010-2021, with patent family stock variables incorporating patenting information going back to 1980 as described above.

3 Descriptive Statistics

Based on our applicant-level dataset, we document several key trends in AI patenting over the past decades, including patents filed by companies, government institutions, universities, and individuals around the world. For simplicity, in the first set of figures, we classify patents as either data-intensive or data-saving, following the taxonomy outlined above.

Figure 1 shows that over time, data-intensive AI patent applications have become much more frequent than data-saving ones. In 2000, data-saving and data-intensive patent applications showed a comparable number of patent families per applicant. The two categories then exhibited parallel growth trajectories through 2012, with only minor variations in relative counts. Beginning in 2013, however, data-intensive patents experienced substantially higher growth rates. This divergence accelerated markedly around 2016—the year Google DeepMind's AlphaGo beat the World Go Champion Lee Sedol. Strikingly, meanwhile, data-saving patenting did not again exceed its 2004 level until 2013. We further note that the recent upsurge in data-saving AI patenting coincides with the implementation of the GDPR in 2018.

Figure 2 shows the same pattern, but this time for disaggregated AI technology classes. Again, the striking increase in patents related to deep learning dwarfs other categories, although we note a marked rise across several categories of AI. Reinforcement learning has established itself as the second most prominent category, experiencing particularly fast growth of 207% since 2018, coinciding with a plateau in knowledge-based systems patents. However, other data-saving approaches have demonstrated significant momentum: between 2018 and 2021, transfer learning patents surged by 185%, while synthetic data generation and Bayesian methods expanded by 86% and 68% respectively. Though these technologies began from modest baselines, their rapid growth indicates rising interest in data-efficient methodologies.

Figure 3 focuses on patents filed by companies, revealing notable temporal shifts in business demographics. Throughout the 2000s and early 2010s, new market entrants and startups under 5 years old collectively accounted for approximately one-third of AI patent applications. However, their share declined markedly to one-fifth in 2013, followed by a further reduction during the 2020 Covid-19 pandemic. In particular, we observe a sharp contraction of data-intensive applications by entrants, which falls abruptly at the onset of the 2020s, after increasing by 40 percentage points in the first decade of the sample. These findings are in line with well-documented trends, such as the decline in business dynamism in the technology sector (Decker et al., 2016), as well as the rise of superstar firms (Autor et al., 2020; Stiebale et al., 2020), whether due to lobbying and protective regulation (Gutiérrez and Philippon, 2019), or changes in technology (Tambe et al., 2020).

We next examine AI patenting trends across geographies and applicant types. In 2014, the United States and China filed a comparable number of AI patents (Figure 4). However, following the launch of Xi Jinping's "Made in China 2025" industrial policy initiative, which designated AI as a priority technology, China's AI patenting activity experienced a dramatic surge, particularly in the public sector (Figure 5). Among universities and government institutions, Chinese applicants account for over 86% and 54% of global AI patents, respectively (compared to 3% and 4%, respectively, for their U.S. counterparts). This dominance is especially pronounced in the data-intensive category, where recent research highlights the role of government procurement in generating training data for private sector use (Beraja et al., 2021; Beraja, Yang and Yuchtman, 2023). Reflecting this, Chinese companies represent 37% of data-intensive patents filed by private entities.

In contrast, data-saving AI sees U.S. companies leading the field, accounting for 45% of global patent applications. While the European Union is notably underrepresented across all categories, it plays a more significant role in private sector data-saving AI, contributing 12% of global data-saving patents filed by companies, and about 13% of total private sector AI patents filed within the EU (compared to 5% in China, where specialization in data-intensive AI is much more skewed; Figure 6). A similar pattern emerges among individual applicants, with the U.S. and EU showing particularly strong shares in data-saving technologies. Finally, we highlight the prominent role of South Korea, which stands out in the public sector, accounting for 29% of global patents filed by government institutions. The prominent share of patents filed by the public sector in countries like China and South Korea underlines their respective traditions of state-led development.

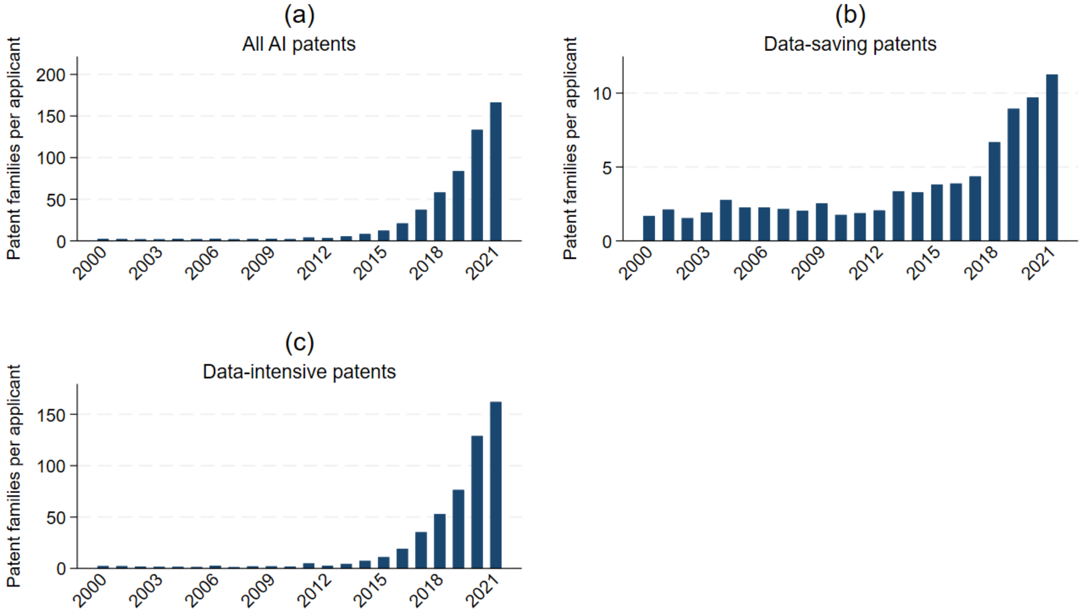
We lastly take a more detailed look at the key players active within disaggregated AI technology classes. Strikingly, China is by far most active across all technology classes, synthetic data being a noteworthy exception (Figure 7). However, we caution against interpreting the findings above as evidence of China leading in AI. Indeed, studies suggest that China's dramatic surge in patent filings may overstate actual innovation. Local government incentives for patent generation, including preferential tax policies and subsidies, have demonstrably reduced patent quality, as reflected in lower renewal rates and low licensing revenue (Long and Wang, 2016).[3] We also note that Europe, while lagging across

[3] Despite China's substantial patent volume, its intellectual property receipts in 2023 amounted to just $11 billion compared to the United States' $126 billion, suggesting that Chinese patents generate limited commercial value. See https://data.worldbank.org/indicator/BX.GSR.ROYL.CD?end=2021&locations=US-CN&start=199. Accessed 19 August, 2024.

all categories of AI, is particularly absent in the domain of data-hungry deep learning systems. Conversely, Europe is better positioned in traditional knowledge-based systems, Bayesian methods, as well as in synthetic data.

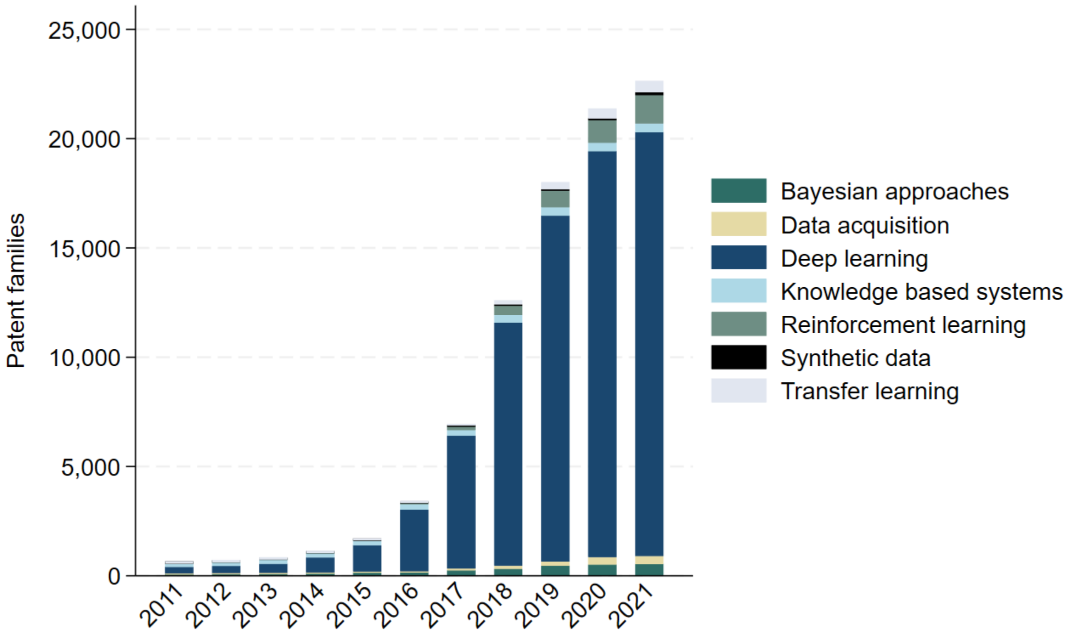
Examining the organizations driving these trends (Figure 8), we observe the extraordinary volume of patents filed by Chinese universities and the Chinese Academy of Sciences (CAS), spanning multiple AI technologies. We further note that established global companies such as IBM and Siemens, along with public organizations like South Korea's Electronics and Telecommunications Research Institute (ETRI), continue to dominate in the old domain of knowledge-based systems. Meanwhile, the contributions of some relatively new players, such as DeepMind (now Google DeepMind), are evident in the emerging field of reinforcement learning.

Figure 1: Average number of AI patent families per applicant



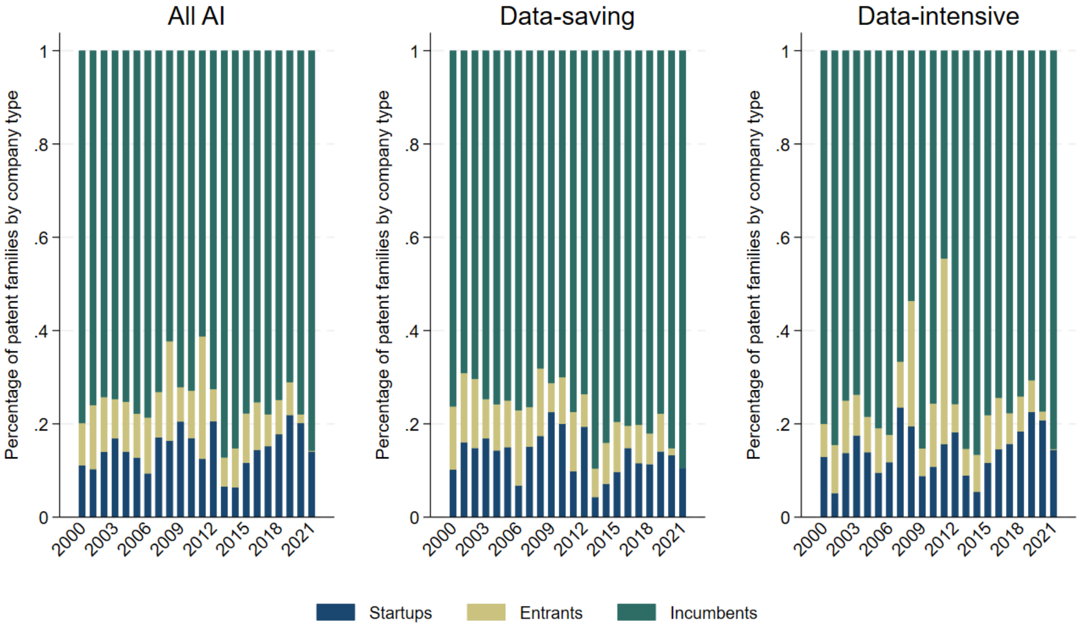
Notes: Each panel shows the average number of patent family applications per applicant across different AI categories based on our full sample of patent-families filed during 2000-2021. The data-saving category (Panel B) entails knowledge-based systems, Bayesian methods, transfer learning, and synthetic data. The data-intensive category (Panel C) comprises deep learning patents. All AI patents (Panel A) include both categories plus reinforcement learning and data acquisition methods.

Figure 2: Patenting across AI technology classes per year



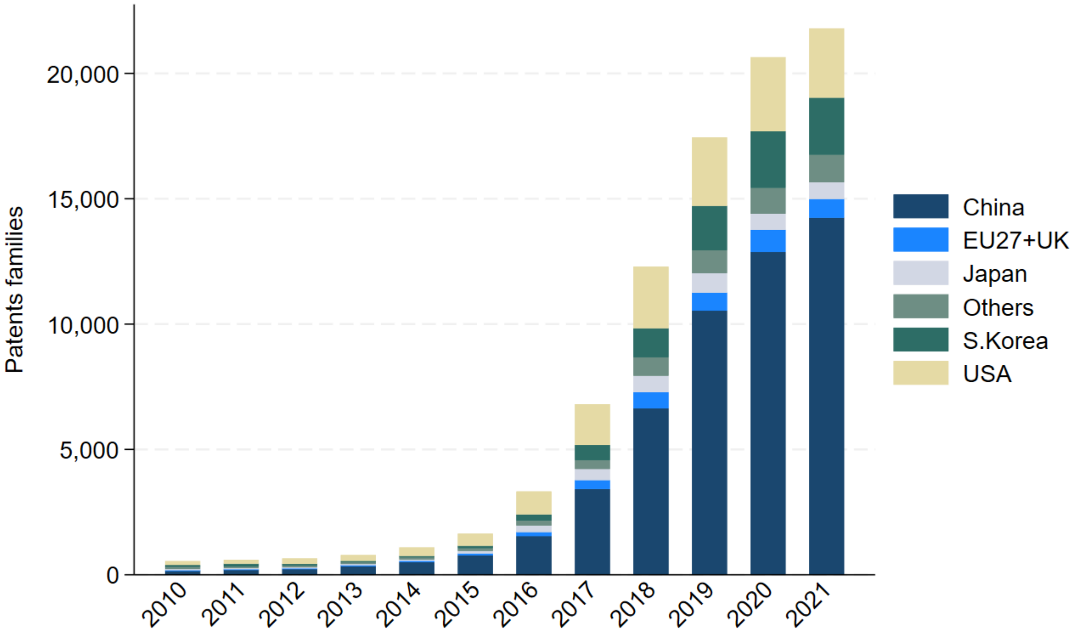
Notes: The figure shows the total number of AI patent families filed each year, classified by different AI categories based on our full sample of patent-families filed during 2011-2021.

Figure 3: Dynamism in AI-related companies



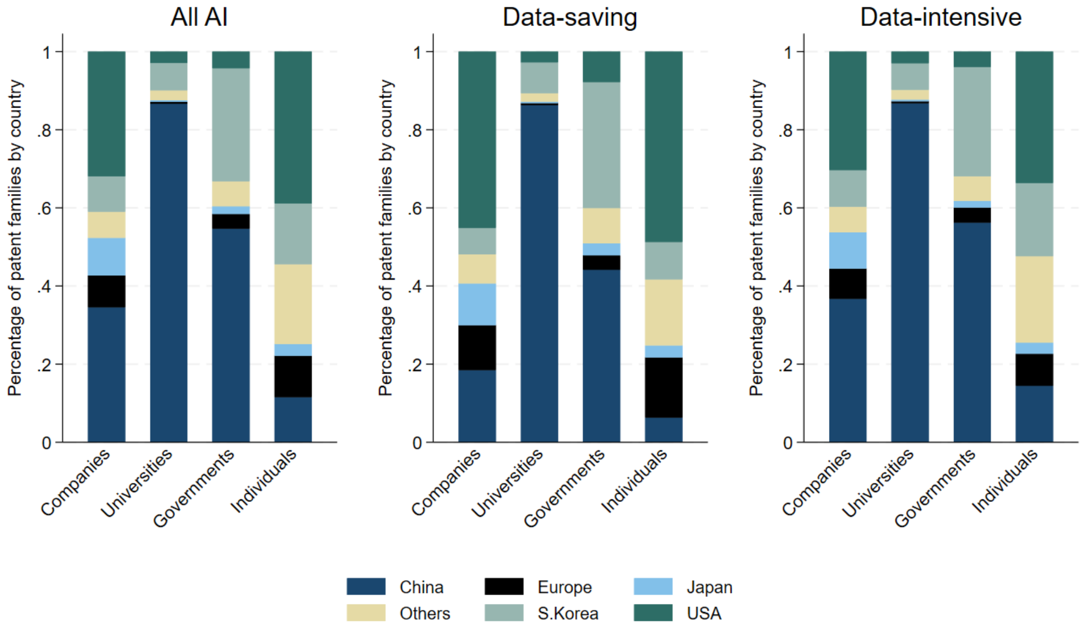
Notes: This figure is based on a sample of patents filed by companies only between 2000-2021. Each panel shows the fraction of total AI patent family applications filed by different types of companies. Entrants are firms with 0 years of existence, startups have between 0 and 5 years, and all others are classified as incumbents.

Figure 4: AI patent families filed by country



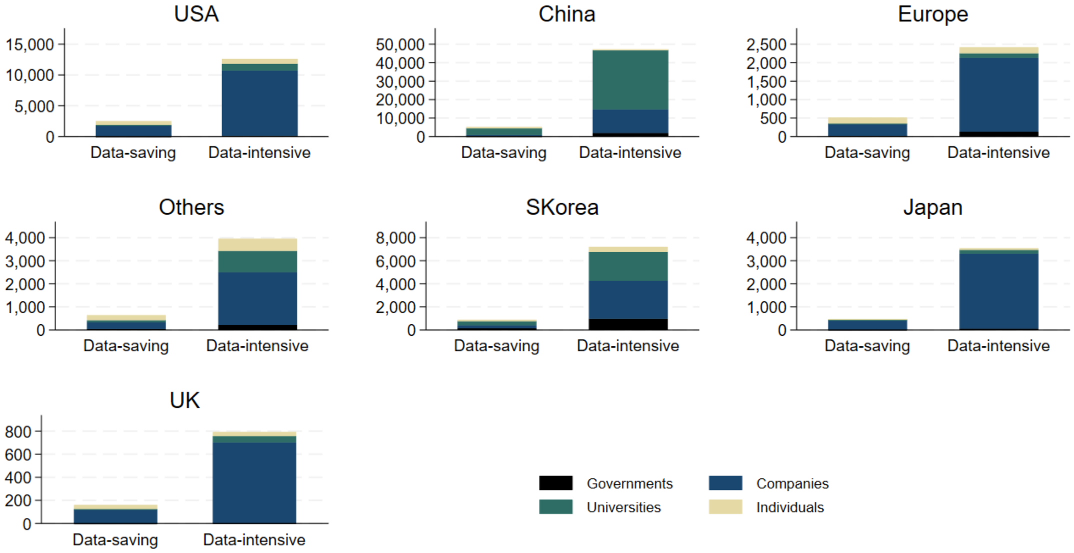
Notes: The figure shows the total number of patent family applications filed worldwide, categorized by the geographical origin of the applicants, based on a sample of patent-families filed during 2010-2021.

Figure 5: AI patent families by institution and country



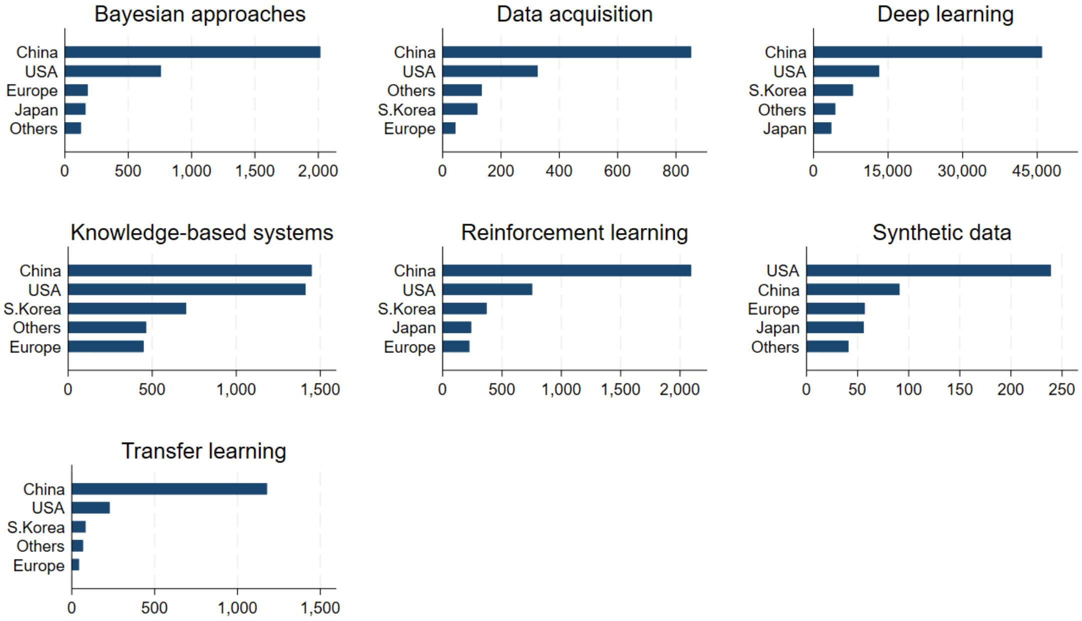
Notes: Each panel represents a different AI category based on our full sample of patent-families filed during 2000-2021. The data-saving category (Panel B) entails knowledge-based systems, Bayesian methods, transfer learning, and synthetic data. The data-intensive category (Panel C) comprises deep learning patents. All AI patents (Panel A) include both categories plus reinforcement learning and data acquisition methods. Each bar corresponds different applicant institutions recorded in PATSTAT, segmented by the applicant's country of origin, indicating the fractional contribution of each country within that institution type.

Figure 6: AI patent family counts by technology class and country



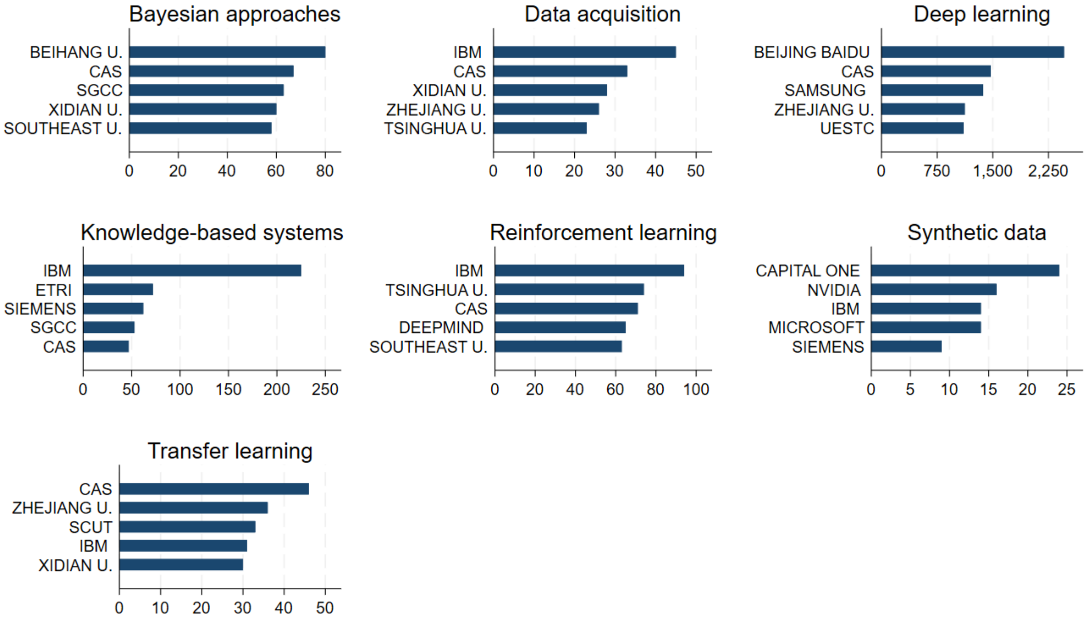
Notes: The figure shows the composition of institution types originating data-saving and data-intensive patent family application counts for each country based on our full sample of patent-families filed during 2000-2021.

Figure 7: Top 5 countries by AI category



Notes: For each AI category, the figure displays the top 5 countries of origin with the highest number of patent family applicants based on our full sample of patent-families filed during 2000-2021.

Figure 8: Top 5 institutions by AI category



Notes: For each AI category, the figure displays the top 5 institutions with the highest number of patent family applicants based on our full sample of patent-families filed during 2000-2021.

4 Data-Biased Technical Change

In this section, we explore how privacy regulation has shaped the trajectory of AI development across space. We begin with a brief overview of the General Data Protection Regulation (GDPR) and its implications for AI patenting. Next, we outline our approach to measuring applicants' exposure to the GDPR. Finally, we detail our empirical strategy and present evidence of varying trends in AI innovation among firms, public institutions, universities, and individual applicants with different levels of exposure to privacy regulation.

4.1 The General Data Protection Regulation

Organizations operating within the European Union must adhere to the GDPR, which delineates guidelines for processing the personal data of EU residents, including browser cookies and IP addresses. Importantly, for our purposes, the GDPR extends its jurisdiction to entities incorporated outside EU countries if they target consumers residing in the European Union. Although enacted in April 2016, enforcement of the regulation began in May 2018, giving businesses a two-year transition period.

The GDPR has increased the cost of storing and processing data in several ways (Johnson et al., 2022; Goldberg et al., 2024; Frey and Presidente, 2024). For instance, since its implementation, websites cannot share user data with third parties without obtaining explicit consent from each user. This severely limits the ability of companies to harvest personal data and to benefit from digital marketing tools based on user profiling and cookie analytics, services often outsourced by online vendors to third-party specialized entities. In addition, the GDPR grants EU residents the rights to access, update, correct, delete, and transfer their personal data. Institutions targeting EU residents are also required to encrypt and anonymize any stored personal data and must regularly audit their data handling processes for compliance. This includes the mandatory appointment of a data protection officer to supervise data management. These compliance requirements impose significant costs on companies and other organizations, particularly those whose business models depend on using highly data-insensitive AI models for prediction.

4.2 Exposure to Privacy Regulation

To measure patent applicants' exposure to privacy regulation, we turn to exploit two key features of the GDPR. First, the regulation concerns *personal* data, meaning that it mainly affects business-to-consumer transactions. Second, its scope extends beyond entities operating within the European Union, applying to organizations—including firms, universities, and individuals—based outside the EU but targeting its consumer market. As a result, the GDPR exerts a global influence, with its reach shaped by the extent to which foreign entities depend on EU consumers.

Our exposure variable is thus based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by patent applicant f based in country c operating in $I > 0$ 2-digit industries indexed by i in the base year 2010. On average, each applicant in our sample operates in 4.48 2-digit industries, which allows us to aggregate the shares of ICT sales sold to EU final consumers by country-industry at the applicant-level, which we denote by

$$S_a^{EU} \equiv \sum_{i_a} \alpha_{i_a} S_{c,i}^{EU}$$

where α_{i_a} is the weight assigned to each NACE2 code associated with a psn_id and $person\_ctry\_code$ pair (recall that the weights sum up to 1 across the applicant's NACE2 codes). In turn, $S_{c,i}^{EU}$ are calculated based on the inter-industry connections provided by the OECD Inter-Country Input-Output Tables (ICIO). These include information on international trade flows across

36 2-digit ISIC Rev. 4 industries and 64 countries.[4] This approach is crucial as the GDPR pertains solely to personal data. By utilizing the breakdown of the ICIO Tables into goods sold to businesses for intermediate use versus final household consumption, we can exclude business-to-business transactions, which are less likely to be impacted by the regulation.

By focusing on the ICT sector ("J" in ISIC rev. 4)—arguably the most relevant in our framework—we are restricting i in $S^{EU}_{c,i}$ to vary across the following three 2-digit industries: (i) "58T60", including "Publishing activities", "Motion picture, video and television program production, sound recording and music publishing activities", and "Programming and broadcasting activities"; (ii) "61 - Telecommunications", and (iii) "62T63", which includes "Computer programming, consultancy and related activities" and "Information service activities". In other words, identification is based on within-ICT sector variation in exposure, as well as on contrasting the patenting activity of entities in the ICT sector to patent applicants in other industries.

4.3 Empirical Strategy

To mitigate endogeneity concerns we restrict our sample to applicants patenting before the start of the study period: NACE2 codes for each applicant are collected based on patents filed during the period 2000-2010. The need for using pre-period industries arises because our exposure measure is based on industries applicants patent in, which are themselves likely to be affected by the exposure of those industries to the GDPR. A drawback with this approach is that using pre-period industries limits our analysis to incumbents and prohibits us from exploring potential effects of the GDPR on firm entry. This is in contrast

[4] ICIO are similar to standard input-output tables, but they include cross-country industry data on both imports and exports. The tables are based on 819 industries, comprising all sectors of economic activity, then aggregated at the 2-digit level by the OECD. For example, "Manufacturing" and "Information and Communications" (ICT) correspond to the 1-digit industries C and J in ISIC rev. 4, which in turn consist of several 2-digit industries.

to the descriptive statistics above, where we use the full sample, including new entrants.

To quantify the differences between firms at various levels of exposure to the GDPR, we estimate the following Two-Way Fixed Effect model:

$$Y_{a,t} = \delta_0 + \delta_1 \, S_a^{EU} \times GDPR + u_f + u_t + \epsilon_{f,t} \tag{1}$$

where $Y_{a,t}$ is the outcome of applicant a in year t and $GDPR$ is a dummy equal to one from 2018 on. $\epsilon_{a,t}$ is an error term. We cluster errors at the applicant level to match the variation in S_a^{EU}.

To ease the interpretation of the results, we normalize S_a^{EU} to have zero mean and unitary standard deviation in the sample, so that the coefficients in the tables correspond to the impact of GDPR for applicants one standard deviation above the average sample exposure.

In the below, we first present the baseline results based on the full sample of patent applicants—including firms, individual applicants, public institutions, and universities. The rationale for including all applicant types is simple: patenting is inherently a commercial activity, and the GDPR affects not only firms but also individual applicants and universities aiming to license and commercialize their innovations. We then focus specifically on corporate applicants to analyze how the GDPR's effects vary based on firm characteristics, like size and age.

4.4 Results

Table 2 presents our baseline results using the full sample. Column 1 shows that the impact of the GDPR on overall AI patenting activity globally is negative but not statistically significant. We further note a small but statistically significant compositional shift: applicants exposed to the GDPR increased their investment in data-saving AI (column 2), while reducing patenting in data-intensive technologies by a similar amount (column 3). Given that the average share of data-saving patents is 0.44, while the average share of deep-learning patents is 0.52, our estimates imply an average change of roughly 1% in absolute value. These findings align with the literature on directed technical change (e.g. Acemoglu, 1998, 2002; Hanlon, 2015; Acemoglu et al., 2015; Hassler et al., 2021), suggesting that when factors of production become scarce—in our case data regulation makes data collection and processing more costly—it will redirect inventive efforts to limit dependence on that factor.

Table 2: Baseline Results

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	-0.001	0.005*	-0.006*
	(0.003)	(0.003)	(0.003)
Observations	53,519	36,249	36,249
R-squared	0.894	0.943	0.939
Applicant FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between AI as a share of the overall patent family stock and each type of AI as a share of the AI family stock in the full sample including individuals, universities, and companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by patent applicant in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Standard errors are clustered at the applicant level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

Given their stronger reliance on European consumer markets, EU-based applicants aiming to bring their patents to market are expected to experience greater impacts from regulatory changes (Frey and Presidente, 2024). Table 3 presents the results of estimating an interaction term between our exposure variable and a dummy flagging EU-based applicants. In column 1, we observe a negative statistically significant impact of the GDPR on

Table 3: Effect on EU27 applicants and others

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	-0.000	0.003	-0.004
	(0.003)	(0.003)	(0.004)
GDPR exposure × post-2018 × EU27	-0.011***	0.016***	-0.015***
	(0.004)	(0.005)	(0.005)
Observations	53,519	36,249	36,249
R-squared	0.894	0.944	0.939
Applicant FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the full sample including individuals, universities, and companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by patent applicant in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. EU27 is a dummy variable equal to 1 if an applicant is based in a EU27 country. Standard errors are clustered at the applicant level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

patenting among applicants located in the EU. Moreover, the coefficients in columns 2 and 3 suggest that the above observed patterns are entirely driven by EU-based entities.[5] As expected, the magnitude of our estimated coefficients is significantly larger: GDPR-exposed EU applicants increased their investment in data-saving AI technologies by 1.6 percentage points more than others (column 2) while reducing patenting in data-intensive technologies by 1.5 percentage points following the regulation's enforcement in 2018 (column 3). We take this finding to reflect a strategic shift of EU entities towards more data-efficient AI practices, influenced by the new regulatory demands, albeit at the expense of the region's overall inventive capacity.

We next explore the role of companies in driving this shift. Column 1 of Table 4 shows that companies were more responsive to the regulation, increasing their overall AI

[5] EU27 includes the United Kingdom until 2019, tagged as non-EU27 thereafter.

patenting relative to universities and individual applicants—who instead decreased their AI patenting activity. Columns 2 and 3 further show that companies were the key drivers of the compositional shift in AI patenting, disproportionally increasing their shares of data-saving AI, while reducing their reliance of data-intensive AI compared to individuals and universities.

4.4.1 Firm Heterogeneity

For this reason, we further restrict the sample to focus on companies. Table 5 presents our results. We note that in response to the introduction of the GDPR, EU companies increased their share of data-saving families by over 30% and reduced their deep learning-related families by more than 20%, with no statistically significant impact on overall AI patenting. Given previous research showing GDPR's disproportionate burden on smaller firms (Frey and Presidente, 2024; Johnson et al., 2023; Peukert et al., 2022), we further examine heterogeneity by firm age and size.[6] Table 8 shows that older firms drive the bulk of the effect: older companies increased patenting in data-saving AI relative to their younger counterparts, while disproportionally reducing AI patenting in data-intensive AI. Table 9, which focuses on size rather than age, shows a similar pattern, although a positive correlation between the GDPR and data-saving AI is also in place for smaller companies. This leads us to conclude that the patterns observed in Table 5 are primarily driven by the oldest and largest EU companies, consistent with evidence of European incumbents' enduring market dominance (Biondi et al., 2023).

[6] This analysis comes with the caveat that in order to limit endogeneity issues, we restrict the sample to companies that were already active in 2010. Thus, our age and size comparison cannot capture entrants.

Table 4: Effect on companies and other applicants

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	-0.106***	-0.012	-0.016
	(0.040)	(0.031)	(0.030)
GDPR exposure × post-2018 × company	0.175***	0.307***	-0.228***
	(0.036)	(0.062)	(0.063)
Observations	44,240	30,356	30,356
R-squared	0.888	0.943	0.938
Applicant FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the full sample including individuals, universities, and companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by patent applicant in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. The dummy "company" takes value 1 if the applicant is a private company. Standard errors are clustered at the applicant level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

Table 5: Effect on companies only

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	0.037	0.140**	-0.110*
	(0.034)	(0.055)	(0.062)
Observations	15,141	9,088	9,088
R-squared	0.791	0.897	0.887
Firm FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the restricted sample focusing on companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

4.4.2 Technology Heterogeneity

Finally, we turn to explore heterogeneity by AI technology class. Table 6 presents the results for our full sample, while Table 7 restricts our sample to firms. We note that across the full sample, applicants appear to have increased patenting activity in knowledge-based systems and Bayesian methods—both of which rely on limited or no data. However, only the increase in Bayesian methods is statistically significant when evaluated separately. Turning to the subsample of companies in Table 7, we note a reduction reinforcement learning patents, likely reflecting the GDPR's impact on development costs for applications like recommendation systems and personalized advertising that rely on individual user interactions. Finally, we observe that companies responded by accelerating innovation in synthetic data, effectively seeking substitutes for increasingly costly personal data. Notably, these effects are statistically significant only when we focus on companies, which have become increasingly narrowly focused on data-intensive and capital-intensive deep learning methods (Klinger et al., 2020). This underscores the uneven impact of the GDPR, with pronounced effects on companies reliant on data-intensive methods.

4.4.3 Pre-trends

A significant limitation of our findings is their descriptive nature; we cannot discount the possibility of unobserved variables influencing AI patenting alongside changes in privacy regulation. But although our data do not support causal claims, we can still probe the robustness of the patterns documented above by examining whether our exposure variable captures trends that are obviously unrelated to the GDPR. To that end, we replace the post-2018 dummies with dummies marking previous years.

The results are presented in Table 10. Reassuringly, we find no pre-trends in 2012. However, for data-saving patent families, the coefficient turns positive and significant in

Table 6: Effect on all applicants by type of AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Data	Exploiting	Knowledge-based	Reinforcement	Synthetic	Transfer
	Bayesian	acquisition	structure	systems	learning	data	learning
Firm exposure × post-2018	0.004***	-0.001	0.006*	0.002	0.003	-0.000	-0.000*
	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.001)	(0.0002)
Observations	36,397	36,397	36,397	36,397	36,397	36,397	36,397
R-squared	0.949	0.962	0.943	0.953	0.891	0.951	0.782
Applicant FE	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the full sample. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

Table 7: Effect on firms by type of AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Data	Exploiting	Knowledge-based	Reinforcement	Synthetic	Transfer
	Bayesian	acquisition	structure	systems	learning	data	learning
Firm exposure × post-2018	0.012	-0.005	0.105	0.093	-0.028**	0.025**	-0.006
	(0.040)	(0.012)	(0.065)	(0.063)	(0.013)	(0.012)	(0.006)
Observations	9,107	9,107	9,107	9,107	9,107	9,107	9,107
R-squared	0.900	0.842	0.893	0.895	0.805	0.899	0.538
Firm FE	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the restricted sample focusing on companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

Table 8: Effect on firms by age

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	-0.043	-0.111	0.055
	(0.048)	(0.098)	(0.102)
GDPR exposure × post-2018 × age	0.004***	0.012***	-0.008**
	(0.001)	(0.004)	(0.004)
Observations	15,141	9,088	9,088
R-squared	0.792	0.897	0.888
Firm FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the restricted sample focusing on companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Company age is proxied by the difference between the first year the company is observed patenting in PATSTAT and the current year. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

Table 9: Firm size effects

	(1) All AI	(2) Data-saving	(3) Data-intensive
GDPR exposure × post-2018	0.034	0.115*	-0.081
	(0.034)	(0.060)	(0.068)
GDPR exposure × post-2018 × size	0.009***	0.045***	-0.052***
	(0.002)	(0.012)	(0.011)
Observations	15,141	9,088	9,088
R-squared	0.792	0.897	0.889
Firm FE	yes	yes	yes
Year FE	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the restricted sample focusing on companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2018 is a dummy variable equal to 1 for the years since 2018. Company size is proxied by the total stock of AI and non-AI patents held by the company. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

2014. This is perhaps not surprising, given that in 2014 the European Parliament demonstrated strong support for the GDPR by voting in plenary with 621 votes in favour, 10 against and 22 abstentions.[7] In other words, there was wide political consensus about the policy implementation, in turn implying that most stakeholders—including the private sector—where already informed about the upcoming regulation. Moreover, while the GDPR was enforced in 2018, it was adopted by the European Parliament in 2016. Still, firms only responded by reducing patenting in data-intensive technologies following the implementation of the GDPR in 2018. A possible explanation is that applicants proactively invested in data-saving innovations as a low-risk strategic hedge when the GDPR appeared likely, but only reduced data-intensive patent applications after implementation when necessary, reflecting their reluctance to abandon existing investments.

Table 10: Full sample results—alternative post-treatment periods.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
VARIABLES	All AI	Data saving	Deep learn	All AI	Data saving	Deep learn	All AI	Data saving	Deep learn
GDPR exposure × post-2012	0.001	0.003	-0.002						
	(0.001)	(0.003)	(0.003)						
GDPR exposure × post-2014				-0.001	0.004**	-0.003			
				(0.002)	(0.002)	(0.002)			
GDPR exposure × post-2016							-0.001	0.006**	-0.004
							(0.003)	(0.003)	(0.003)
Observations	53,753	36,397	36,397	53,753	36,397	36,397	53,753	36,397	36,397
R-squared	0.894	0.943	0.939	0.894	0.943	0.939	0.894	0.943	0.939
Applicant FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes

Notes: This table shows the correlation between the share of AI in patent family stocks and the share of each type of AI in AI family stocks in the restricted sample focusing on companies. GDPR exposure is based on the share of sales of Information and Communications products and services (ICT) sold to EU final consumers by company in a given country and 2-digit industry in the base year 2010. Post-2012/2014/2016 are dummy variables equal to 1 starting from the relative years. Standard errors are clustered at the company level. The coefficients with *** are significant at the 1% level, with ** are significant at the 5% level, and with * are significant at the 10% level.

[7] See https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en.

4.5 Conclusions

This paper introduces a novel AI taxonomy, applied to patent data, to analyze key trends in the data-intensity of AI development over time and space. We find that over the past decade, AI innovation has increasingly shifted from knowledge- or rules-based systems toward data-intensive deep learning. However, recent years—following the introduction of the EU's General Data Protection Regulation (GDPR) in 2018—have also seen growing interest in data-saving methods, such as transfer learning, synthetic data generation, and Bayesian approaches, possibly influenced by rising data costs and privacy concerns.

The potential influence of the regulatory environment on this trend is further highlighted by significant geographic disparities in AI patenting activity and technological specializations. Since the launch of China's "Made in China 2025" industrial policy initiative, for example, Chinese AI patenting activity has surged dramatically, particularly in the public sector, with universities and government institutions contributing 86% and 54% of global AI patents in these categories, respectively—far outpacing the 3% and 4% shares of their U.S. counterparts. This dominance is especially pronounced in data-intensive AI, supported in part by government procurement efforts that generate critical training data for commercial applications (Beraja et al., 2021; Beraja, Yang and Yuchtman, 2023). In contrast, the U.S. leads in data-saving AI innovation, with American companies responsible for 45% of global data-saving patent applications. Although the EU lags in overall AI patenting activity, it is also more active in data-saving approaches in relative terms: 13% of total private sector AI patents filed within the EU where data-saving compared to just 5% in China.

Thus, while all regions demonstrate higher patent activity in data-intensive AI, the relative balance between approaches varies substantially across jurisdictions. China shows the strongest skew toward data-intensive innovation, while U.S. applicants maintain a more balanced portfolio across both domains. European applicants, though also filing more data-intensive patents in absolute terms, show a comparatively stronger emphasis on data-saving technologies relative to other regions.

Building on these insights, we formally explore how privacy regulation has shaped the trajectory of AI innovation across regions. Our baseline estimates—including patents filed by firms, universities, government institutions, and individuals—show that GDPR-exposed applicants reduced data-intensive AI patents while increasing data-saving applications. The average effect is driven by EU-based applicants. This regulatory impact manifests heterogeneously across applicant types. Companies, particularly larger and older EU firms, drove the bulk of this shift: EU companies increased data-saving applications by over 30% while reducing deep learning patents by more than 20%. We conclude that the EU's emphasis on privacy protection—although reducing AI patenting overall—has fostered innovation in data-saving technologies. This regulatory-driven specialization suggests that policy choices not only affect the volume of innovation but fundamentally shape its direction.

Our findings carry important implications for both innovation policy and market competition. The GDPR appears to redirect technological development toward more privacy-preserving approaches, demonstrating how regulation can shape innovation trajectories. However, the stronger response among established firms suggests that privacy regulation may inadvertently reinforce incumbent advantages. Indeed, a substantial body of work has documented the adverse impacts of the GDPR for smaller companies and innovation more broadly, leading to rising levels of market concentration (Frey and Presidente, 2024; Peukert et al., 2022; Johnson et al., 2023). The forthcoming EU AI Act may exacerbate this trend by increasing compliance burdens on smaller firms, while potentially further shifting technological development toward less data-intensive approaches, as it emphasizes explainability—a challenge for deep learning technologies.[8] Investigating the impact of the EU AI Act on both the volume and direction of technological change in AI is a promising avenue for future research.

[8] On the EU AI Act, see L. Garicano, ”The Strange Kafka World of the EU AI Act: The Regulation Needs Repealing.” Silicon Continent, 30 Oct. 2024. https://www.siliconcontinent.com/p/the-strange-kafka-world-of-the-eu. Accessed on November 30, 2024.

References

Acemoglu, D. (1998), 'Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality', *The Quarterly Journal of Economics* **113**(4), 1055–1089.

Acemoglu, D. (2002), 'Directed Technical Change', *The Review of Economic Studies* **69**(4), 781–809.

Acemoglu, D. (2023*a*), 'Distorted Innovation: Does the Market Get the Direction of Technology Right?', *AEA Papers and Proceedings* **113**, 1–28.

Acemoglu, D. (2023*b*), 'Harms of AI', *The Oxford Handbook of AI Governance* p. 660–706.

Acemoglu, D., Aghion, P., Bursztyn, L. and Hemous, D. (2012), 'The Environment and Directed Technical Change', *American Economic Review* **102**(1), 131–66.

Acemoglu, D., Gancia, G. and Zilibotti, F. (2015), 'Offshoring and Directed Technical Change', *American Economic Journal: Macroeconomics* **7**(3), 84–122.

Acemoglu, D. and Johnson, S. (2023), *Power and progress: Our thousand-year struggle over technology and prosperity*, Hachette UK.

Acemoglu, D. and Restrepo, P. (2020), 'The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand', *Cambridge Journal of Regions, Economy and Society* **13**(1), 25–35.

Allen, R. C. (2009), *The British Industrial Revolution in Global Perspective*, Cambridge University Press.

Aridor, G., Che, Y.-K. and Salz, T. (2020), 'The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR', *The RAND Journal of Economics* **54**, 695–730.

Autor, D., Dorn, D., Katz, L. F., Patterson, C. and Van Reenen, J. (2020), 'The Fall of the Labor Share and the Rise of Superstar Firms', *The Quarterly Journal of Economics* **135**(2), 645–709.

Babina, T., Fedyk, A. and Hodson, J. (2020), 'Artificial Intelligence, Firm Growth, and Industry Concentration', *Journal of Financial Economics* **151**.

Beraja, M., Kao, A., Yang, D. Y. and Yuchtman, N. (2021), 'AI-tocracy', *The Quarterly Journal of Economics* **138**, 1349–1402.

Beraja, M., Kao, A., Yang, D. Y. and Yuchtman, N. (2023), Exporting the Surveillance State Via Trade in AI, Working Paper 31676, National Bureau of Economic Research.

Beraja, M., Yang, D. Y. and Yuchtman, N. (2023), 'Data-intensive Innovation and the State: Evidence from AI Firms in China', *The Review of Economic Studies* **90**(4), 1701–1723.

Biondi, F., Inferrera, S., Mertens, M. and Miranda, J. (2023), Declining business dynamism in europe: The role of shocks, market power, and technology, Technical report, Jena Economic Research Papers.

Bonney, K., Breaux, C., Buffington, C., Dinlersoz, E., Foster, L. S., Goldschlag, N., Haltiwanger, J. C., Kroff, Z. and Savage, K. (2024), Tracking firm use of ai in real time: A snapshot from the business trends and outlook survey, Working Paper 32319, National Bureau of Economic Research.

Campbell, J., Goldfarb, A. and Tucker, C. (2015), 'Privacy Regulation and Market Structure', *Journal of Economics & Management Strategy* **24**(1), 47–73.

Cockburn, I. M., Henderson, R. and Stern, S. (2019), The impact of artificial intelligence on innovation, *in* A. Agrawal, J. Gans and A. Goldfarb, eds, 'The Economics of Artificial Intelligence: An Agenda', University of Chicago Press.

Decker, R. A., Haltiwanger, J., Jarmin, R. S. and Miranda, J. (2016), 'Declining Business Dynamism: What We Know and the Way Forward', *American Economic Review* **106**(5), 203–207.

Edler, J., Blind, K., Kroll, H. and Schubert, T. (2023), 'Technology Sovereignty as an Emerging Frame for Innovation policy. Defining Rationales, Ends and Means', *Research Policy* **52**(6), 104765.

Edquist, C. (2013), *Systems of Innovation: Technologies, Institutions and Organizations*, Routledge.

Fagerberg, J. and Srholec, M. (2008), 'National Innovation Systems, Capabilities and Economic Development', *Research policy* **37**(9), 1417–1435.

Frey, C. B. (2019), *The Technology Trap*, Princeton University Press.

Frey, C. B. and Presidente, G. (2024), 'Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally', *Economic Inquiry* **62**, 1074–1089.

Giczy, A. V., Pairolero, N. A. and Toole, A. A. (2022), 'Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset', *The Journal of Technology Transfer* **47**(2), 476–505.

Goldberg, S., Johnson, G. and Shriver, S. (2024), 'Regulating Privacy Online: An Economic Evaluation of the GDPR', *American Economic Journal: Economic Policy* **16**(1), 325–358.

Goldfarb, A. and Tucker, C. (2012), 'Privacy and innovation', *Innovation policy and the economy* **12**(1), 65–90.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep learning*, MIT press.

Gutiérrez, G. and Philippon, T. (2019), The failure of free entry, Working Paper 26001, National Bureau of Economic Research.

Habakkuk, H. J. (1962), *American and British Technology in the 19th century*, Cambridge University Press.

Hall, B. H., Jaffe, A. and Trajtenberg, M. (2005), 'Market value and patent citations', *Rand Journal of Economics* pp. 16–38.

Hanlon, W. W. (2015), 'Necessity is the Mother of Invention: Input Supplies and Directed Technical Change', *Econometrica* **83**(1), 67–100.

Hassler, J., Krusell, P. and Olovsson, C. (2021), 'Directed Technical Change as a Response to Natural Resource Scarcity', *Journal of Political Economy* **129**(11), 3039–3072.

Hornbeck, R. and Naidu, S. (2014), 'When the Levee Breaks: Black Migration and Economic Development in the American South', *American Economic Review* **104**(3), 963–90.

Johnson, G. (forthcoming), "Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond, *in* A. Goldfarb and C. Tucker, eds, 'The Economics of Privacy'.

Johnson, G. A., Shriver, S. K. and Goldberg, S. G. (2023), 'Privacy and market concentration: intended and unintended consequences of the gdpr', *Management Science* **69**(10), 5695–5721.

Johnson, G., Shriver, S. and Goldberg, S. (2022), 'Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR', *Management Science,* **69**(10), 5695–5721.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. et al. (2021), ‘Highly accurate protein structure prediction with alphafold’, *Nature* **596**(7873), 583–589.

Kasy, M. (2022), The Political Economy of AI Regulation: Towards Democratic Control of the Means of Prediction, Discussion paper 16948, Institute of Labor Economics.

Klinger, J., Mateos-Garcia, J. and Stathoulopoulos, K. (2020), A Narrowing of AI Research?, Working paper, Available at SSRN.

Long, C. X. and Wang, J. (2016), ‘Evaluating Patent Promotion Policies in China: Consequences for Patent Quantity and Quality’, *Economic Impacts of Intellectual Property-Conditioned Government Incentives* pp. 235–257.

Marcus, G. F. (2018), ‘Deep Learning: A Critical Appraisal’, *ArXiv* **abs/1801.00631**.

Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (2013), *Machine Learning: An Artificial Intelligence Approach*, Springer Science & Business Media.

Miller, A. R. and Tucker, C. (2009), ‘Privacy protection and technology diffusion: The case of electronic medical records’, *Management Science* **55**(7), 1077–1093.

Miller, A. R. and Tucker, C. (2018), ‘Privacy Protection, Personalized Medicine, and Genetic Testing’, *Management Science* **64**(10), 4648–4668.

Miller, A. R. and Tucker, C. E. (2011), ‘Can Health Care Information Technology Save Babies?’, *Journal of Political Economy* **119**(2), 289–324.

Mitchell, M. (2019), *Artificial Intelligence: A Guide for Thinking Humans*, Penguin UK.

Nelson, R. R. (1993), *National Innovation Systems: A Comparative Analysis*, Oxford University Press.

Nelson, R. R. and Nelson, K. (2002), 'Technology, Institutions, and Innovation Systems', *Research policy* **31**(2), 265–272.

Peukert, C., Bechtold, S., Batikas, M. and Kretschmer, T. (2022), 'Regulatory Spillovers and Data Governance: Evidence from the GDPR', *Marketing Science* **41**(4), 746–768.

Presidente, G. (2023), 'Institutions, Holdup, and Automation', *Industrial and Corporate Change* **32**(4), 831–847.

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

Stiebale, J., Suedekum, J. and Woessner, N. (2020), 'Robots and the Rise of European Superstar Firms', *International Journal of Industrial Organization* **97**.

Tambe, P., Hitt, L., Rock, D. and Brynjolfsson, E. (2020), Digital Capital and Superstar Firms, Working paper 28285, National Bureau of Economic Research.

WIPO (2019), *Technology Trends 2019: Artificial Intelligence: Data collection method and clustering scheme: Background paper*, World Intellectual Property Organization, Geneva.

Wooldridge, M. (2020), *The Road to Conscious Machines: The Story of AI*, Penguin UK.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020), 'A Comprehensive Survey on Transfer Learning', *Proceedings of the IEEE* **109**(1), 43–76.