The Rise of China's Technological Power: the Perspective from Frontier Technologies*

Antonin Bergeaud Cyril Verluise

January 2023

Abstract

We develop a new method to identify patent in specific technologies and use it to study the contribution of the US, Europe, China, and Japan to frontier innovation. We find that China's contribution to frontier technology has become quantitatively similar to the US in the late 2010s while overcoming the European and Japanese contributions. Although China still exhibits stigmas of a catching up economy, these stigmas are on the downside. The quality of frontier innovation published at the Chinese Patent Office has leveled up to the quality of patents published at the European and Japanese patent offices.

JEL classification: O30, O31, O32, O43
Keywords: Frontier Technologies, China, Patent Landscaping, Machine Learning, Patents

*Addresses: Bergeaud: HEC Paris, CEP-LSE and CEPR; Verluise: Collège de France and PSE. We are grateful to Juliette Coly for her help during the first stage of the project. We are thankful to Philippe Aghion, Bronwyn Hall, Adam Jaffe, Gaétan de Rassenfosse, John Van Reenen and Abhijit Tagade for useful comments and help. We acknowledge support from Google Cloud Platform (GCP research credits programme grant).

1 Introduction

Modern growth theory (Romer, 1990; Aghion and Howitt, 1992) recognizes the crucial role of technological progress in long-term economic growth, but also emphasizes that the nature of technological progress depends on a country's level of development (Acemoglu, Aghion, and Zilibotti, 2006). Developing countries often catch up to more advanced economies by making incremental adjustments to adapt technologies previously developed by the most advanced economies. As they approach the technological frontier, growth increasingly relies on frontier innovation, which in turn requires institutional transformations such as competition policy (Zilibotti, 2017), research education (Krueger and Lindahl, 2001; Aghion et al., 2009; Goldin and Katz, 2010), external finance (Diallo and Koch, 2018; Rajan and Zingales, 1998), and improved management practices (Bloom and Van Reenen, 2007). Failure to implement these favorable institutions can hinder a country's ability to fully converge with advanced economies and can trap it in a "middle-income trap" (**worldbank2018**). A country's ability to produce, improve, and disseminate frontier innovation is essential for developing countries to join the ranks of developed economies and for developed countries to maintain their economic competitiveness.

Despite the central place that frontier innovation takes in growth theory, the empirical characterization of the diffusion of a specific novel technologies remain empirically tricky. Economists often rely on patent data to study innovation, but patents do not come with immediate ways to delineate a specific frontier technology, all the more when comparing different countries. As a result, the now well-documented dramatic increase in patenting in China (in 2019, the Chinese Patent Office filed 1.4 million patents, that is 43% percent of the world's total applications) is challenging to assess in terms of China's contribution to advancing the global technological frontier and the quality of these patents. By applying new statistical methods to the patent corpus, our work aims to address this issue.

From a methodological perspective, researchers might be tempted to use standard technological classes (e.g. CPC, IPC, USPC) assigned by patent offices to patents in order to identify patents contributing to a given technology and study its development. However, the economic literature has long cautioned against this approach. Griliches (1990) famously referred to this issue as the "patent classification problem". The classifications provided by patent offices are designed to facilitate the search for prior art and are therefore primarily based on techniques, which may not align with economists' understanding of technology. Additionally, a given technology is often characterized by a complex combination of different classes. For example, Schmookler (1966) notes that a subclass related to the dispensing of solids contained patents on both manure spreaders and toothpaste tubes.

Our first contribution is to introduce a new general methodological approach for accurately

and consistently retrieving a large set of patents related to specific technologies. We extend the algorithm of Abood and Feltenberger (2018), which uses machine learning operations to emulate human curation at scale, by adding a tractable amount of human supervision to improve the accuracy and consistency of our results. We apply this approach to six novel and representative technologies: additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage, and self-driving vehicles. These technologies were carefully chosen to cover a broad range of economic sectors and to ensure conceptual homogeneity. However, our new methodological approach can be easily extended to any technology using patent data and therefore used to track down the development of a specific field over time.

Our approach allows us to study the role and contribution of any country with a patent system in the development of these technologies. In this paper, we focus on four regions (United States, Europe, Japan, and China) which are at the epicenter of radical innovation production. Our second contribution is to use these six technologies to illustrate China's recent rise as a technological power. We find that, despite their differences in nature and maturity, these technologies provide a consistent and clear picture. The Chinese Patent Office's contribution to frontier innovation patenting has rapidly increased since the early 2000s, making China the second largest actor of frontier innovation and rapidly catching up with the US. Our results also suggest that China still exhibits stigmas of a former catching-up economy. In particular, Chinese's universities and firms account for a very small share of the basic research used in the development of the patents included in each of the six technologies. However, these stigmas seem to be vanishing. The quality and novelty of patents published at the Chinese Patent Office in these technologies has been rapidly increasing since the 2000s, reaching the level of patents published at the European and Japanese patent offices by the end of the 2010s. During this time, frontier technology patenting in China has also been increasingly supported by domestic patentees and has become more influential internationally, indicating the growth of domestic capabilities (Furman, Porter, and Stern, 2002).

Background and Literature Review

Our paper utilizes patent data to examine the evolution of specific technologies in diverse countries, with a focus on China, and to evaluate the quality of these contributions. As such, it is relevant to various strands of the literature on the subject.

First, there are a number of papers that measure the rise of China as a new technological power and explore potential explanations. This literature typically reports that China is an important player, if not the leader, in terms of many indicators of innovativeness. China

concentrated 159 unicorn companies in 2021 according to CB Insight for a total valuation of more than 500 billion dollars.[1] This is much more than Europe (89 for a valuation of 314 billion dollars, including respectively 32 and 138 billion dollars for the UK alone) and Japan (5 for a valuation of $6.8 billion dollars) but still far behind the US (405 for a valuation of $1,353 billion dollars). In terms of scientific effort, the importance of China has been rising for the past 20 years as measured by the number of top cited articles, which places the country as second scientific powerhouse behind the US.[2]

This catch up in terms of scientific publications is even more dramatic when it comes to Artificial Intelligence research (Baruffaldi et al., 2020). This tends to support the view that China has built the capacity to innovate in the technology of tomorrow, making its catch up more likely to be sustainable, and avoid the middle-income trap (Fan, 2014). This has been partly made possible by trade and foreign direct investments (Aghion et al., 2019; Hu and Jefferson, 2009), but also by subsidies and reforms of property rights (Dang and Motohashi, 2015).[3] However, other scholars argue that China still suffers from stigmas that penalize its capacity to innovate without the collaboration of other countries (Aghion et al., 2022). Abrami, Kirby, and McFarlan (2014) explain that while China does not lack the number of entrepreneurs, inventors or scientists, its institutions are not well-suited to encourage the development of frontier technologies. For example, every company larger than 50 employees is required to have a Chinese Communist Party (CCP) representative and a party liaison. In 2020, Xi Jinping, the general secretary of the CCP openly opposed the IPO of Ant Group, a large innovative financial company. Aghion, Dewatripont, and Stein (2008) and Murray and Stern (2007) show that academic freedom and functional Intellectual Property (IP) institutions are two critical requirements for the production of original research and which could question the capacity of China to compensate for their lack of freedom with mass investment in R&D.[4]

We contribute to this literature by looking as objectively as possible at the relative importance of China in the development and diffusion of recent frontier technologies that were chosen and identified without any preconceptions.

We also speak to a recent literature that exploits the patent corpus to study the diffusion of frontier technologies. These technologies are typically characterized by their radicalness,

[1] See CB Insights for a list.

[2] See the OECD Science, Technology and Industry Scoreboard 2017.

[3] These state subsidies and incentives to file patent applications have led experts to cast some doubt on the relevance and quality of the average Chinese patent (see e.g. He, 2021). In the empirical analysis, we take this possibility into account.

[4] Song, Storesletten, and Zilibotti (2011) and König et al. (2020) use structural estimations of dynamic heterogeneous firm models and report that R&D investment in China appears less productive than in other countries (namely Taiwan).

novelty, pervasiveness and their capacity to diffuse quickly and to have large impacts but are also highly uncertain and risky (Rotolo, Hicks, and Martin, 2015). Webb et al. (2018) look at the evolution in the number of patents filed in the US for a number of modern technologies such as Artificial Intelligence, Machine Learning, Semiconductor, Drones... They focus on the 1970-2015 period and find that most of these technologies have experienced a boom in the number of patents and inventors in the past decades mostly driven by US and Japanese multinationals. They also report a modest but growing contribution of Chinese inventors and firms to the rise of high tech patenting in the US. In a subsequent work, Bloom et al. (2021) also used patent data to study the diffusion of 29 disruptive technologies and their adoption by firms and labor markets in the US. Their findings suggest that there are long term impacts on the areas that hosted the initial development of these frontier technologies. These two studies focus on the US and attempt to have an overall view on the role and impact of high tech patenting. In contrast, Bessen and Hunt (2007) use patent data to analyze specifically the rise in software patenting in the US and compare the role of increased R&D spending and changes in IP legislation to explain this phenomenon. Other studies typically conducted by patent offices apply a combination of different methods to look at the development of patenting in a specific technology and a specific region.[5] For example IP Australia (2019) has analyzed the significant increase in patenting related to Machine Learning. We contribute to this literature by considering six technologies that cover various subjects and consider patents from the four main global technology contributors. This allows us to compare countries over time since the birth of these technologies.

Finally, we also contribute to a methodological literature which aims at delimiting technologies using patents. Historically, Trajtenberg (1990) tackled the classification problem by manually curating US patents belonging to the Computed Tomography Scanners technology. This method delivers precise results but is of course too labor-intensive to be extended to a larger corpus and multiple technologies. Other studies have used a number of rules combining keywords and Cooperative Patent Classification (CPC) classes to define a technology and constitute groups. This is the methodology applied for by Webb et al. (2018)[6] and by the patent landscaping literature. For example, the European Patent Office (EPO) has published a report on patenting in the field of automated vehicles (EPO, 2018). The rules used in these analysis are typically *ad hoc* and require a high level of expertise. Recently, Abood and Feltenberger (2018) introduced a new methodology that aims at circumventing this difficulty. Their approach, which we present in more details in Section 3.2, allows to emulate human-made technology classification using only a small number of representative patents as an input. Related approaches have leveraged Natural Lan-

[5]For a list of such study, see WIPO (2021).
[6]See Section 2.2 for more details on the selection process.

guage Processing and clustering algorithms to construct groups of patents (see Bergeaud, Potiron, and Raimbault, 2017 for a review). For example, the Fung Institute proposes an application of automatic labeling using machine learning to automated vehicles.[7] Similarly Giczy, Pairolero, and Toole (2021) have applied a slightly modified version of the Abood and Feltenberger (2018)'s algorithm to identify patents related to AI. These methodological works however do not attempt to measure and compare the diffusion of technologies across countries and time. We build on their method and adapt both the selection process of the imputed set of patents and the way the algorithm expands from this initial seed. Ultimately, our methodology combines a small amount of human work and automated landscaping to select patents related to a given technology with a high degree of precision and with no limitation in the geographical coverage and has been designed with the view of being easily extended to other technologies.

The remaining of this paper is organized as follows: Section 2 details our technology definition and selection procedure; Section 3 presents the automated patent landscaping approach and how we extend it; Section 4 evaluates the internal and external validity of the results generated by our algorithm on each of the six technologies; Section 5 documents the rise of China's technological power.

2 Technology definition and selection

The interpretation of the results we present in this paper are determined by two fundamental questions. First, what do economists mean by "technology"? Second, how to select a set of frontier technologies? We address these two key preliminary questions in this section.

2.1 Definition

Technology is a widely used term and can refer to many different concepts. In the economic and innovation literature, we classified its main usages into three categories which we refer to as "technique", "functional application" and "application field". A *technique* is a set of processes sharing a common methodological paradigm. Two distinct techniques can share a common goal. For example, TALENs, Zinc Fingers and CRISPR are all distinct techniques pursuing the same goal of editing the genome. A *functional application* is a high level goal which is directly targeted by one or several techniques in the course of their developments. Examples include computer vision and genome editing. The range of their

[7] See the webpage of the Fung Institute Capstone Project.

market applications can vary and usually exceed a single market. Eventually, an *application field* is an existing or newly created economic market which can leverage functional application to develop new or improve existing products. Examples of application fields include smartphones, nuclear power generation, etc...

In this paper, we work at the *functional application* level. This comes as a natural choice since we are interested in frontier innovation which has the potential to give advanced economies a significant growth momentum. Hence, our focus is on technologies which, like General Purpose Technologies, have the ability to infuse progress in a large range of applications. Function applications can be characterized by a set of tasks (for example, one of the task of autonomous vehicles is to enable cars to make autonomous decisions) which we will use to guide our selection procedure.

2.2 Selection

There are two main ways to define a set of technologies of interest: the supervised and unsupervised approaches. The most common approach, the "supervised", is based on human curation of technology-related documents. This is the approach followed by Webb et al. (2018) who define a list of technologies in the high-tech segment from prior knowledge. The second and more recent approach, the "unsupervised", combines text mining (specifically "topic modelling" techniques) and technology-related corpus to identify technologies (e.g. topics) without any use of prior knowledge. Such a method is implemented by Bloom et al. (2021) who use earnings conference call transcripts to uncover technologies which are the most frequently cited for their contribution to companies' momentum (see also Lenz and Winker, 2020 for an application to scientific fields).

Although extremely appealing, the unsupervised approach presents two limitations in our context. First, and most importantly, relying on past financial and corporate documents will invariably miss frontier technologies with still nascent market applications. Second, existing topic modeling techniques cannot guarantee that the identified "topics" (here technologies) are conceptually homogeneous. Without any supervision, selected technologies might (and will) include techniques, functional applications and application fields indifferently. While manual curation can address this issue to some extent, this however attractive approach is not suitable for our specific setting.

We opted for the supervised approach but designed a methodology to minimize our own biases and discipline the selection process. In particular, we sought to restrict to technologies that are considered as impactful and radical by many different institutions of different nature and geographical location. Do to so, we first screened a large number of reports and articles published at different time and dedicated to breakthrough technologies. These

articles have various sources: international institutions (OECD, 1998; 2016, EPO, 2020) , national agencies (Tarasova and Shparova, 2021, Kennedy, 2015), industry associations (BDI, 2011), experts (Review, 2021) and consulting companies (McKinsey, 2021; Deloitte, 2021). We took care to include sources from both developed and developing countries. From those documents, we listed without any *a priori* more than 30 technologies in a broad sense. Then we classified these items into the three aforementioned categories (technique, functional application and application field) and kept only those entering the "functional application" category. Eventually, we reviewed the remaining candidates (goals, recent breakthroughs, expected economic impact, and development stage) with two main objectives in mind: 1) only keep technologies that have already proven to have market applications or are expected to do so in the near future and 2) cover a large number of distinct application fields. From our initial list of technologies, we ended up with six frontier technologies: additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage and self-driving vehicles. See Appendix A for more details about how we selected the six technologies.[8]

2.3 Six different technologies

Before moving to the description of the automated patent landscaping methodology, we briefly discuss the characteristics of the six technologies considered in this article and why they constitute a relevant panorama of frontier technologies at the dawn of the 21^{st} century. A brief individual description and discussions about market potential are available in Appendix A.

Additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage, and self-driving vehicles are all technologies that are seen as having the potential to fundamentally disrupt our daily lives, are growing rapidly, and are receiving large investments. They are however at different stage of their development. Additive manufacturing, and computer vision are technologies that have been developed for decades with existing commercial applications. It is usually acknowledged that the first 3D-printing patents are filed in the first half of the 1980s[9] (Forsberg, 2020) and the history of computer vision starts with

[8]These technologies have been chosen among a large list that include other potential candidate (to name a few: natural language processing, vertical farming, cultured meat etc...). We do not claim that these six technologies alone are representative of the entirety of frontier technologies at the dawn of the 21^{th} century. Our goal is to illustrate the development of innovation in China and other countries using these examples that are constructed using the methodology presented in this paper. Hence, these technologies were chosen based on their prevalence and relevance to the study, and were not intended to be exhaustive. However, our methodology can be easily applied to other technologies.

[9]Although some sources consider previous patents to be related to 3D printing

the development of digital image scanner in the 1960s. Self-driving vehicles have been the subject of significant research at least since the 1970s, but the process of developing a fully autonomous commercial vehicle is not yet complete. Finally, hydrogen storage, genome editing and blockchain are more recent technologies, even if in some case, research started many years ago. Figure D1 in the Appendix shows the number of patent publications in each of these technologies each year (these patents have been selected with a methodology that we detail in the next section).

These technologies also differ in their development. While Additive manufacturing, computer vision and self-driving vehicles are the subject of massive investment by large industrial groups for several years, startups play a big role in pushing the blockchain technologies which is very recent and allows firms to scale-up without the need of massive investment in tangible capital. The development of genome editing technologies remains closely linked to university laboratories, with an important coordination effort (see e.g. Williams, 2013). Using a simple classifier based on the name of the assignee, we find that in 2019 about 12% of patents in genome editing are filed by a university of a public research institution. This number is below 5% in all other five technologies.[10]

Last but not least, these six technologies have applications (or potential applications) in a wide varieties of sectors. Additive manufacturing is already adopted in many different industrial sectors, blockchain has implication in data processing but also in finance, computer vision is an important brick of the development of AI systems, genome editing is mostly concentrated in the pharmaceutical sector, hydrogen storage in energy and self-driving vehicle in transport.

3 Automated patent landscaping with humans in the loop

In this section, we introduce automated patent landscaping, how it relates with existing approaches in economics, what are its limitations and how we address them.

3.1 The traditional approach

Determining the scope and boundaries of a technology using the patent corpus, or organizing patents into clusters, has been a longstanding challenge. A variety of methods have been attempted in order to address this issue. The three main tools that have been utilized

[10]This classifier is based on a simple model that assign patents in a category "academic institution" or not based on the name of the assignee. The model can be found here.

are technological classifications, citations, and keywords. While each of these tools can provide useful insights, they are also prone to introducing a significant amount of noise and variability into the analysis. In this section, we provide qualitative intuitions on these limitations. Section 4 will further quantify them. Technological classes are based on technical principles which are only partially related to the concept of technology we are looking for (functional application). Citations between patents have clear limitations in this case as well. Patent-to-patent citations are generated in order to define the scope of the technological monopoly granted to the patentees and to assess the validity of a patent over prior art. Proximity in the sense of functional application is then just one of the many reasons to generate a citation. Besides, the network of citations is very sparse and a large number of patents are never cited (Hall, Jaffe, and Trajtenberg, 2005). Finally, keywords can help identify patents dealing with a technology. However, language is highly variational: there are many ways to mention the same idea and at the same time a given word can have many different meanings. Hence, one can expect neither comprehensiveness nor accuracy from keywords alone. In this context, following Trajtenberg (1990), manual patent curation might appear to be the most accurate way to delineate a technology in the patent corpus.[11]

3.2 Automated patent landscaping

That is where the *automated* patent landscaping introduced recently by Abood and Feltenberger (2018) makes an important contribution. The authors develop a *semi-supervised* machine learning framework to emulate human-made technology classification. The algorithm only requires a small set of patents as input – the *seed* – which must be representative of the technology of interest. The algorithm then *expands* to "likely related" patents using both technological classes and citations (forward and backward). Specifically, it first expands to technological classes which are overrepresented in the seed and then expands twice on citations. Importantly, at this stage, we know that the resulting expansion set includes patents unrelated to the target technology or "false positives". The false positives are then *pruned* out using a classification model, based namely on the patent abstract, applied to the expansion set.

More precisely, the classification model is trained to distinguish between patents that belongs to the seed and a set of patents randomly drawn from the universe of patents, outside the expansion set (so-called *anti-seed*) and therefore "likely unrelated" to the target technology. This approach ultimately returns a group of patents in the target technology at virtually no cost, except for the curation of the seed patents. Importantly, no human

[11] Trajtenberg (1990) manually curated "computed tomography scanners" patents granted in the US to measure the value of citations.

intervention is needed to elaborate the set of rules determining whether a patent belongs or not to the target technology: semantic patterns are learned from the data.

The approach described in Abood and Feltenberger (2018) has already demonstrated a high level of potential, but it still exhibits certain limitations that are worth noting.

First, the pruning model is trained on "polar" cases while we would prefer to apply it to "intermediary" cases. The seed patents (positive examples) are selected to be at the "core" of the target technology. On the contrary, anti-seed patents (negative examples) are chosen from the complementary of the expansion set, hence potentially very far away from the target technology. For example, when trying to select patents related to the blockchain technology, the anti-seed might contain patents on drugs, car engines and semi-conductors. Hence, even if the algorithm performs well on the validation set,[12] it is not necessarily indicative of its performance when applied to patents in the expansion set, which may contain a significant proportion of "intermediary" examples. These examples may not be directly related to the target technology, but are still relatively close to it in terms of their characteristics or features. Training the model using a large majority of polar cases may therefore affect the overall validity of the classification model and the performance of the algorithm.

Second, the algorithm does not adequately consider the effect of variations in the data, such as the impact of changes to the seed data on the algorithm's output. The robustness of the algorithm, or its ability to produce consistent results despite variations in the input data, is an important factor to consider when evaluating the reliability of the results and the overall interpretation of the analysis. Robustness is a crucial aspect to consider when assessing the confidence we can place in the results and the conclusions that can be drawn from them.

3.3 A new extended approach

Our extended approach seek to address these two limitations. First, we *augment* the anti-seed with "harder" examples. These harder examples naturally arise from the human labeling of the seed patents that we performed for each technology. We start by inspecting existing attempts by the literature to landscape our technologies of interest using traditional methods. Formally, we use this literature and their reported selection rules (usually based on technological classes and/or keywords)[13] to generate a set of representative patents,

[12] The validation set is typically a 20-50% random split of the learning set (here, the seed and anti-seed patents) which is not used for training the model.

[13] See for example IP Australia (2018), Clarke, Jürgens, and Herrero-Solana (2020) and IIPRD (2017) for Blockchain. A full list of the sources we used is given in Appendix B.2.

keywords and CPC classes to be included in the seed for each technology. Sections B.3.1 to B.3.3 details these rules for each technology. From these rules, we randomly draw a set of potential candidate patents and manually and carefully label them as belonging to the technology or not from reading their titles and abstracts (see Table B1). Importantly, we keep the rejected patents as they provide "hard examples". Although they matched one or more rules used by previous attempts to landscape the technology, a human annotator have chosen to exclude them based on their abstracts. These are typically the "intermediary" examples we want our classification model to learn from and to be ultimately able to distinct from patents actually belonging to the target technology we are trying to delineate. We call this set of examples the *augmented anti-seed*. The model is ultimately trained using both the anti-seed *a la* Abood and Feltenberger (2018) and the augmented anti-seed to constitute the negative examples. See Appendix B for more details about how we construct the seed.

Second, we address the data variation question by implementing a series of robustness tests based on random variations in the seed. Specifically, we investigate how variations in the seed affect the expansion and the pruning outcomes. Formally, to test the robustness of the expansion, we draw random subsets from the seed, run the expansion using each of these subsets and compare the generated expansion sets. Next, we assess the pruning robustness by iterating over various random train-test splits of the annotated data. Various models are trained on varying sets of training data for each technology. Pruning robustness is ultimately evaluated by looking at models' agreement on a sample of out-of-training patents. Detailed results are reported in Section 4.

4 Algorithm deployment and validation

In this section we go through the main steps of the actual deployment of the algorithm. Next, we show that our results, in addition to being accurate and consistent, also exhibit patterns in line with technology experts' expectations.

4.1 Algorithm deployment

To begin with, it is important to note that contrary to Abood and Feltenberger (2018), we deploy the algorithm at patent family level rather than at patent publication level. A patent family is a collection of patent documents that are considered to cover a single invention in the sense that they share the same priority claims. Their technical contents are identical. Hence, considering only one document per family does not imply any loss of information

while significantly reducing the total number of items considered.[14] This seemingly minor twist has two important practical advantages. First, it enables us to consider all families with at least one publication having a known English abstract. That way, we ultimately cover more than 86% of all publications since 1970, while only 76% of patent publications do have a non-null abstract in our database. Detailed coverage is reported in Figure C1 in Appendix. Second, it minimizes the amount of texts to be classified at the pruning stage. Each family is processed only once, even if it includes more than one patent. This improves the overall computational tractability of the algorithm. Each individual patent then inherits from the characteristics of its family.

Construction of the seed Next, we delve into the algorithm deployment itself. As already discussed in Section 2, our work starts one step before the algorithm described by Abood and Feltenberger (2018). This first step consists in the definition of rules to identify a set of candidates. These candidates are picked out of patents which match at least one of the rules that we were able to find in the specialized literature. These rules include technological classes, keywords and patent similarity.[15] A random set of candidates are then labeled by humans based on the abstract and detailed annotation guidelines (see Table B1 in Appendix B). Annotation guidelines guarantee both transparency and replicability. In practice, we labeled candidates until at least 300 candidates are accepted which constitutes the technology *seed*. Importantly, rule-based candidates systematically included a large proportion of false positives, which were rejected. This set of rejects constituted the *augmented anti-seed*.[16]

Expansion Starting from the seed, the following step is the expansion. Regarding this step, we mainly follow to the Abood and Feltenberger (2018)'s procedure. We first expand to technological classes that were over-represented in the seed and then expand twice using citations (backward and forward). Note however that we had to adapt at the margin to take into account our choice to work at family level rather than publication level. In particular, we expressed citations in terms of the patent family rather than the usual publication format. For each family, we considered all citations received (forward) and sent (backward) by any patent in that family.

[14] There are around 120 million patent publications in the CLAIMS dataset versus 70 million patent families.

[15] There are instances in the specialized literature where specific patents are identified as being particularly representative or significant for a particular technology. In our study, we used the patents most similar to these key patents as defined in the Google Patents database and expand our dataset to include these related patents.

[16] These false positive would have been wrongly included in the set of patents delineating the target technology had a simple rule-based approach been used.

Pruning Finally, our pruning stage also differs from Abood and Feltenberger (2018) along 3 dimensions. First comes the composition of the training data. As already discussed, we add an augmented anti-seed to the seed and anti-seed described in their paper. Second, while our predecessors used not only text but also citations and technological classes as input to the classification model, we only restricted to text. In our view, both technological classes and citations imply potential pitfalls at this stage. Using technological classes in both the expansion and the classification model can generate pathological cases. Assuming that all technological classes in the seed are found important, then the anti-seed and the seed have no technological class in common which makes the classification task trivial. Regarding citations, by construction, patents in the second level of the citation expansion (L2) have no citations in common with the seed. Hence, considering citations in the classification task implies a systematic and uncontrolled bias against patents in the part of the expansion which we find undesirable. Third comes the model itself. We implement 3 different neural network architectures popular for text classification tasks: the multi-layer perceptron (MLP), the convolutional neural network (CNN) and a transformer, specifically a pre-trained Bert encoder. We provide an overview of these architectures in the following sub-section. The actual pruning is performed using the Transformer model which exhibits both the highest performance and consistency.

4.2 Performance and consistency

The most simple architecture we consider is the multi-layer perceptron (MLP). This architecture can be seen as a stack of logistic regressions and treats tokens or groups of tokens independently. Although it can be successful at identifying key phrases, it is unable to handle context and might eventually be seen as a sophisticated phrase matcher. We then turn to a second model and implement a Convolutional Neural Network (CNN). This architecture leverages the sequential nature of text through the use of feature maps (masks). These feature maps are there to detect sequences of tokens with a common and discriminant "meaning". CNN performances usually dominate those of MLP models thanks to this enriched understanding of language. However, they lack "memory" and cannot handle long context as feature maps typically focus on 3 to 5 token-long spans of text. Finally, we consider the Transformer architecture which was recently introduced (Vaswani et al., 2017) and has achieved spectacular results in many natural language processing (NLP) tasks, including text classification. Transformers rely on a core mechanism called *attention* which enables them to "understand" tokens in the context of neighboring tokens. Transformers are very large models trained at masked language completion on very large texts and eventually fine-tuned on specific tasks (e.g. text classification). This pre-training allows downstream users to start from a model that already embodies a large "understanding"

of language. A limited number of examples is then enough to adjust weights and achieve high performances on more specific tasks in specific contexts. This is especially well-suited when annotating examples is costly. The main drawback of using Transformers is their high computational costs.[17]

Performance We then train all these models. The task is a standard binary text classification. Specifically, we train and evaluate each model on ten distinct train-test sets for each technology. We implement this approach as a cross-validation method to have an estimate of the impact of random variations of the training set on both the performance of the model and its out of (training) sample predictions - later called *consistency*. Let us first focus on performance before moving to consistency later. We report the median *precision*, *recall* and *F1-score* for each technology and model architecture in Table 1. These metrics were all computed on the test set, that is, on examples not used to train the model. The precision is the share of texts that the model assigns to the seed and which are indeed part of it. The recall is the share of texts in the seed which were indeed predicted to be part of it. The F1-score is the arithmetic mean of the precision and recall. We observe that MLP and CNN architectures tend to exhibit similar F1-score. However, MLP models have higher precision and lower recall than CNN. This relates to the fundamental nature of MLP. As stated earlier, MLP can be seen as a sophisticated keyphrase matcher which usually has high precision but low recall. In any case, the transformer outperforms both of the models and achieves around 90% of median F1-score for all technologies except for self-driving vehicles (79%).[18] In the rest of the paper, we will use results from this latter model.

Table 1: Models performance

	MLP			CNN			TRF		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Additive Manufacturing	0.89	0.79	0.84	0.79	0.85	0.81	0.86	0.92	0.89
Blockchain	0.90	0.81	0.86	0.83	0.88	0.86	0.97	0.98	0.97
Computer Vision	0.89	0.81	0.85	0.86	0.87	0.87	0.87	0.95	0.90
Genome Editing	0.89	0.87	0.88	0.87	0.91	0.88	0.86	0.94	0.89
Hydrogen Storage	0.86	0.73	0.80	0.76	0.83	0.78	0.92	0.98	0.93
Self-driving Vehicle	0.79	0.65	0.71	0.69	0.73	0.71	0.75	0.85	0.79

Notes: Reported performance metrics were computed on the test set - unseen during training. Performance metrics are reported as follows: P for precision, R for recall and F1 for F1-score.

[17] Transformers are almost intractable using traditional Central Processing Unit (CPU) and require Graphics Processing Unit (GPU).

[18] This technology is indeed harder to classify even for humans. The very same technology can be used to automate driving or to assist human driving. In the former case, we would accept a patent while in the latter it would be rejected.

Comparison with rule-based approaches Using our candidate annotation exercise, we can compare those results with the performance that would have been obtained based on the rules used by existing attempts to landscape our six technologies of interest. Specifically, it enables us to obtain performance metrics for rule-based approaches using technological class, keywords and patent similarity. Although our approach does not enable us to compute all the performance metrics reported before, we can compute the precision of simpler approaches (for example using only a set of relevant keywords). We find that rule based candidate selection delivers both low and variable precision performances across technologies. Specifically, precision from CPC-class rule-based patent selection ranges from 0.01 (blockchain) to 0.34 (additive manufacturing). Precision from keyword rule-based selection goes from 0.09 (blockchain) to 0.89 (genome editing) for an average of 0.32. Precision from patent similarity ranges from 0.02 (additive manufacturing) to 0.57 (genome editing). All these metrics are reported in Table D1 in Appendix. It clearly appears that our approach to delineate technologies from the corpus of patents not only achieves good performance but also outperforms traditional rule-based methods. Hence, the set of patents selected using this new approach is both more precise and more complete than those of most existing attempts.

Consistency As already discussed, although performance *per se* matters, it is also crucial to understand how variations in the seed data can affect the results of the algorithm. We identify two channels. First, data variations can affect the expansion. The latter depends on the seed and has a critical role. It determines the set of documents which will be considered by the pruning model. Second, data variations can affect the pruning itself. The pruning model depends on the seed, the anti-seed and the augmented anti-seed and ultimately determines which documents in the expansion are to enter the technology or not. Robustness to random variations in the data is then crucial to ensure that algorithm results can be exploited rigorously. To investigate the consistency of the expansion, we generate random subsets of the seed. Specifically, we consider 3 different sizes: 90%, 70% and 50% of the initial seed and draw 10 subsets for each size. We then proceed to the full expansion starting from these distinct seeds and compute the pairwise family overlap of the generated expansion sets for each technology and seed size. Detailed results are reported in Table 2. We find that the average pairwise family overlap exceeds 89% in all cases. This remarkably high number indicates a high level of consistency for the expansion step and reassure regarding the relevant of the delimited technology.

Next, we looked at how the pruning stage is affected by variations in the training data. As discussed above, we trained the same architectures on 10 different train-test splits (of respective size 80%-20%) for each technology as a way to emulate natural variations in

Table 2: Median pairwise expansions overlap

	90%	70%	50%
Additive manufacturing	0.99	0.93	0.89
Blockchain	0.99	0.98	0.96
Computer vision	0.99	0.96	0.92
Genome editing	0.99	0.99	0.98
Hydrogen storage	0.99	0.97	0.95
Self-driving vehicle	0.99	0.97	0.95

Notes: For each size (90%, 70% and 50%), we drew 10 random subsets of the seed and proceeded to an expansion. For each pair, we computed the share of families in the two expansions. We report the median share of overlapping families across all expansion pairs.

the data. We then apply these models to a set of 10,000 out-of-training-sample documents randomly drawn from the expansion. For each technology, we then look at the standard deviation of the ten scores (each score ranging between 0 and 1) for each document and report its median in Table 3. We find that the standard deviation of the predicted scores is usually very low, most of the time below 0.05 which supports the consistency of the pruning step.

Table 3: Models robustness (Median dispersion in predicted scores)

	MLP	CNN	TRF
Additive manufacturing	0.029	0.082	0.017
Blockchain	0.008	0.047	0.003
Computer vision	0.015	0.029	0.010
Genome editing	0.003	0.001	0.004
Hydrogen storage	0.015	0.037	0.005
Self-driving vehicle	0.039	0.091	0.011

Notes: For each model architecture, we trained 10 models using distinct random subsets (80%) of the training set. Each model was then applied to a set of 10,000 texts (out of training set). We report the median standard deviation (at the sample level) of the predicted scores across models.

To summarize, our evaluation of the performance and consistency of the extended patent landscaping is very encouraging. In the next section, we take a first look at the set of patents that constitute each of the six technologies and consider the external validity of our approach.

4.3 External validation

We now use the output of the algorithm to investigate whether our results make sense. To do so, we first consider the top assignees and top inventors as reflected by the total number

of patents they hold.[19] We do it for each studied technology and then confront these results with prior insights from technology-specialized literature as well as background checks.[20] These lists of top assignees and inventors are reassuringly consistent with our priors and existing information. They also provide insights about the main actors of the different technologies considered. Finally, we also use the PatCit dataset (Rassenfosse and Verluise, 2020) and look at the top 3 most cited academic articles by patents in each technology.

4.3.1 Top 10 assignees by technology

Top panel of Table 4 reports the top 10 assignees for each technology by the number of patents they were granted worldwide.

A first observation is that most of the obvious players in each technology are present. For the sake of brevity, we focus on some remarkable high-ranked agents for each technology and explain why they were indeed expected. Starting with additive manufacturing, Xerox and Hewlett-Packard are two large companies that traditionally developed printers and which naturally moved to 3D printing technologies. In the field of blockchain, Alibaba, Intel, nChain and IBM are also in the top list of assignees in the expert-based landscaping of blockchain innovation proposed by Clarke, Jürgens, and Herrero-Solana (2020). The most prolific assignees in the field of Computer vision include firms that build and sell electronic devices, including cameras (Canon, Sony etc...). Interestingly, the top assignees in the field of genome editing are universities such as University of California Berkeley, Harvard University and University of Pennsylvania. As explained in Section 2.3, this technology as the characteristics of being very tightly connected to the academic world and breakthrough advances have been made in the laboratories of famous universities. Nevertheless, the list also reports large companies that develop chemistry and pharmaceutical products like Regeneron and Dupont. Overall, these findings are consistent with results from an overview of patenting in the genome editing technology field proposed by Benahmed-Miniuk et al. (2017). The field of hydrogen storage technologies is mostly dominated by car manufacturers. This naturally comes from the fact that the main usage of this technology is to propel vehicles using hydrogen. Finally, the field of self-driving cars also includes many traditional car manufacturers, including Toyota and Ford that communicate intensively on their progress in the development of autonomous vehicles. The list of top assignees also

[19] We used the harmonized name of assignees and inventors from the IFI CLAIMS dataset (available through Google Patent public data). This harmonization does not always guarantee that two different names of the same entities are actually merged in the same entity (e.g. Toyota Motor Co Ltd and Toyota Motor Corps).

[20] Note that, while the landscaping is done at the family level, analytical results are at the patent publication level.

includes automotive equipment suppliers such as Bosch and Denso Corp.[21]

On top of very large firms that spread over a large number of different technologies such as IBM and Samsung, we also note the presence of a number of firms that are much more specialized in a specific field. This is notably the case of Air Liquide for hydrogen storage, nChain for blockchain, ASML for additive manufacturing, Regeneron pharma for genome editing and Denso Corp for self-driving cars.

[21]Toyota, Ford and Bosch are mentioned as the top assignees in the field by WIPO (2019) (Chapter 3).

Table 4: Top 10 assignees and top 10 inventors

	Additive manufacturing	Blockchain	Computer vision	Genome editing	Hydrogen storage	Self driving vehicle
Top assignees						
1	Samsung Electronics Co Ltd	Alibaba Group Holding Ltd	Canon KK	Univ California	Toyota Motor Co Ltd	Toyota Motor Co Ltd
2	Hewlett Packard Development Co	IBM	Sony Corp	Pioneer Hi Bred Int	Honda Motor Co Ltd	Bosch Gmbh Robert
3	Xerox Corp	Qualcomm Inc	Samsung Electronics Co Ltd	Du Pont	Nissan Motor	Honda Motor Co Ltd
4	Asml Netherlands BV	Samsung Electronics Co Ltd	Koninkl Philips Electronics NV	Regeneron Pharma	Toyota Motor Corp	Nissan Motor
5	Gen Electric	LG Electronics Inc	Matsushita Electric Ind Co Ltd	Genentech Inc	Matsushita Electric Ind Co Ltd	Ford Global Tech LLC
6	Eastman Kodak Co	Sony Corp	Sharp KK	Monsanto Technology LLC	Sanyo Electric Co	Denso Corp
7	Canon KK	NChain Holdings Ltd	Seiko Epson Corp	Harvard College	Hyundai Motor Co Ltd	Toyota Motor Corp
8	Fujifilm Corp	Huawei Tech Co Ltd	Lg Electronics Inc	Hoffmann La Roche	Air Liquide	Hyundai Motor Co Ltd
9	Siemens AG	Intel Corp	Qualcomm Inc	Univ Pennsylvania	Panasonic Corp	Mitsubishi Electric Corp
10	IBM	Ericsson Telefon Ab L M	IBM	Centre Nat Rech Scient	GM Global Tech Operations Inc	Bayerische Motoren Werke AG
Top inventors						
1	Silverbrook Kia	Karczewicz Marta	Karczewicz Marta	Murphy Andrew J.	Ovshinsky Stanford R.	Tabata Atsushi
2	Lapstun Paul	Zhang Li	Zhang Li	Macdonald Lynn	Ukai Kunihiro	Shimizu Yasuo
3	Ng Hou T.	Zhang Kai	Nishi Takahiro	Mcswiggen James	Edlund David J.	Nordbruch Stefan
4	Vermeersch Joan	Wright Craig Steven	Kondo Tetsujiro	Zhang Feng	Fetcenko Michael A.	Hayakawa Yasuhisa
5	Van Damme Marc	Qiu Honglin	Wang Ye-kui	Rosen Craig A.	Taguchi Kiyoshi	Lynam Niall R.
6	Lewis Thomas E.	Yang Xinying	Chen Ying	Stevens Sean	Wakita Hidenobu	Watanabe Kazuya
7	Zhao Lihua	Wang Yue	Chen Jianle	Ruben Steven M	Maenishi Akira	Yasui Yoshiyuki
8	Patibandla Nag B.	Liu Hongbin	Yamazaki Shunpei	Wilson James M.	Young Kwo	Liu Jun
9	Ganapathiappan Sivapackia	Wang Zongyou	Kadono Shinya	Ni Jian	Nishio Koji	Breed David S.
10	Ye Jun	Fukushima Shigeru	Sugio Toshiyasu	Gurer Cagan	Reichman Benjamin	Matsuno Koji

Notes: Assignees and inventors are ranked based on the total number of patents for each technology over the whole corpus of patents. The harmonization of assignees' names is taken from the CLAIMS dataset.

4.3.2 Top 10 inventors

Moving from firms to people, the bottom panel of Table 4 reports the top 10 inventors for each technology by the number of patents they were granted worldwide.

As previously, for the sake of brevity we focus on the most emblematic and high-ranked inventors. We can note the presence of Marta Karczewicz in both Blockchain and Computer Vision. M. Karczewicz is a prolific inventor working at Qualcomm Technologies, Inc.. She is famous for having developed many technologies related to data compression which facilitates the transfer of important mass of information. The methods she developed are very central for many computer-related technologies such as computer vision and blockchain. As a recognition for her contributions, the EPO named her one of the three finalists for the award of European inventor of the year 2019.[22] Considering additive manufacturing, the most prolific inventor in the field is Kia Silverbrook. He is also a famous inventor who holds more than 9,000 patents worldwide.[23] K. Silverbrook founded Silverbrook Research, a company that developed digital printing and 3D printing technologies, among other inventions. In the field of genome editing, our top inventor is Andrew Murphy. He is the vice president in charge of research of Regeneron, a biotechnology company that develops different drugs and recently made important progress in new therapies using CRISPR (Gillmore et al., 2021). We also note the presence of Feng Zhang, a Professor at MIT and researcher at the Broad Institute. He is well known for his role in the development of optogenetics and CRISPR. He is also famous for his ongoing patent dispute with Chemistry Nobel Prize recipients J. Doudna and E. Charpentier over CRISPR-cas9 human application priority. Next, regarding hydrogen storage, Stanford R. Ovshinsky was a prolific inventor and engineer who contributed enormously to various fields, including energy science, and own hundreds of patents. In particular, he developed solid hydrogen storage technologies and founded the company Ovshinsky Innovation LLC at the end of his life to continue to explore alternative sources of power. Finally, in self-driving vehicle technology, Atsushi Tabata is an engineer at Toyota who published several articles related to the automation of driving controls.

4.3.3 Top academic publications

As a last exercise, we use the PatCit database (Verluise et al., 2020; Rassenfosse and Verluise, 2020) to look at the most cited academic papers by technology. PatCit is a tool that lists all citations from patents to research articles (also known as Non Patent Literature citations)

[22] See EPO (2019).
[23] See Wikipedia (2021).

that were used as a source. We report these articles along with the corresponding journal title in Table 5. To save space, we only report the top 3 for each technology but a longer list of doi is available in Table D2. As expected, the most cited articles, i.e. those that were the most pivotal in producing the ideas used in the development of the patents of each technology, are published in journal that are related to the technology. These journal can have a direct and clear link, for example, the International Journal of Hydrogen Energy is mentioned for hydrogen storage and the Proceedings Eighth IEEE International Conference on Computer Vision for computer vision.

However, the links may also seem less obvious, reflecting the complexity of externalities from academic research to the development of innovations. For example, the second most cited article for self-driving vehicle is a 1982 research that discusses CO2 concentration in the atmosphere. Since one of the goals of autonomous cars is to reduce the carbon footprint of transportation, this topic is often mentioned and discussed in the relevant patents.

Table 5: 3 most cited articles by technology

	Title	Journal	DOI
Additive Manufacturing			
1	Immersion lithography at 157 nm	Journal of Vacuum Science & Technology B	10.1116/1.1412895
2	Diaryliodonium Salts. A New Class of Photoinitiators for Cationic Polymerization	Macromolecules	10.1021/ma60060a028
3	Intelligent paper	Electronic Publishing, Artistic Imaging, and Digital Typography	10.1007/bfb0053286
Blockchain			
1	The design and implementation of a log-structured file system	ACM SIGOPS Operating Systems Review	10.1145/121133.121137
2	Scale and performance in a distributed file system	ACM SIGOPS Operating Systems Review	10.1145/37499.37500
3	A case for redundant arrays of inexpensive disks (RAID)	Proceedings of the 1988 ACM SIGMOD international conference on Management of data	10.1145/50202.50214
Computer Vision			
1	Overview of the H.264/AVC video coding standard	IEEE Transactions on Circuits and Systems for Video Technology	10.1109/tcsvt.2003.815165
2	Rapid object detection using a boosted cascade of simple features	Proceedings of the 2001 IEEE Computer Society Conference	10.1109/cvpr.2001.990517
3	Robust real-time face detection	Proceedings Eighth IEEE International Conference on Computer Vision	10.1109/iccv.2001.937709
Genome Editing			
1	Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells	Nature	10.1038/35078107
2	Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate	The EMBO Journal	10.1093/emboj/20.23.6877
3	RNA interference is mediated by 21- and 22-nucleotide RNAs	Genes & Development	10.1101/gad.862301
Hydrogen Storage			
1	Compact methanol reformer test for fuel-cell powered light-duty vehicles	Journal of Power Sources	10.1016/s0378-7753(97)02724-9
2	Steam reforming of natural gas with integrated hydrogen separation for hydrogen production	Chemical Engineering & Technology	10.1002/ceat.270100130
3	A safe, portable, hydrogen gas generator using aqueous borohydride solution and Ru catalyst	International Journal of Hydrogen Energy	10.1016/s0360-3199(00)00021-5
Self-driving Vehicles			
1	Adaptive Cruise Control System Aspects and Development Trends	SAE Technical Paper Series	10.4271/961010
2	Atmospheric CO2 Content in the Past Deduced from Ice-Core Analyses	Annals of Glaciology	10.3189/s0260305500002822
3	Advanced public transport information in Munich	International Conference on Public Transport Electronic Systems	10.1049/cp:19960454

Notes: Top 3 academic papers cited in the patents in each six technologies in the frontpage retrieved using PatCit (Rassenfosse and Verluise, 2020).

5 The rise of China

In this section, we examine the contributions of the United States, Europe, Japan, and China in terms of patents related to each of the six technologies that were selected using the procedure described earlier. The results we present are primarily descriptive in nature and do not aim to provide an explanation for the observed trends. However, these results do offer new insights into China's emergence as a technological power, as seen through the lens of these six illustrative technologies.

5.1 The bi-polarization of frontier innovation by the US and China

To measure the respective contribution of each region, we first count the number of utility patent filed by innovative actors in this region. An ideal way of doing so would require a way to assign patents using the address of the patentees. However, this would lead to a dramatic underestimation of the number of patents filed by Chinese inventors and assignees as these patents are less likely to be associated with an address in standard patent database (IFI CLAIMS, Patstat etc...).[24] To address this issue, we will focus on the earliest publications of a given patent family at the US (US), European (EP)[25], Japanese (JP) and Chinese (CN) patent offices. As a result, we are assuming that companies typically file their initial patent application in their home country before potentially filing additional applications in other countries. This assumption is based on the idea that firms generally prioritize protecting their intellectual property in their domestic market before seeking protection in other markets.[26]

For each technology , we start the analysis from the first year for which we could find at least 500 published priority patents (1989 for additive manufacturing, 1998 for blockchain, 1974 for computer vision, 1983 for genome editing, 1992 for hydrogen storage and 1974 for self-driving vehicle) and report the share of the four patent offices in the patent publication count for each technology in Figure 1. Of course, simply counting the number of patent applications ignore the well-known fact that all patents are not created equal and should be qualified by some measure to weight their quality. Nevertheless, Figure 1 already strik-

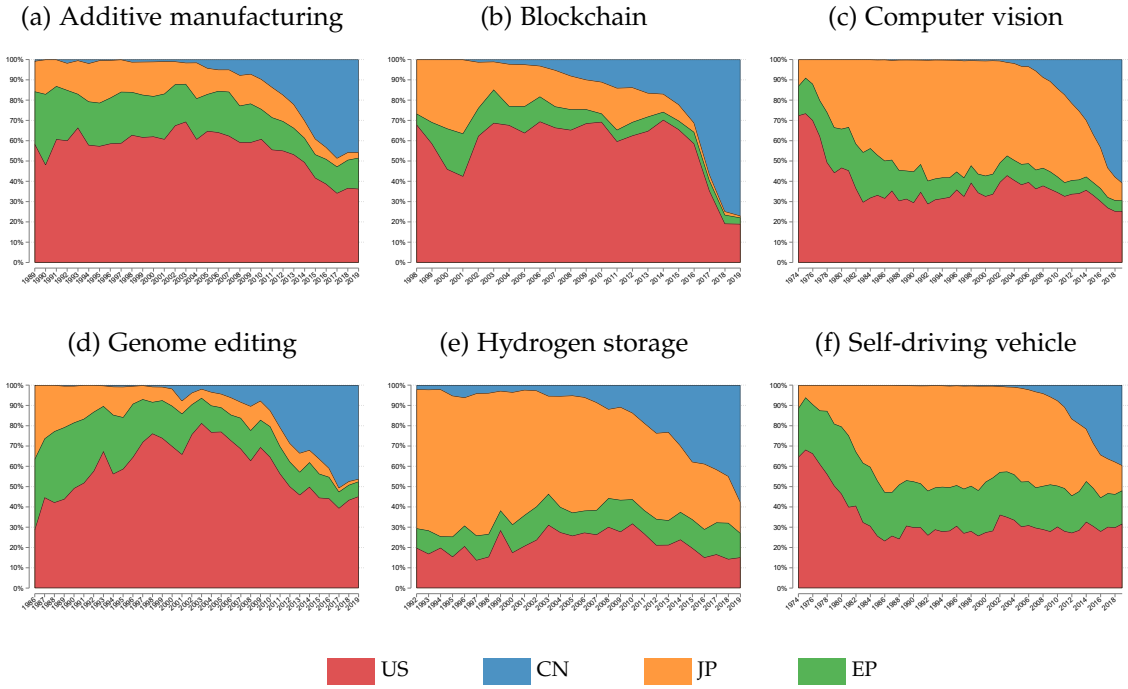
[24] In IFI CLAIMS, in 2021, 99.99% of USPTO patents are associated with a an address for the inventors while in the CNIPA inventors are not geolocated prior to 2009.

[25] We include individual national patent offices from all EU countries to which we add the British, Swiss and Norwegian patent offices on top of patents filed at the EPO under the label "EP". Since we only keep the oldest publication in a given patent family, we do not risk double counting patents.

[26] It is reasonable to assume that firms will first seek to protect their intellectual property in their domestic market before seeking protection in other countries. Looking at USPTO data between 2000 and 2019 where we can observe the country of the assignees and inventors, we find that a little more than 95.5% of patent families in which at least one assignee is located in the US first file at the USPTO.

ingly show the generalized growth of the share of Chinese patents across all technologies considered from the early 2000s. While it used to be almost insignificant in the early 2000s, at the end of the 2010s, the Chinese office represents at least a third of patent publications for all the frontier technologies considered. This share even exceeds 70% in the case of blockchain and 50% for computer vision.

Figure 1: Relative contribution to frontier technologies



Notes: Patent counts in the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP) as a share of the total patent count for each technology. The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

It should also be noted that this relative growth in China's technological power is taking place against a backdrop of markedly heterogeneous trajectories within other regions. In particular, the share of Japanese patents collapsed to a very low level at the end of the period in favor of Chinese patents. That Japan lost its position as a central hub for both production and innovation in Asia since the 2000s is a well documented fact (Criscuolo and Timmis, 2018; Ito et al., 2019), and it seems to be particularly striking for these six frontier technologies. Meanwhile, the United States has maintained a relatively high level of activity in all of the technologies while Europe holds a significant share of patents in self-driving vehicle and hydrogen storage but at the same time is almost nonexistent in blockchain and computer vision.

Looking in more details about the dynamics in Europe, we consider individual countries in Figure D2. Overall, Germany holds most of the patents in all technologies with more than 50% of European priority fillings. Two exceptions are worth noticing: Blockchain and Genome editing where the UK is dominating at the end of the period. This finding is not

surprising as Germany is the most important manufacturing hub in Europe. On the other hand, the UK is one of the world leader for blockchain technologies and home of leading research universities that are important drivers of innovation in genome editing. France accounts for about 10% of patents in hydrogen storage, computer vision and self-driving vehicles and other European countries share amount to about 10% to 15% of patents in all technologies.

Overall, the frontier technology landscape, which was once led by the US, Japan, and Europe, has become more polarized between the U.S. and China.

5.2 The Chinese catch up in quality

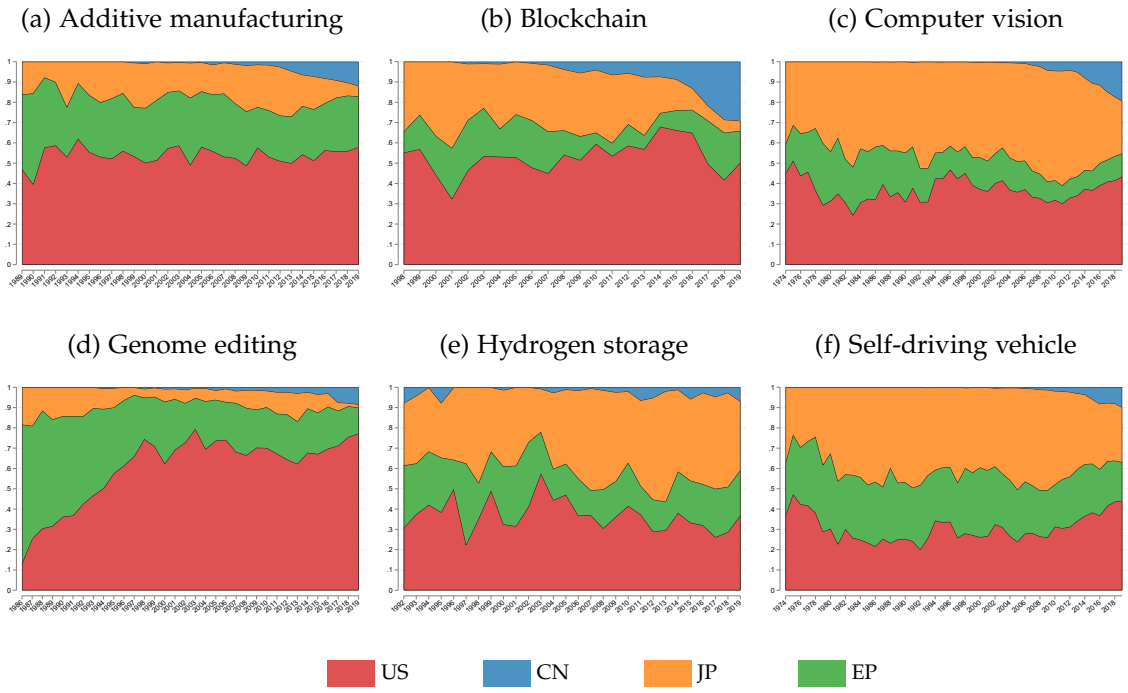
Next, we examine the quality of patents published in the four aforementioned patent offices. As previously discussed, measuring contributions to frontier innovation using only the raw number of patents published in frontier technologies can be misleading. Not all patents are created equal. In particular, there is evidence to suggest that the volume of patents in China, in particular, may not accurately reflect technological advancement due to concerns about the quality of these patents As discussed by He (2021), patent applications in China reflect diverse incentives which have sometimes little to do with invention. These incentives include government subsidy or job promotion, reputation building for individuals or universities and institutions, or acquiring certification as national high-tech enterprises. In this context, Hudson (2021)further emphasized the limitations of using patent counts as a reliable indicator of technological leadership, citing the case of 5G standards for broadcast cellular networks.

To account for this, we first filter patent families and keep only these that have at least one patent publication, within a given technology, in two of the four main patent offices considered (CNIPA, USPTO, JPO and EPO). By limiting the analysis to patents that have been published in multiple offices, we aim to exclude patents that have remained purely domestic and thus may not accurately reflect technological advancement. This approach has two main effects. Firstly, it allows to restrict to patent families having a minimum level of quality.[27] Second, it increases the comparability of patent count as it requires the family to have at least one patent application accepted in an other patent office. Results are presented in Figure 2.

With this restriction, the share of China is clearly less predominant at the end of the time

[27]The literature has established a link between the geographical coverage of a patent family and its quality, see e.g. Squicciarini, Dernis, and Criscuolo (2013). Restricting to family with a patent in each four patent offices will result in a noisy picture, yet with similar trends, due to the small number of observations for some technologies.

Figure 2: Relative contribution to frontier technologies - restricting on international applications



Notes: Patent counts in the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP) as a share of the total patent count for each technology. Restriction on patent family with at least one publication in two of the main patent offices (USPTO, CNIPA, EPO and JPO). The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

period. China now holds less than 30% of all patents and nearly none in hydrogen storage. Conversely, the relative importance of the US is now much larger and approaches 50% in all technologies. However, while the levels are different, the trends remain similar. In particular, China is growing since the 2000s relatively faster than other countries, and this is especially striking in the case of blockchain and computer vision.

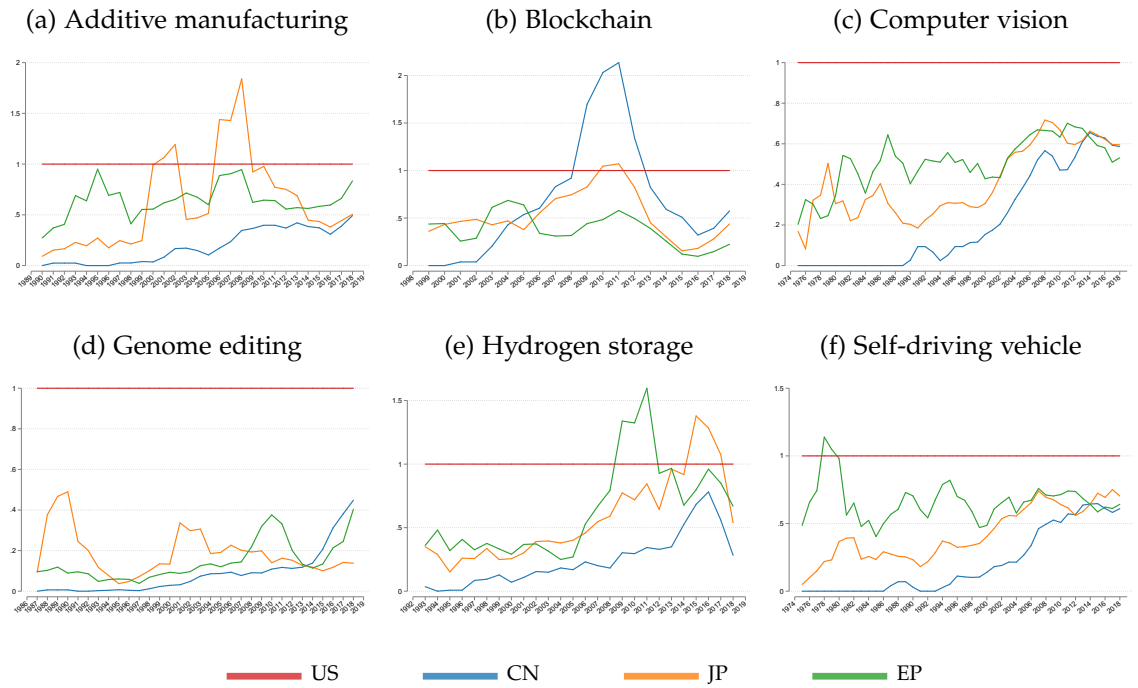
Do the previous results suggest that China is catching-up in terms of creating breakthrough innovations in these different technologies? We investigate this by looking at a common measure of patent quality, the number of citations received by patents.[28] Comparing different countries using citation counts is not straightforward because the propensity to cite or to be cited is highly dependent on the intellectual property offices' specific rules and customs. In addition, the home bias, i.e. the propensity to cite more naturally patents from the same patent office, mechanically increases the number of citations as the number of domestic patents increases. We however replicates the exercise of Figure 1 but weighting the number of patents by the number of citations received from foreign patent offices. The

[28]Note however that the very notion of patent quality is multi-faceted. The various measures used to apprehend patent quality can be inconsistent as evidenced by Higham, De Rassenfosse, and Jaffe (2021).

results are presented in Appendix D, Figure D3 and are consistent with that of Figure 2, suggesting that China is indeed catching-up in terms of quality.

A natural way to abstract from the home bias is to use one common origin for patent citations. We do this using citations received from Patent Cooperation Treaty (PCT) applications. PCT applications are international application that provides a common procedure to file a patent applications in all member states (which include more than 150 countries). In this procedure, an International Searching Authority will be in charge of searching for prior art which limits the risk of home bias. In addition, we chose to focus on the upper tail of the distribution of citations since the most cited patents are also those which are expected to have the largest impact. More precisely, we consider the average number of citations received by the top 10% most cited patents in each technology, year and country. The results are presented in Figure 3. In order to account for the fact that the average number of citations is not stationary, we report each number standardized by the US corresponding value. We can see that China is clearly exhibiting an upward trend in terms of the average citations received by its top patents and has clearly caught up with Europe and Japan in all technologies, and in some cases is very close to the US.

Figure 3: Average citations received from PCT applications by top 10% most cited patents



Notes: Average citations received by the top 10% most cited patents each year and in each technology and country from PCT applications. Level relative to the US. Top 10% patents are selected in the distribution of patents with at least once citation. Each series has been smoothed using 3 year rolling window centered around current year.

Although we have only analyzed citations from PCT applications, it is still possible that this measure may be influenced by the high number of patents filed by the CNIPA or that PCT applications may have a tendency to cite patents from different offices differently. To

address this, we introduce an alternative measure of a patent's contribution to its technology, which we refer to as radicalness. To do so, we follow the idea developed by Kelly et al. (2021) who utilize the semantic content of a patent to define a quality index based on how unique the patent is compared to its predecessors, but closely related to its successors. Kelly et al. (2021) applied this method to USPTO patents, which have full text available, and demonstrated that this quality index effectively captures a patent's technological value in that these patents are both novel and impactful. We have made several modifications to adapt this method to our needs. First, since the full text of patents is not usually available for documents outside the USPTO, we instead use the embedding representation of patent publications provided by Google Patent (see Srebrovic, 2019 for more details). Second, we calculate our measure of radicalness, within each of our six technologies instead of comparing a patent with the universe of other publications. That is, for each technology, we assign to each patent a measure between 0 and 1 qualifying its contribution to the field. More details are provided in Appendix C.2.

Figure 4 presents the relative share of the yearly number of patents, weighted by our measure of radicalness, from 2000 to 2014.[29] Our results support the observation that China is closing the gap with the US, which remains the technological leader, in terms of its contribution to these six technologies. Meanwhile, Europe has a relatively minor impact, particularly in the areas of computer vision and blockchain. Japan continues to make significant contributions to hydrogen storage and self-driving vehicles.

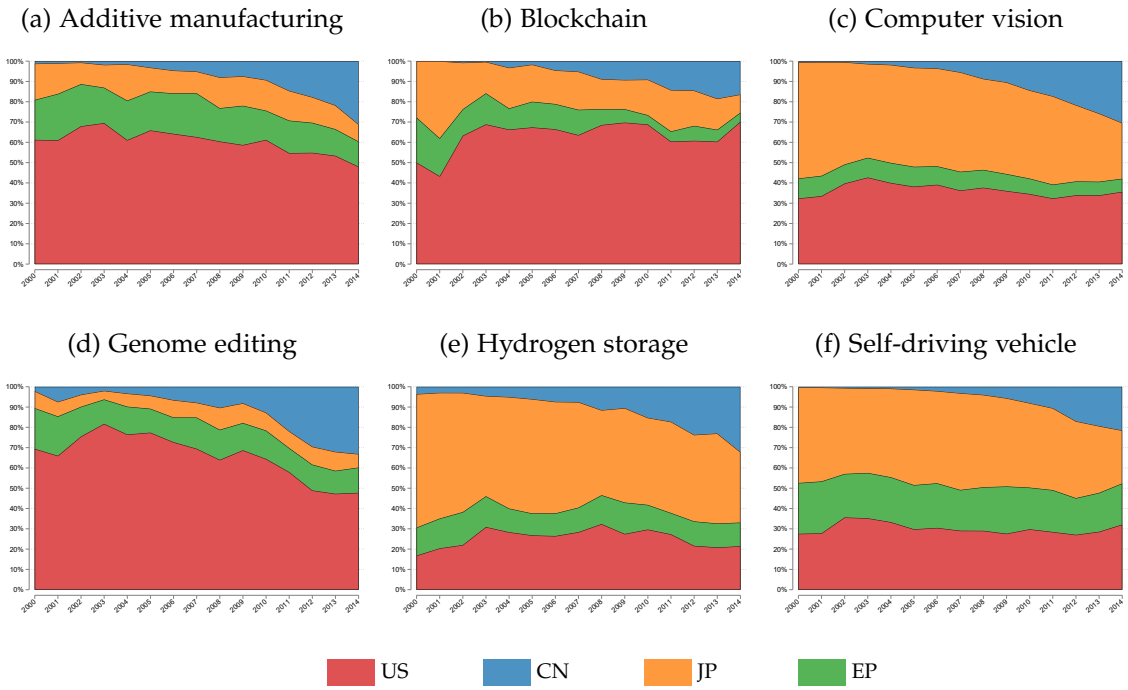
5.3 China's frontier innovation domestic capacities build up

In light of the previous results, one natural question is whether China can continue to catch-up with the US and push the technological frontier further. To explore this possibility, we examine the extent to which the increase in patenting (adjusted for quality) at the Chinese patent office reflects the actual development of domestic innovation capabilities.

To investigate this, we first look at the origin of the priority applications in each family of patents. To the extent that the office of priority filing, that is the office where the first patent of a family was filed, is a good proxy for the country of residence of the inventor or the assignee, then by examining the number of Chinese patents that claim priority to a Chinese patent versus a patent filed in another country, we can gauge the growth of local frontier innovation in China. This is important because domestic innovation is necessary for a country to transition out of the middle-income trap. If most Chinese patents are

[29] We start in 2000 and stop in 2014 due to the need to use 5 year window to calculate the index of radicalness and to keep enough data to calculate a reference point (see Appendix C.2).

Figure 4: Relative contribution to frontier technologies - weighting by radicalness



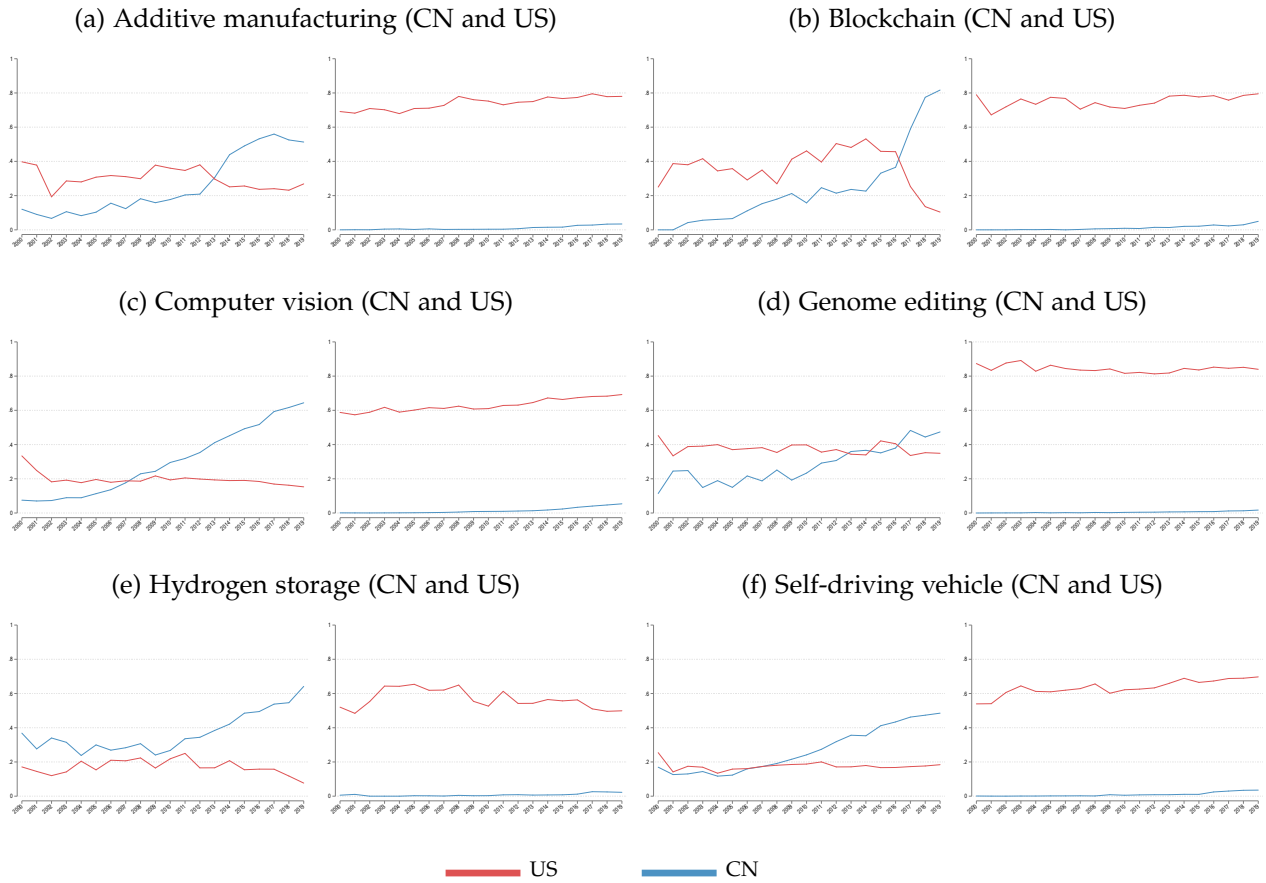
Notes: Patent counts in the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP) as a share of the total patent count for each technology. Each patent is weighted by a measure of its radicalness as defined in Appendix C.2. The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

subsequent applications of inventions originally protected in the United States, it could indicate that US firms are interested in protecting their products or processes in China and that the Chinese market is becoming more attractive to foreign technology owners. Hence, an increase in the share of Chinese patents that claim priority to the Chinese patent office would suggest that the development of these products and processes is increasingly coming from domestic innovation efforts.

In Figure 5, we plot the share of patents filed at CNIPA respectively with a priority filing also at CNIPA and at the USPTO (left-hand side panels). We see that the share of domestic priority filing is increasing since 2000 when priority filings were mainly coming from the US (and also from Japan, see Figure D4 for a more complete picture). By 2019, the majority of Chinese patents were claiming priority to domestic applications. For comparison, we conducted the same analysis for the USPTO (right-hand side panels and Figure D5) and found that the USPTO tends to have a consistently high proportion of US priority filings, typically hovering around 80%.

These findings suggest that China is strengthening its innovative capacity, but at the same time does not seem to account for a large share of USPTO priority claims. To further examine the global influence of Chinese patents, particularly in the US, we analyze the origin of citations received by CNIPA patents in six different technologies. Figure 6 shows

Figure 5: Origin of priority filings, US and China



Notes: Share of patents with a priority filing in China and in the US respectively filed at the CNIPA (left-hand side panel in each subfigure) and at the USPTO (right-hand side panel in each subfigure). Time period 2000-2019.

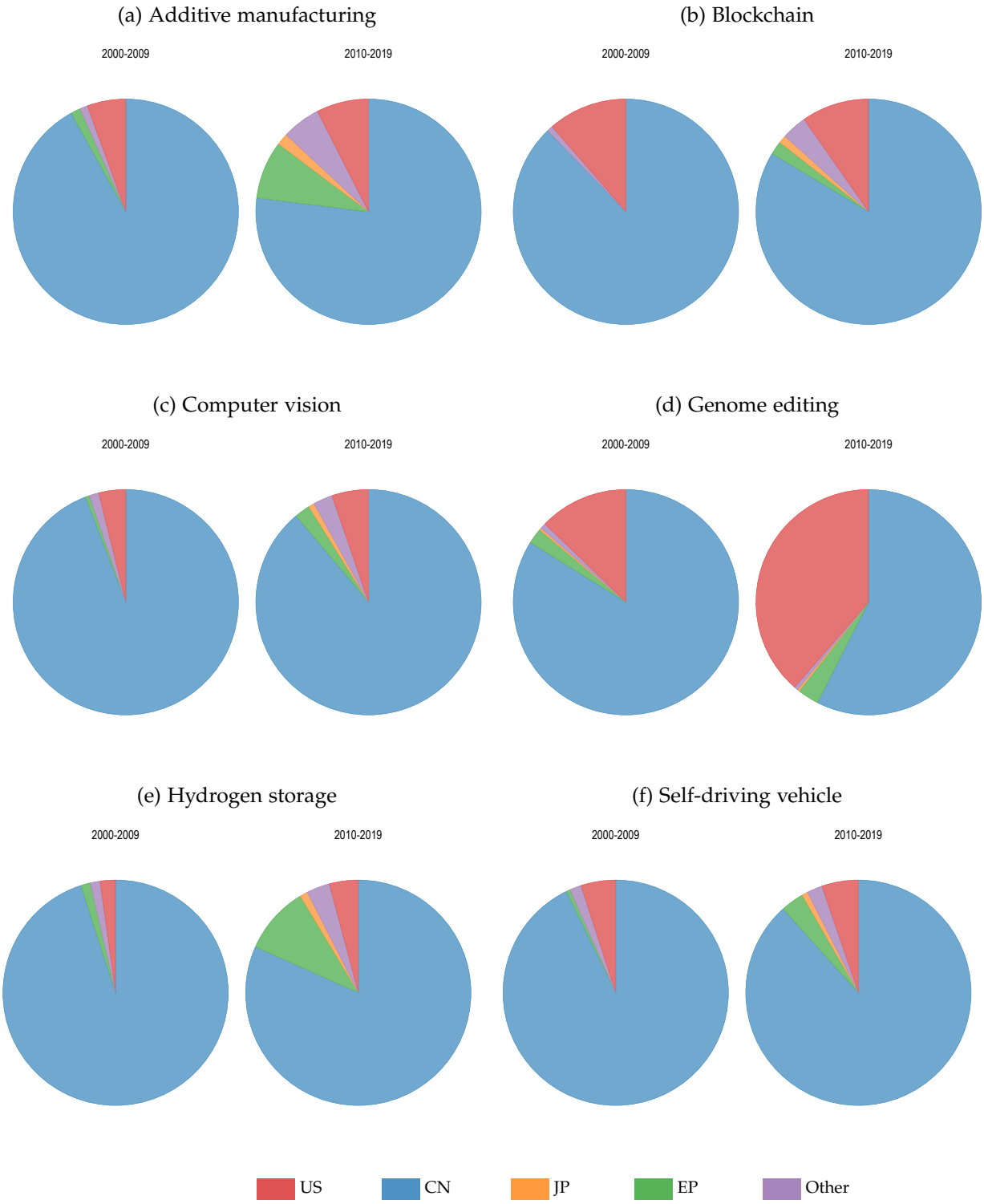
the distribution for two subperiods (2000-2009 and 2010-2019) and reveals that the influence of Chinese patents, as measured by the number of forward citations received, is more internationalized in the recent period than it was in the early 2000s. In particular, over one third of the citations received by CNIPA in the genome editing technology are from foreign patent offices, primarily the USPTO.

To complete our investigation on the building of domestic capabilities, we analyze the origin of the knowledge incorporated in the patents of our six technologies. To do this, we combined PatCit (Rassenfosse and Verluise, 2020) with the Elsevier Scopus API. This allowed us to track the affiliation history of all the authors of the academic articles referenced in a given patent.[30] We apply this approach to a random sample of patents in the six technologies under consideration.[31] The results by technology are reported in Table 6.

[30]Scopus does not allow us to select the affiliation at a specific time, so it is not possible to determine with certainty which laboratory a given researcher was affiliated with when a paper was published. As a result, we will use both the full affiliation history and the most recent affiliation as imperfect proxies in the following analysis.

[31]Scopus limits the number of requests that can be made each month so we could only use a small sample

Figure 6: Origin of citations received by Chinese patents



Notes: Distribution of the origin of citations received by priority filing of patents filed at the CNIPA by patent offices. EP includes all patents filed at the EPO and at European national patent offices. The number of citations has been standardized by the total number of citations made by all patents in each corresponding patent offices. Average over two time periods: 2000-2009 and 2010-2019. Citations from PCT applications are excluded.

of the total number of patents. Currently, we were able to retrieve 8854 articles published by 25,258 distinct

Specifically, this table shows the proportion of all researchers cited by these patents who are affiliated with a laboratory located in the US, Japan, China, or a European country. We defined the location either by considering the current affiliation or by considering all affiliations in the researcher's history with equal weight. In both cases, we found that China only represents a small share of the production of the basic knowledge used in the development of patents in each of the six technologies. The US accounts for the largest share, but European laboratories also have a significant presence with nearly 30% of all researchers.

Table 6: Share of academic paper by country and technology

	Current Affiliation				Affiliation History			
Technology	USA	Japan	China	Europe	USA	Japan	China	Europe
Additive Manufacturing	51.1%	5.7%	2.7%	28.0%	44.5%	6.2%	3.5%	31.3%
Blockchain	53.7%	4.5%	3.9%	22.7%	46.9%	4.9%	4.1%	26.9%
Computer Vision	53.6%	5.3%	2.5%	26.5%	45.7%	6.0%	3.0%	30.9%
Genome Editing	57.3%	4.8%	1.3%	29.3%	45.1%	5.3%	1.7%	35.9 %
Hydrogen Storage	34.9%	11.6%	6.3%	29.4%	32.3%	10.7%	6.3%	32.2%
Self Driving Vehicle	49.0%	6.1%	1.7%	28.2%	42.2%	6.72%	2.1%	31.9%

Notes: This table reports the share of each of the four regions in the affiliation of researchers whose article has been cited by patents in each technology. The location of the affiliation is either taken as the most recent affiliation or by considering all affiliations with equal weight for each researcher. Results are based on a sample of 8854 articles randomly selected from the list of non-patent literature citations extracted using Patcit (Rassenfosse and Verluise, 2020). Affiliation has been retrieved using Elsevier Scopus API.

Overall, these findings consistently point to the same conclusion: while China has made significant progress in terms of patent quality and has to some extent developed domestic capabilities, its influence in the development and diffusion of the six technologies considered remains limited, particularly when considering the origin of the academic research that drives breakthrough innovation. While it is true that China has made significant strides in recent years, our analysis suggests that the US continue to play a leading role in driving technological advancement in these areas.

6 Conclusion

In this paper, we have extended the automated patent landscaping approach from Abood and Feltenberger (2018) to accurately and consistently delineate a group of frontier technologies from the worldwide corpus of patents. Our first contribution is to show that this methodology, which can be easily applied to any technology, delivers consistent and precise results. We then used these six representative technologies to investigate the contribution of the United States, Europe, Japan, and China to the production of frontier innovation.

authors from a sample of about 18,000 patents.

Based on the evidence presented, we clearly see that China's technological strength has been increasing since the late 2000s, accounting for a significant share of patents. As a result, the technology landscape that was dominated by the United States, Europe, and Japan in the early 2000s is now much more polarized by U.S. and Chinese offices.

Digging deeper, we observed that the patents published by the Chinese patent office used to be of lower quality than their European, Japanese and American counterparts. However, the gap is closing and, at the same time, China is building up its domestic capabilities.

So can China continue to make a significant contribution to the technology frontier and catch up with the United States? In light of this, the answer seems to be yes. However, three important points should be made. First, the Japanese example shows that the innovation capacities of a country can never be taken for granted. What will happen to China's innovative power in the next decades is out of the scope of this paper but China's spectacular take-off since the 2000s does not necessarily foreshadow the next decades. Second, we have shown that China is increasingly contributing to frontier technologies which were pioneered before China's technological take-off. This leaves the question of China's ability to pioneer a new frontier technology untouched. Third, the question of China's ability to further adapt its institutions, especially in terms of research teaching and academic freedom, to contribute even more along the whole knowledge chain remains open. Recent results from Aghion et al. (2022) indeed suggest that academic research done in China continues to be too dependent of the US, consistently with our own findings.

Our approach delivers consistent insights to study innovation through the lens of patents. Most importantly, they open at least two important avenues for further research. First, delving into the characteristics (assignees, inventors, patentees locations, etc) of frontier technology patents filed in China and other developing countries appears to be a promising way to better assess the role of the various technology diffusion channels. Second, another promising avenue would be to delve into business dynamics (entry and exit) which take place within technologies themselves and might well have sound implications for the rest of the economy, including the fall of the labor share in the US and the rise of superstar giant innovators, as discussed by Autor et al. (2020).

References

Abood, Aaron, and Dave Feltenberger. 2018. "Automated patent landscaping." *Artificial Intelligence and Law* 26 (2): 103–125.

Abrami, Regina M, William C Kirby, and F Warren McFarlan. 2014. "Why China can't innovate." *Harvard business review* 92 (3): 107–111.

Acemoglu, Daron, Philippe Aghion, and Fabrizio Zilibotti. 2006. "Distance to frontier, selection, and economic growth." *Journal of the European Economic association* 4 (1): 37–74.

Aghion, Philippe, Celine Antonin, David Stromberg, and Xueping Sun. 2022. "Is Chinese innovation dependent on the US?" Manuscript, London School of Economics.

Aghion, Philippe, Antonin Bergeaud, Timothee Gigout, Mathieu Lequien, and Marc Melitz. 2019. "Spreading Knowledge across the World: Innovation Spillover through Trade Expansion." Manuscript, Harvard University.

Aghion, Philippe, Leah Boustan, Caroline Hoxby, and Jerome Vandenbussche. 2009. "The causal impact of education on economic growth: evidence from US." *Brookings papers on economic activity* 1 (1): 1–73.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C Stein. 2008. "Academic freedom, private-sector focus, and the process of innovation." *The RAND Journal of Economics* 39 (3): 617–635.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60 (2): 323–351.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen. 2020. "The fall of the labor share and the rise of superstar firms." *The Quarterly Journal of Economics* 135 (2): 645–709.

Baruffaldi, Stefano, Brigitte van Beuzekom, Hélène Dernis, Dietmar Harhoff, Nandan Rao, David Rosenfeld, and Mariagrazia Squicciarini. 2020. *Identifying and measuring developments in artificial intelligence: Making the impossible possible.* Working Paper 2020/05. OECD, Sciences, Technology and Innovation Directorate.

BDI. 2011. *Germany 2030. Future perspectives for value creation.* Technical report. Https://espas.secure.europarl.europa.eu/orbis/document/germany-2030-future-perspectives-value-creation.

Benahmed-Miniuk, Fairouz, Mat Kresz, Jitendra K Kanaujiya, and Christopher D Southgate. 2017. "Genome-editing technologies and patent landscape overview." *Pharmaceutical patent analyst* 6 (3): 115–134.

Bergeaud, Antonin, Yoann Potiron, and Juste Raimbault. 2017. "Classifying patents based on their semantic content." *PloS one* 12 (4): e0176310.

Bessen, James, and Robert M Hunt. 2007. "An empirical look at software patents." *Journal of Economics & Management Strategy* 16 (1): 157–189.

Bloom, Nicholas, Tarek Alexander Hassan, Aakash Kalyani, Josh Lerner, and Ahmed Tahoun. 2021. *The Diffusion of Disruptive Technologies.* Technical report w28999. National Bureau of Economic Research.

Bloom, Nicholas, and John Van Reenen. 2007. "Measuring and explaining management practices across firms and countries." *The quarterly journal of Economics* 122 (4): 1351–1408.

Clarke, Nigel S, Björn Jürgens, and Victor Herrero-Solana. 2020. "Blockchain patent landscaping: An expert based methodology and search query." *World Patent Information* 61:101964.

Criscuolo, Chiara, and Jonathan Timmis. 2018. "GVCS and centrality," no. 12.

Dang, Jianwei, and Kazuyuki Motohashi. 2015. "Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality." *China Economic Review* 35:137–155.

Deloitte. 2021. *Future of the Tech Sector in Europe.* https://www2.deloitte.com/uk/en/pages/technology-media-and-telecommunications/articles/future-of-tech-in-europe.html. Accessed: 2021-05-26.

Diallo, Boubacar, and Wilfried Koch. 2018. "Bank concentration and Schumpeterian growth: theory and international evidence." *Review of Economics and Statistics* 100 (3): 489–501.

EPO. 2018. *Patents and self-driving vehicles. The inventions behind automated driving.* Report. EPO.

———. 2019. *Honouring a prolific inventor's dedication to advancing video compression: Marta Karczewicz named European Inventor Award 2019 finalist.* https://www.epo.org/news-events/press/releases/archive/2019/20190507n.html. Accessed: 2021-05-26.

———. 2020. *Top 10 Emerging Technologies 2020.* https://www.weforum.org/reports/top-10-emerging-technologies-2020. Accessed: 2021-05-26.

Fan, Peilei. 2014. "Innovation in China." *Journal of Economic Surveys* 28 (4): 725–745.

Forsberg, Brett. 2020. *Complete Beginner's Guide to Additive Manufacturing.* https://edgy.app/what-additive-manufacturing-beginners-guide-3d-printing. Accessed: 2022-10-08.

Furman, Jeffrey L, Michael E Porter, and Scott Stern. 2002. "The determinants of national innovative capacity." *Research policy* 31 (6): 899–933.

Giczy, Alexander V, Nicholas A Pairolero, and Andrew Toole. 2021. *Identifying artificial intelligence (AI) invention: A novel AI patent dataset.* Technical report.

Gillmore, Julian D, Ed Gane, Jorg Taubel, Justin Kao, Marianna Fontana, Michael L Maitland, Jessica Seitzer, Daniel O'Connell, Kathryn R Walsh, Kristy Wood, et al. 2021. "CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis." *New England Journal of Medicine* 385 (6): 493–502.

Goldin, Claudia, and Lawrence F Katz. 2010. *The race between education and technology.* harvard university press.

Griliches, Zvi. 1990. "Patent Statistics as Economic Indicators: A Survey." *Journal of Economic Literature* 28 (4): 1661–1707.

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg. 2005. "Market value and patent citations." *RAND Journal of economics,* 16–38.

He, Alex. 2021. *What Do China's High Patent Numbers Really Mean?*

Higham, Kyle, Gaétan De Rassenfosse, and Adam B Jaffe. 2021. "Patent quality: towards a systematic framework for analysis and measurement." *Research Policy* 50 (4): 104215.

Hu, Albert Guangzhou, and Gary H Jefferson. 2009. "A great wall of patents: What is behind China's recent patent explosion?" *Journal of Development Economics* 90 (1): 57–68.

Hudson, Institute. 2021. *5G Technological Leadership.*

IIPRD. 2017. *Sample patent landscape study - blockchain.* Report. IIPRD.

IP Australia. 2018. *Blockchain Innovation: A Patent Analytics Report.* Report. IP Australia.

———. 2019. *Machine Learning Innovation. A Patent Analytics Report.* Report. IP Australia.

Ito, Keiko, Kenta Ikeuchi, Chiara Criscuolo, Jonathan Timmis, and Antonin Bergeaud. 2019. "Global value chains and domestic innovation." In *RIETI Discussion Paper 19-E-028, April.* Research Institute of Economy, Trade / Industry. Tokyo.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights* 3 (3): 303–20.

Kennedy, Scott. 2015. *Made in China 2025.* https://www.csis.org/analysis/made-china-2025. Accessed: 2021-05-26.

König, Michael, Zheng Michael Song, Kjetil Storesletten, and Fabrizio Zilibotti. 2020. *From imitation to innovation: Where is all that Chinese R&D going?* Technical report w27404. National Bureau of Economic Research.

Krueger, Alan B, and Mikael Lindahl. 2001. "Education for growth: Why and for whom?" *Journal of economic literature* 39 (4): 1101–1136.

Lenz, David, and Peter Winker. 2020. "Measuring the diffusion of innovations with paragraph vector topic models." *PloS one* 15 (1): e0226685.

McKinsey. 2021. *The top trends in tech.* https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-top-trends-in-tech. Accessed: 2021-05-26.

Murray, Fiona, and Scott Stern. 2007. "Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior & Organization* 63 (4): 648–687.

OECD. 1998. *21st Century Technologies.* 170.

———. 2016. *Future technology trends.*

Rajan, Raghuram G., and Luigi Zingales. 1998. "Financial dependence and growth." *American Economic Review* 88 (3): 559.

Rassenfosse, Gaétan de, and Cyril Verluise. 2020. *PatCit: A Comprehensive Dataset of Patent Citations.* V. 0.15, March.

Review, MIT Technology. 2021. "10 Breakthrough Technologies 2021." *MIT Technology Review.*

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, Part 2): S71–S102.

Rotolo, Daniele, Diana Hicks, and Ben R Martin. 2015. "What is an emerging technology?" *Research policy* 44 (10): 1827–1843.

Schmookler, Jacob. 1966. *Invention and Economic Growth.* Harvard U.P.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti. 2011. "Growing like china." *American economic review* 101 (1): 196–233.

Squicciarini, Mariagrazia, Hélène Dernis, and Chiara Criscuolo. 2013. *Measuring Patent Quality: Indicators of Technological and Economic Value.* OECD Science, Technology and Industry Working Papers 2013/3. OECD Publishing, June.

Srebrovic, Rob. 2019. "Expanding your patent set with ML and BigQuery." Google Cloud Data Analytics https://cloud.google.com/blog/products/data-analytics/expanding-your-patent-set-with-ml-and-bigquery.

Tarasova, Nina N., and Polina Shparova. 2021. *Top 15 Digital Technologies in Manufacturing Industry.* https://issek.hse.ru/en/news/494926896.html. Accessed: 2021-09-05.

Trajtenberg, Manuel. 1990. "A Penny for Your Quotes: Patent Citations and the Value of Innovations." *The RAND Journal of Economics* 21 (1): 172–187.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.

Verluise, Cyril, Gabriele Cristelli, Kyle Higham, and Gaétan de Rassenfosse. 2020. "The missing 15 percent of patent citations." Available at SSRN 3754772.

Webb, Michael, Nick Short, Nicholas Bloom, and Josh Lerner. 2018. *Some Facts of High-Tech Patenting.* Technical report. National Bureau of Economic Research.

Wikipedia. 2021. *List of Prolific Inventors.* https://en.wikipedia.org/wiki/List_of_prolific_inventors. Accessed: Sept. 2021.

Williams, Heidi L. 2013. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy* 121 (1): 1–27.

WIPO. 2019. *The Geography of Innovation: Local Hotspots, Global Network.* Technical report, World Intellectual Property Report.

———. 2021. *Patent Landscape Reports.* https://www.wipo.int/patentscope/en/programs/patent_landscapes/. Accessed: 2021-05-26.

Zilibotti, Fabrizio. 2017. "Growing and slowing down like China." *Journal of the European Economic Association* 15 (5): 943–988.

Appendix

A Selection of technologies

A.1 Choice criteria

The term "technology" is ambiguous. It is used in many contexts to refer to distinct concepts. To make sure that we adopt a consistent approach, it's key to clearly distinguish between these concepts.

- **Technique:** We call *technique* a set of processes sharing a common methodological paradigm. Importantly, two distinct techniques differing by the methods involved can still share the same goals. E.g. statistical learning, deep learning, fuzzy logic for Natural Language Processing; TALENs, Crispr, Zinc fingers for genome editing; etc.
- **Functional application:** We call *functional application* a high level goal which are directly targeted by one or more techniques in the course of their development. They are not necessarily related to immediate market outcomes and the range of their market applications can vary. E.g. Computer vision, Natural Language Processing, Cultured meat, 3D printing, Genome editing, Bio plastic, etc.
- **Application field:** We call application field an existing or newly created economic sector which can leverage functional application to develop new/improve existing goods and services. E.g. Agriculture, Telecommunication, Transportation, etc. In some cases, the application field can be confounded with a family of *devices/goods/services* (e.g. smartphone)

We are interested in the so-called "breakthrough technologies". Using the above framework, breakthrough technologies correspond to *functional applications* which are expected to have a large *impact* on one or more application fields. Many *techniques* might be competing to become dominant at this functional application.

To select candidate technologies, we refer to various sources listing the potential technologies of the 21^{th} century, while keeping in mind that we need a technology to have the following features:

- Advancing rapidly or experiencing breakthrough that drive accelerated rates of change or discontinuous capability improvements
- Having a potential broad impact, i.e. touching various companies and industries and affecting (or giving rise to) a wide range of machines, products, or services.
- Having a high economic impact
- Being potentially disruptive, i.e. able to transform how people live and work, create new opportunities and businesses

- Being sufficiently discussed in the expert literature so we can access existing attempt at landscaping the technology
- Being sufficiently "new" to ensure that we do not consider technologies whose advanced is well planned and the results of an industry consensus (in particular the 4G, 5G, 6G communication protocol).

Our initial list of technologies is presented below along with the type of technology (A for application field, F for functional application, T for technique).

- **Computer Science:** Quantum computing (F/T), Blockchain (F), Three dimensional chip (T), Application-specific integrated circuit (T), Neuromorphic chips (T), Grid computing (T), Cloud computing (T), Field Programmable Gated Array (T), Edge computing (T)
- **Biotechnoloy:** Genome engineering (F), Personalized medicine (F), mRNA vaccines (F/T)
- **Information & Communication:** Internet of Things (F), Mega constellation (F), 5G/6G (T)
- **Energy:** Smart grid (F), Wind energy (F), Solar energy (F), Marine & Tidal energy (F), Internet of energy (F), Hydrogen storage (F), Fusion Power (T), Hydrogen battery (T), Advanced energy storage (T), Organic solar cell (T)
- **Transportation:** Self driving vehicles (F), Drones (F), Electric vehicle (A/F)
- **Agriculture:** Cultured meat (F), Vertical farming (F)
- **Materials:** Bioplastic (F), Additive manufacturing (F), Graphene material (T), Carbon nanotubes material (T)
- **Human-machine interface:** Exoskeleton (F), Brain computer interface (F)
- **Artificial Intelligence:** Computer vision (F), Natural Language processing (F), Speech processing (F), Machine translation (F)

A first selection based on selecting only functional application and removing technologies that are either at a too early or uncertain stage or without enough documentations in terms of patent landscaping boils down to the following list: hydrogen energy storage, blockchain, genome engineering, cultured meat, additive manufacturing, computer vision, natural language processing. While all these technologies would fit our criteria, we further reduce this list to 6 technologies. We remove natural language processing and cultured meat. The former because we already included computer vision as a technology within AI and the latter because the existing patent landscaping documents did not allow us to define a clear frontier.

A.2 Description of the selected technologies

A.2.1 Additive Manufacturing

A brief description Additive manufacturing, or 3D printing, is the construction of a three-dimensional object from a Computer-aided Design (CAD) model or a digital 3D model. Contrary to standard manufacturing techniques, additive manufacturing does not start from an existing block that would be cut and shaped but builds from a raw material, layer to layer. The very concept of 3D printing appeared in the 1950s (then called molecular spray) and the first patents filed are usually dated in the 1970s (depending on the sources, either by Charles W. Hull or by Johannes F Gootwald and in 1974 the term of 3D printing was coined in the New Scientist.

The term 3D priting encompass a large variety of underlying printing methods, the most commonly used being known as Fused deposition modeling (or FDM) uses a continuous filament of a thermoplastic that is directed by a head to create the desired shape. Among its advantages, 3D priting generates little waste and allows more customization and flexibility in creating complex shapes.

3D printing is still predominantly used in prototyping (40%) and some small and large scale finished goods production (30%) as well as research and education purposes (10%) in various sectors in particular automotive, aerospace and machine industry (EPO, 2020). While the technology is already well diffused in the industry, several challenges remain. First the cost of material is 10 to 200 more expensive than their non-printing equivalent. In addition, 3D printing is still too slow compared to other prototyping technologies. Second, there is an important need to extend the ability of current 3D printers to support more than 1 material at a time. Third, investment are needed to improve 3D printing of metallic device.

See Zastrow (2020) for more details.

Market potential In 2019, estimates of the additive manufacturing market is estimated at $10.9 worldwide (EPO, 2020). While it represents only 1% of valued added in manufacturing at this date, it could go up to 5% as the tech further mature. Many different industry are likely to adopt this process, from textile and in particular sportswear, aicraft and aerospace manufacturers to the design of very specific medical-device. Not surprisingly, its growth rate is expected to reach up to 20 percent per year during the next decade.

A.2.2 Blockchain

A brief description Blockchain is a distributed database (or ledger) shared across a public or private network. Each computer of the network gets a copy of the full ledger as a way to prevent system failure. The database itself is a growing list of records (called blocks) linked together using cryptography. Each block contains: i) a cryptographic hash of the previous block, ii) a timestamp and iii) transaction data. Consensus and or validation protocols are used to validate a new block before it can be added to the chain. This prevents fraud without the need of a central authority.

The development of blockchain is tied to the Bitcoin, but does not limit to the support of crypto-currencies. Indeed, the range of applications or potential application of this technology is very large (financial transactions more generally, but also any type of record and verification system such as patents, land titles etc...). Fundamentally, blockchain can be viewed as a way to ensure transactions in a broad sense in a low-trust environment without the need of a supervising actor.

The technology has been developed since the 1990s but experienced several breakthrough since the 2010s. In 2012, King and Nadal (2012) introduced the proof of stake which might be used as a replacement of the proof of work used, for example, as part of the bitcoin blockchain. The proof of stake overcomes a major limitation of early versions of the blockchain: energy consumption (due to many miners performing the same operation). It is notably used by the crypto-currency Eutherium.

In 2014, the Ethereum's white paper described Bitcoin as a weak version of smart contract - a transaction protocol intended to automatically execute, control or document legally relevant events and actions according to the terms of a contract or an agreement. Although smart contracts were first proposed in the early 1990s by Nick Szabo, envisioning blockchain as a support for smart contract in general considerably widens its potential impact and fields of applications by ascertaining trust between unknown parties.

See Zheng et al. (2017) and Zheng et al. (2018) for more details.

Market potential It is still difficult to assess the size of the market for blockchain. Some estimates suggest that the growth rate of total sales from blockchain could reach 50% per year and reach more than $40 billion by 2027. In any case, according to Carson et al. (2017), the potential developments of blockchain are very pervasive and broad and are likely to represent several billion in investment.

A.2.3 Computer Vision

A brief description Computer vision aims to give computers the ability to "understand" digital images and videos. "Understanding" corresponds to the transformation of visual images into descriptions of the world that are meaningful to thought processes and can prompt appropriate action. Computer vision is a field of Artificial Intelligence and has a wide variety of applications (face recognition, live translation of a text, autonomous vehicles...)

Computer vision started as early as the 1950s and distinguished from "rough" image processing by the desire to extract 3D representation from image. Recent resurgence in the field has been supported by considerable progress in machine learning and even more in deep learning. Deep learning algorithms have achieved accuracy close and in many application above, human performance on a set of benchmark tasks.

See Voulodimos et al. (2018) and Demush (2019) for more details.

Market potential Marr (2019) estimates the market size of computer vision to reach $48 billion in 2022. This size is expected to continue to grow given that computer vision has

(and is expected to have even more in the future) a large range of industrial applications. Automatic inspection of production (in manufacturing), event detection (e.g. wild fire), object modeling (3D printing), navigation (autonomous vehicle), information organization (automatic labeling/organization of databases), etc.

Even though modern computer vision has already found many industrial use cases, the recent domination of deep learning methods (Karpathy et al., 2014) promises additional extension to a number of industrial applications in the coming years.

A.2.4 Genome Editing

A brief description Genome editing (or genome engineering), is a type of genetic engineering in which DNA is inserted, deleted, modified or replaced in the genome of a living organism. Unlike early genetic engineering techniques that randomly inserts genetic material into a host genome, genome editing targets the insertions to site specific locations.

Genome editing was pioneered in the 1990s, its use was limited by low efficiencies of editing but has rapidly evolved in the 2000s. The three competing technologies in the field are zinc fingers, TALENs and CRISPR-Cas9. As described by Ledford (2015), researchers initially relied on zinc fingers, a class of enzymes, in order to accurately edit genomes. However, such enzymes were rather expensive. In 2012, CRISPR-Cas9 (or simply CRISPR) was introduced. It relies on an enzyme called Cas9 that uses a guide RNA molecule to home in on its target DNA, then edits the DNA to disrupt genes or insert desired sequences. In addition to being more efficient and easy to use, it is also much cheaper than previous technologies, including TALENs, the third competing method. As an order of magnitude, CRISPR costs about 150 times less than zinc fingers. It is now widely used and a very active subject of research and invention.

See Ledford (2015), Travis (2015), and Cohen (2017) for more details.

Market potential Market specialist Market and Markets projects the market size of genome editing at $11.7 billion by 2026. With countries moving to adjust the regulation to favor the development of genome editing applications, the growth of this technology is likely to be very high (Smyth and Wesseler, 2021).

Indeed, genome editing is expected to have a large impact in gene therapy in general, either by replacing existing treatments or treating illness which could not be cured so far (e.g. Down syndrome). Genome engineering is also said to have the potential to eradicate diseases by disrupting the genes encoding the production of a virus receptor surface (e.g. HIV, herpes and hepatitis B) or by removing disease predisposition genes (e.g. cancer).

A.2.5 Hydrogen Storage

A brief description Hydrogen energy storage denotes a set of technologies aiming at storing dihydrogen (H_2), in any form for later use. Traditionally, hydrogen generation is done by electrolysis using surplus energy production from renewable energy. The resulting hydrogen is then either used on-site or compressed and stored in tanks for transport and later use. However, recent interest in using hydrogen for energy storage on board clean

transportation vehicles has led to the development of new storage methods that are safer, smaller and more easily integrated to mobile units.

Hydrogen is an interesting source of energy: it has the highest energy per mass of any fuel and its combustion does not generate CO2. Another interesting feature is that, unlike electricity, hydrogen can be stored for extended period of time. It is however rather inefficient in terms of energy per unit of volume, in particular due to its very low boiling points (20.3K or -253°C). It is therefore very important to develop advanced storage methods that have potential for higher energy density.

The most important existing hydrogen storage methods include physical storage methods based on either compression or cooling or a combination of the two (hybrid storage). More recently, the use of nanomaterials has been proposed as an alternative option. Carbonaceous materials are currently being considered for onboard storage systems due to their versatility, multifunctionality, mechanical properties and low cost with respect to alternatives. The introduction of nanomaterials in onboard hydrogen storage systems is viewed as a major turning point for the future of hydrogen storage for the automotive industry.

For more details, see energy.gov

Market potential Various recent estimations of the market potential and future development of hydrogen storage are available. While the numbers vary, most experts concur that this technology should continue to grow in the next years.

Market analysis specialists such as Market Data Forecast or Allied Market Research forecast aggregate sales ranging from 19 to 25 billion dollars in 2027. The Hydrogen Council, a consortium of firms with stakes in the hydrogen market, project that "total investments will reach more than $300 billion in spending through 2030" (Hydrogen Council, 2021). Similarly, another group of market players, The Energy Transitions Commission, claimed that to reach zero net emission by 2050, an investment of $80 billion per annum will be required between 2020 and 2050 "for hydrogen production facilities and transportation & storage" (Energy Transitions Commission, 2021).

A.2.6 Self-driving Vehicle

A brief description A self-driving vehicle (or autonomous vehicle) is a vehicle that is capable of sensing its environment and moving safely with little or no human input. The technology can be divided into 2 broad sectors. First automated vehicle platform: items/hardware (e.g. sensors) and proceedings/software (e.g. algorithms) enabling the vehicle to make autonomous decisions. Second, smart environment which enables vehicles to interact with each other and their surrounding. Cars are classified into six different levels of autonomy. From no autonomy at all (level 0) to total autonomy, which makes human driving commands optional (level 5).

Since their diffusion in the early 20^{th} century, cars have become progressively more and more autonomous. However, as of 2020, only a marginal number of products have reached level 3 (vehicle that can be driven with no need for human action, except in some specific cases which requires some level of attention). Waymo, Aptiv and Dena have developed such "robo-taxis" but they are only deployed in a well-known extended neighborhood and

under standard weather conditions. In December 2020, Waymo opened its service to the public, becoming the world's first robo-taxi service. Similarly, the Tesla autopilot requires constant attention from a human driver but in October 2020, full self driving beta mode was introduced with the ability to navigate previously unseen streets (not only high-speed lanes) in autonomous mode.

Large scale adoption of a fully autonomous vehicle would require important legal, insurance and infrastructure adjustments as well as important guarantees in terms of security, even if the technology is well advanced.

See EPO (2018) for more details.

Market potential For the reasons explained above, self driving vehicles are virtually unavailable on the market but might appear in the coming years with potentially already existing cars "transiting" to self-driving vehicles as software and regulations get updated.

The diffusion of this technology could impact many aspect of society. In addition to converting driving time into leisure, self driving vehicles could open up to a "car as a service" model rather than "ownership" model generating potential savings. Car could also become a non depreciating asset as updates of the car software and modularity could generate continuous improvement of existing cars.

B Construction of the seed

B.1 Annotation guidelines

In order to manually assign one of the candidate patents to the seed or the anti-seed based on its abstract, we defined a series of tasks corresponding to each technologies. These tasks are presented in Table B1.

B.2 List of sources

In this section, we list the sources that we used to select relevant keywords, technological classes and patents to build the seed.

- **Additive Manufacturing:** EPO (2020), van de Kuilen (2015), Anish Mathews et al. (2020), and Zastrow (2020)
- **Blockchain:** IIPRD (2017), IP Australia (2018), Clarke, Jürgens, and Herrero-Solana (2020), and Isaacson (2020)
- **Computer Vision:** WIPO (2019b), WIPO (2019a) and Bo et al. (2021)
- **Genome Editing:** Jefferson et al. (2021)
- **Hydrogen Storage:** Baumann et al. (2021) and Office (2021)
- **Self-driving Vehicle:** EPO (2018) and Cho, Liu, and Ho (2021)

B.3 Criteria

We now detail the criteria by type and technology. The selection of candidate patents that we manually review to include in the seed must match at least one of the following criteria: 1) the patent's abstract contains at least one of the keywords (or keyphrases) listed in Section B.3.1; 2) the patent's CPC codes include at least one code listed in Section B.3.2; 3) the patent is highly similar to a patent listed Section B.3.3. The latter patents are patents known to be at the core of the technology and the similarity is based on Google Patents embedding and are directly provided by Google Patent.

B.3.1 Keywords

Additive Manufacturing 3d-printing, stereolithography, additive manufacturing, three-dimensional objects, rapid prototyping, additive material manufacturing three dimensional printing material, 3d-printing materials photolithography, fuse deposition mode

Blockchain blockchain, digital mining, bitcoin, cryptocoin, cryptocurrency, digital wallet, ethereum, smart contracts, record keeping, distributed ledger, distributed node, private ledger, public ledger, intelligent node, full node, digital signatures, public key, user identity, hashing, consensus methodologies, proof of work, proof of stake, deposition based, ripple

Computer Vision adaboost, xgboost, bayesian network, decision tree, genetic algorithm, gradient tree boosting, logistic regression, random forest, rankboost, support vector machine, multilayer perceptron, hidden markov model, generalized adversarial network, backpropagation, stochastic gradient descent, supervised training, reinforcement learning, neural network, self learning, semi supervised learning, unsupervised training, transfer learning, overfitting, active learning, clustering, data mining, deep learning, expert system, embedding, machine learning, fuzzy logic, feature selection, objective function, target function, regression model, signal processing, computer vision, machine vision, lidar, character recognition, optical character recognition, handwritten character recognition, image to text, text recognition, face recognition, facial recognition, biometric data, biometrics, mass surveillance, face unlock, traffic cameras, object detection, edge detection, obstacle avoidance, motion tracking

Genome Editing dna editing, gene editing, genome engineering, recombinant targeting vectors, homologous recombination, double-strand dna break, homology-directed repair, targeted dna sequence, dna cleavage, fok1, sequence-specific nuclease system, zinc finger nuclease, cys2-his2, transcriptional activator-like effector nuclease, talens, clustered regularly interspaced short palindromic repeat, crispr/cas, cas9, pre-crrna, tracrrna, enzyme rnase, single guide rna, crispr-cpf1, ngago, single-stranded dna-guided argonaute endonuclease, natronobacterium gregoryi argonaute

Hydrogen Storage hydrogen fuel cells, hydrogen storage, liquid hydrogen, solid-state hydrogen storage, compressed hydrogen storage, dehydrogenation reaction, hydrogen gas, hydrogen fuel, hydrogen storage materials, hydrogen-powered device

Self Driving Vehicle self-driving vehicle, autopilot, driverless vehicle, autonomous vehicle, automated vehicles, vehicle connectivity, vehicle-to-vehicle communication, fleet management, vehicle lidar, vehicle sonar, vehicle radar, vehicle camera, object detection, obstacle detection, object classification, cruise control, pedestrian detection, environment mapping, surround view, blind spot detection, park assistance, lane departure, traffic sign recognition, drive assist system, trajectory generation, reactive control, path trajectory planning, manoeuvers planning

B.3.2 CPC classes

Additive Manufacturing B81C2201/0184, G05B2219/49002, G05B2219/49003, G05B2219/49004, G05B2219/49005, G05B2219/49006, G05B2219/49007, G05B2219/49008, G05B2219/49009, G05B2219/49011, G05B2219/49013, G05B2219/49014, G05B2219/49015, G05B2219/49016, G05B2219/49017, G05B2219/49018, G05B2219/49019, G05B2219/49021, G05B2219/49022, G05B2219/49023, G05B2219/49024, G05B2219/49025, G05B2219/49026, G05B2219/49027, G05B2219/49028, G05B2219/49029, G05B2219/49031, G05B2219/49032, G05B2219/49033, G05B2219/49034, G05B2219/49035, G05B2219/49036, G05B2219/49037, G05B2219/49038, G05B2219/49039, A43D2200/60, A23P2020/253, B29C64/10, C08L101/00, B29C67/00, B22F3/00, G05B2219/49013, G03F7/70416, B28B1/001, B33Y10/00, B23K9/04, B23K10/027, B23K15/0086, B23K11/0013

Blockchain H04L009/08, H04L67/00, H04L009/10, H04L009/12, H04L009/14, H04L009/28, H04L29/06, G06Q20/00, G06F21/00, G06F12/14, G06Q20/06, G06Q20/10, G06Q20/20, G06Q20/32, G06Q20/36, H04L2209/00, G09C001/00, G09C001/02, G09C001/04, G09C001/06, H04L63/00, G06Q30/0619, G06F21/00, G06F021/24, G06F021/00, G06F021/02, G06F012/28, G06F012/14, G06F17/00

Computer Vision B25J9/161, G06F17/16, G06N5/003, G06N7/005, G06N7/046, B29C66/965, G08B29/186, F02D41/1405, G01N29/4481, G06F11/1476, G06F17/2282, H02P21/0014, H02P23/0018, H03H2222/04, Y10S128/924, Y10S128/925, B64G2001/247, F05B2270/707, F05B2270/709, F05D2270/709, G10H2250/151, H04L25/03165, H04Q2213/054, H04Q2213/343, B60G2600/1876, B60G2600/1878, B60G2600/1879, E21B2041/0028, F16H2061/0081, F16H2061/0084, G06F2207/4824, G10K2210/3024, G10K2210/3038, H03H2017/0208, B29C2945/76979, G05B2219/33002, G06T2207/20081, G06T2207/20084, G06T2207/20084, H04L2025/03464, H04L2025/03554, H04Q2213/13343, B60W30/06, B60W30/10, B60W30/12, B60W30/14, B60W30/17, G06T9/002, G10L25/30, G06K7/1482, G06T3/4046, B62D15/0285

Genome Editing A01H4/00, A01K67/00, C12N/1500, C12N1/00, C12N5/00, C12N7/00C12Y, C12N5/10, C12Q1/68, C12Q1/70, G01N33/00, A61K48/00, A61K31/7088, C07K14/00

Hydrogen Storage Y02E60/30, Y02E60/32, Y02E60/321, Y02E60/322, Y02E60/324, Y02E60/325, Y02E60/327, Y02E60/328, Y02E60/34, Y02E60/36, Y02E60/362, Y02E60/364, Y02E60/366, Y02E60/368, B01D53/02, C01B3/00-58, F17C2221/012, C22C19/03, C22C22/00, C22C33/00, F25B17/12, H01M4/38, H01M8/06, F17C2221/012, F17C6/00, F17C5/02

Self Driving Vehicle G08G1/02, G08G1/0967, G08G1/0968, G01S7/003, G07B15/063, G07C5/00, G07C5/12, E01F, E01F9/00, E01F9/40, H04W36/00, H04W76/50, B61L3/00, G05D1/0011, G05D1/0027, G05D1/0287, G05D1/0297, G08G1/00, G08G1/01, G08G1/09, G08G1/0968, G08G1/127, G08G1/16, G08G1/164, G08G1/20, G01S13/93, G10S13/931, G01S15/88, G01S15/93, G01S17/88, G01S17/93, G07C5/00, G07C5/01, G07C5/02, G07C5/03, G07C5/04, G07C5/05, G07C5/06, G07C5/07, G07C5/08, E01F9/00, B60L2240/70, B61L25/00, G01S7/00, G01S13/00, G01S15/00, G01S17/00, G01S7/00, G01S7/02, G01S7/52, G01S13/00, G01S13/86, G01S13/87, G01S13/93, G01S15/00, G01S15/025, G01S15/87, G01S15/931, G01S17/00, G06K9/00, G05D1/00, G05D1/0257, B60W2420/52, B60Y2400/3017, B60R19/00, G01S17/023, G01S17/06, G01S17/87, G01S17/88, G01S17/936, G01S7/48, G01S2013/9332, B60W2420/52, G06T1/0007, G06T1/0014, G06T1/20, G06K9/00362, G06K9/00785, G06K9/00791, H04N5/335, B60Y2400/3015, B60W2420/42, B60S1/56, G01C21/00, G01C21/26, G01C21/34, G01S7/52, G01S15/00, G05D1/00, G05D1/0027, G05D1/0088, G05D1/021, G05D1/0212, G05D1/0276, G05D1/0287, G05D1/02, G06T1/0007, G06T1/0014, G06T1/20, G08G1/16, G08G1/161, G08G1/22, H04W4/44, H04W4/46, F16D2500/31, B60L2240/60, B60L2240/62, B60W30/16, B60W2050/008, B60W2550/402, B60W2550/408, B60G17/015,

B60G17/016, B60G17/0195, B60G2800/00, B60K28/04, B60W30/00, B60W40/00, F16D2500/508, G05D1/0088, G05D2201/0212, B60W30/095, B60W50/0097, G05D1/0212

B.3.3 Representative patents

Additive Manufacturing US-4575330-A, US-5534104-A, US-6259962-A, US-5204055-A, US-5182056-A, DE-102013205724-A1, FR-3070302-B1, US-10076875-B2, US-8349239-B2, CN-108868141-A, CN-105569344-A, CN-105604327-A, WO-2018229418-A1, KR-101706473-B1, WO-2016111879-A1, US-20180141274-A1, WO-2008061909-A2, US-20170251713-A1, EP-1352619-B1, EP-3319545-B1, EP-3151782-B1, US-10441426-B2, US-9056017-B2

Blockchain EP-3125489-B1, US-9785369-B1, DE-102016104478-A1, US-9853819-B2, US-9842216-B2, US-9855785-B1, US-20180137465-A1, US-9635000-B1, EP-329562-A1, EP-3295350-B1, CN-105719172-A, CN-105701372-B, US-9836908-B2, US-9818092-B2, US-9824031-B1, US-10643202-B2, CN-105844505-A, US-9298806-B1, CN-105790954-B, US-9858781-B1, US-9853977-B1, US-9641338-B2, US-9641342-B2, EP-325719-B1

Computer Vision US-8953886-B2, WO-2003023696-A1, US-5881172-A, US-20170024607-A1, US-20170169205-A1, US-20170169303-A1, US-20170235931-A1, US-20200175326-A1, US-10872228-B1

Genome Editing WO-2000041566-A9, WO-2003087341-A3, WO-2010079430-A1, WO-2011072246-A2, US-8440431-B2, US-8440432-B2, US-8450471-B2, US-8566363-B2, WO-2014093661-A2, WO-2013176772-A1, US-20170367280-A1

Hydrogen Storage US-20080248355-A1, CN-1322266-C, US-7678362-B2, US-7118611-B2, CN-203500844-U, US-7094493-B2, US-10622655-B2, WO-2019239141-A1, US-8871671-B2, JP-6061354-B2, EP-2554694-B1, FR-2939784-A1, CA-2980664-C, CN-103797142-A, US-7678479-B2, JP-6418680-B2, DE-102009016475-B4, US-7093626-B2, DE-102013203892-A, KR-101107633-B1, JP-4849775-B2, JP-3706611-B2, US-6875536-B2, JP-5338903-B2

Self Driving Vehicle US-20050088318-A1, US-9293045-B2, US-9723457-B2, US-10405215-B2, WO-2019052353-A1, US-10089537-B2, US-10564639-B1, DE-112019000049-T5, US-20190384304-A1, DE-112019000122-T5, US-20170030728-A1, US-20190265703-A1, WO-2019094843-A1

Table B1: Annotation guidelines

Technology	Options
Additive Manufacturing	- Create 3D printable model with computer aided design - Examine stereolithography file for errors and inconsistency - Convert model into a series of thin layers - Manufacture materials for 3D printings - Print 3D model
Blockchain	- Record transactions between two parties - Serve as public transaction ledger of cryptocurrency - Execute or enforce smart contract - Hash tree verification / Verify the authenticity of documents / Proof of work - Analyse transactions in a distributed ledger - Manage Identity System based on the concept of peer-to-peer protocols (IDMS) / Mediate user authentication
Computer Vision	- Process digital images - Analyse digital images - Understand digital images
Genome Editing	- Target DNA sequence - Break DNA sequence - Edit DNA sequence
Hydrogen Storage	- Hydrogen production and compression - Generate power from hydrogen gas - Design vessel containment that is resistant to hydrogen permeation and corrosion (+ thermal management) - Manufacture fuel cell using hydrogen - Provide hydrogen to a hydrogen-powered device (fill, tank)
Self-driving Vehicle	- Enable vehicles to make autonomous decisions - Automate vehicle handling - Vehicle-to-vehicle communication - Communication between vehicle and rest-of-the-world

Notes: Human annotator accepts or rejects a candidate patent depending on whether the patent's abstract clearly discusses one or more of the options listed.

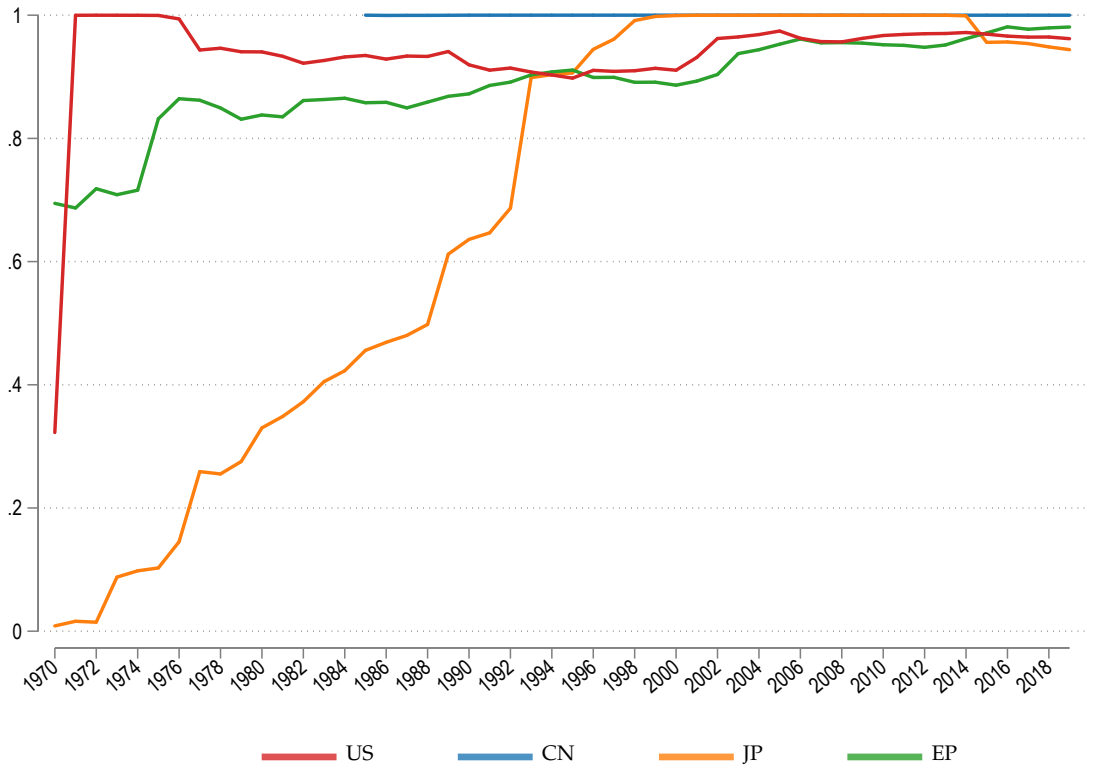
C Data

C.1 Selection of patents

In this Appendix we detail how we selected the first priority publication associated with a patent family and thus avoid double counting.

We use the IFI CLAIMS dataset which is available through Google's BigQuery. This dataset contains bibliographical information on a very large number of patents as well as the abstract translated in English when available. The share of patents with an abstract that we could use increases in time but is essentially stationary and above 90% since 2000 in all patent offices considered (see Figure C1).

Figure C1: Share of patent families with an abstract in English



Notes: Share of patents with at least one patent in the same family with an abstract in English in each of the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP). The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

Each patent publication is association with a unique number and belongs to a family which corresponds to a group of publications that share the same priority claims. As explained in Section 4.1, the landscaping is deployed at the family level. Hence, all patents belonging to a family that is assigned to one of the technologies we are interested in will also be assigned to this technology. Since the analysis is done at the patent publication level, we avoid multiple counting similar patents in a given family by restricting to the earliest priority application in the family.

The final number of publications considered by technologies and patent office is given in Table C1

Table C1: Number of unique patent publications considered by patent office and technology

	USPTO	CNIPA	EPO	JPO
Additive Manufacturing	17,242	9,014	5,519	3,286
Blockchain	6,409	7,972	952	1,219
Computer Vision	208,720	144,831	50,038	222,177
Genome Editing	22,302	6,470	5,376	2,707
Hydrogen Storage	3,372	2,589	1,905	6,703
Self-driving Vehicle	62,127	33,142	40,960	68,257

Notes: Number of observations in each of the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP). National European patent offices include all EU countries, UK, Norway and Switzerland.

C.2 Construction of the index of radicalness

Measuring the novelty, impact and radicalness of patents is a complicated task. Initially, Hall, Jaffe, and Trajtenberg (2005) considered the difference between the set of technological classes in citing patents and the set of technological classes in cited patents (see also Squicciarini, Dernis, and Criscuolo, 2013). Intuitively, a radical innovation would be different from existing (but related) knowledge and would influence subsequent developments (and potentially makes the existing technological classifications unsuitable).

One limitation of this approach is that it ignores potential radical innovation within a well defined technology. Recently, Kelly et al. (2021) have proposed a new measure that relies on the text of patent publications. Formally, they use natural language processing to compare the occurrence of words and group of words in a given patent with previous publications made in the 5 year window before publication. This define a distance that they then compare with the similar distance constructed with patents published in the 5 year window following its publication.

We use and adapt their approach. First, because we want to construct such a measure of radicalness for all patents, and not only for USPTO's, we cannot use the text. Instead, we rely to the embedding vector representation provided by Google Patent Research (see Srebrovic, 2019). The GP embedding is a 64-dimensional vector that was constructed using machine learning with the goal of measuring distance between two patents (this is what Google Patent use to provide its list of "similar patent" that we use in the construction of the seed, see Appendix B. Each coordinate of the vector is a continuous variable between -1 and +1. It therefore provides a simple algebraic representation from which we can compute simple distances by taking the scalar product between the two corresponding vectors.

Formally, for each patent p, we denote by $\mathcal{E}(\mathbf{p})$ the corresponding embedding vector. We then define the distance between a patent p and a patent q as :

$$d(p,q) = \mathcal{E}(\mathbf{p}).\mathcal{E}(\mathbf{q}) \in [0,1]\,.$$

The second adjustment that we do is that we calculate our measure of radicalness by comparing patents *within* each of our six technologies. Hence, we assign a quantity between 0 and 1 to each patent that measure to what extent it was pivotal in the development of the

corresponding technology. By contrast, Kelly et al. (2021) compare a given patent with all existing publications as their goal is to exhibit the birth of new technologies.

Formally, we proceed as follow. Let $t(p)$ denotes the year of publication of patent p and $\mathcal{P}(t,k,X)$ the set of patents published between year t and $t+k$ in technology X (X equal additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage or self-driving vehicles). Then we first define:

$$I(p) = \sum_{q \in \mathcal{P}(t(p)+1,5,X)} \frac{\mathcal{E}(p).\mathcal{E}(q)}{|\mathcal{P}(t(p)+1,5,X)|}.$$

$I(p)$ is a measure of the *impact* of patent p and is defined as the dot product between the embedding vector of p and the average embedding vector of all patents published in the same technology in the next 5 year.

Similarly, we define:

$$N(p) = \sum_{q \in \mathcal{P}(t(p)-6,5,X)} \frac{\mathcal{E}(p).\mathcal{E}(q)}{|\mathcal{P}(t(p)-6,5,X)|},$$

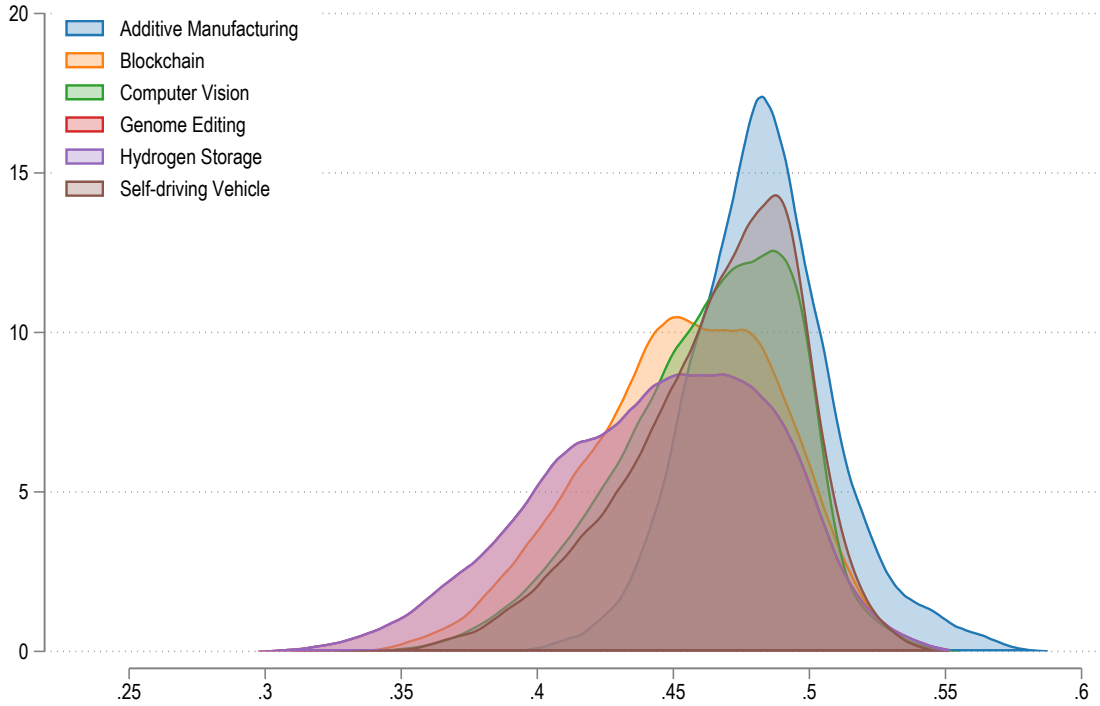
a measure of the (inverse) novelty of the patent, obtained by comparing patent p with patents published in the past 5 years.

From N and I we can define the measure of radicalness of a patent by taking their geometric average:

$$R(p) = \sqrt{I(p)(1-N(p))}.$$

Note that in order to accommodate the need to calculate the impact measure we cannot compute radicalness for patents issued after 2014. We also start in 2000 so that we have enough patents. Figure C2 plots the distribution of novelty by technology for all countries and years from 2000 to 2014.

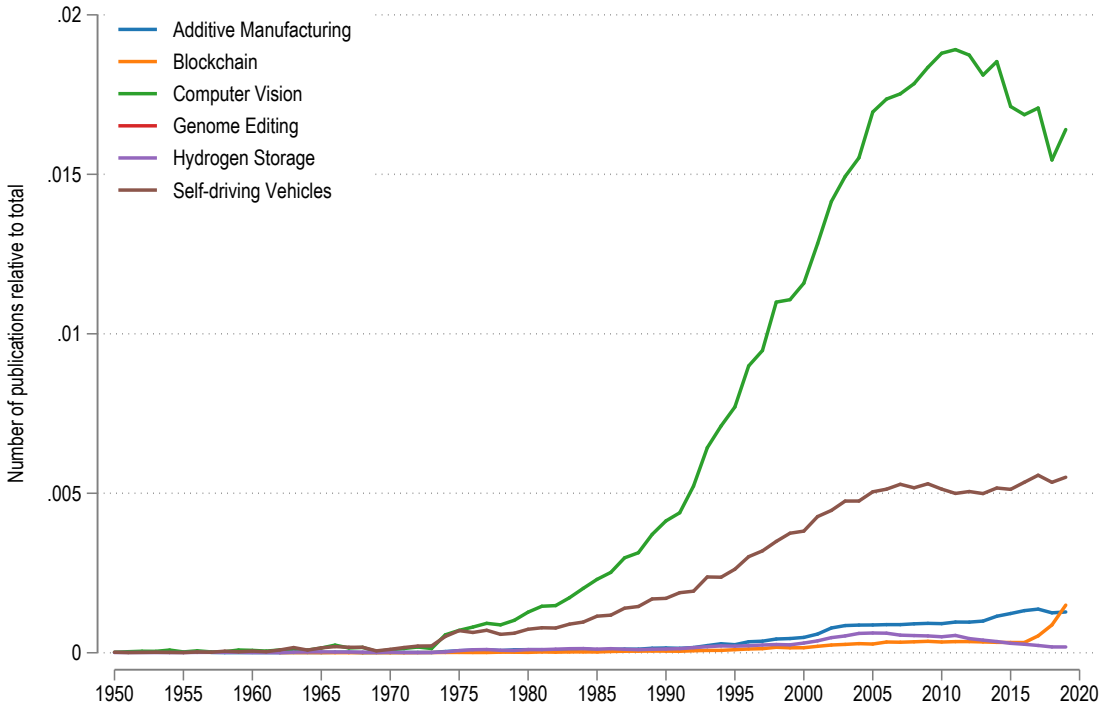
Figure C2: Distribution of radicalness indicator by technology



Notes: Kernel density estimations of the radicalness indicator by technology. All years and countries are pooled together. See Appendix C.2 for more details.

D Additional results

Figure D1: Number of patent publications by technology



Notes: Total number of patents published yearly at the USPTO, CNIPA, EPO and European national patent offices and JPO in each of the six technologies considered. This number has been standardized by the total number of patents published in these patents office in any technology. The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

Table D1: Precision from simple rule-based classification

	Additive Manufacturing	**Blockchain**	**Computer Vision**
CPC	0.34 (224)	0.01 (458)	0.26 (298)
Keywords	0.16 (218)	0.09 (456)	0.20 (254)
Patents	0.02 (42)	0.09 (53)	0.4 (5)
	Genome Editing	**Hydrogen Storage**	**Self driving Vehicle**
CPC	0.05 (172)	0.14 (211)	0.12 (222)
Keywords	0.89 (158)	0.24 (221)	0.36 (239)
Patents	0.57 (7)	0.16 (32)	0.42 (24)

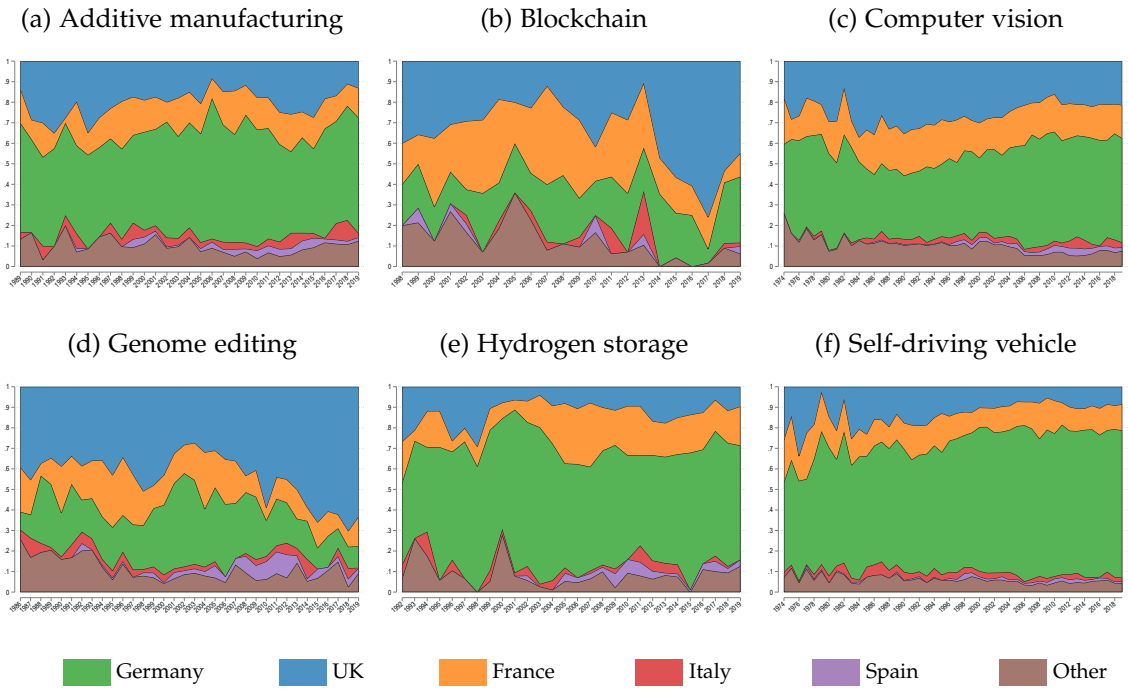
Notes: Precision computed on the test set - unseen during training using simple rule-based classification (either using only CPC, only keywords or similar patents from the set of manually added patents). Number in parentheses corresponds to the number of patents.

Table D2: 10 most cited articles by technology

	Additive manufacturing	Blockchain	Computer vision	Genome editing	Hydrogen storage	Self driving vehicle
1	10.1116/1.1412895	10.1145/121133.121137	10.1109/tcsvt.2003.815165	10.1038/35078107	10.1016/s0378-7753(97)02724-9	10.4271/961010
2	10.1021/ma60060a028	10.1145/37499.37500	10.1109/cvpr.2001.990517	10.1093/emboj/20.23.6877	10.1002/ceat.270100130	10.3189/s0260305500002822
3	10.1007/bfb0053286	10.1145/50202.50214	10.1109/iccv.2001.937709	10.1101/gad.862301	10.1016/s0360-3199(00)00021-5	10.1049/cp:19960454
4	10.1109/vlsit.2007.4339708	10.1109/tcsvt.2003.815173	10.1109/tcsvt.2003.815173	10.1038/35888	10.2172/460349	10.1109/ivs.2000.898318
5	10.1145/237170.237191	10.1109/tcsvt.2003.815165	10.1109/iccv.1999.790410	10.1126/science.2315699	10.1002/cjce.5450690503	10.1109/imtc.2001.929558
6	10.1145/311535.311556	10.1109/2.16	10.1007/bfb0053007	10.1126/science.1231143	10.1002/cjce.5450690504	10.1109/irds.2002.1041378
7	10.1145/37402.37422	10.1109/tcsvt.2003.814963	10.1016/b978-0-08-051581-6.50024-6	10.1126/science.1232033	10.1246/cl.1993.41	10.1109/imtc.1999.776736
8	10.1116/1.591000	10.1109/dcc.1998.672152	10.1109/34.888718	10.1038/332323a0	10.1016/0360-3199(95)00131-x	10.1049/cp:19980155
9	10.1117/12.968423	10.1145/989.990	10.1109/tcsvt.2007.905532	10.1038/327070a0	10.1016/s0378-7753(97)02760-2	10.1007/s10967-005-0069-2
10	10.1002/pol.1979.170170410	10.1109/tcsvt.2003.815175	10.1889/1.3256703	10.1038/313810a0	10.1016/s0378-7753(97)02796-1	10.1016/s0034-6667(99)00031-7

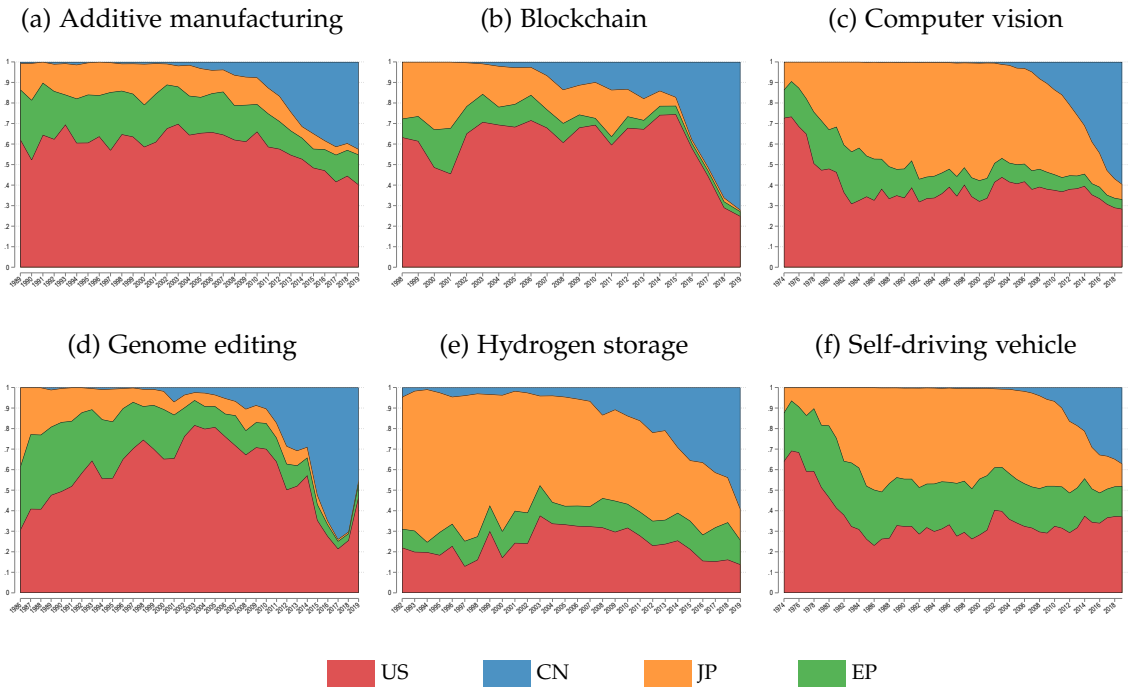
Notes: Top 10 academic papers cited in the patents in each six technologies in the frontpage retrieved using PatCit (Rassenfosse and Verluise, 2020).

Figure D2: Relative contribution to frontier technologies. European patents



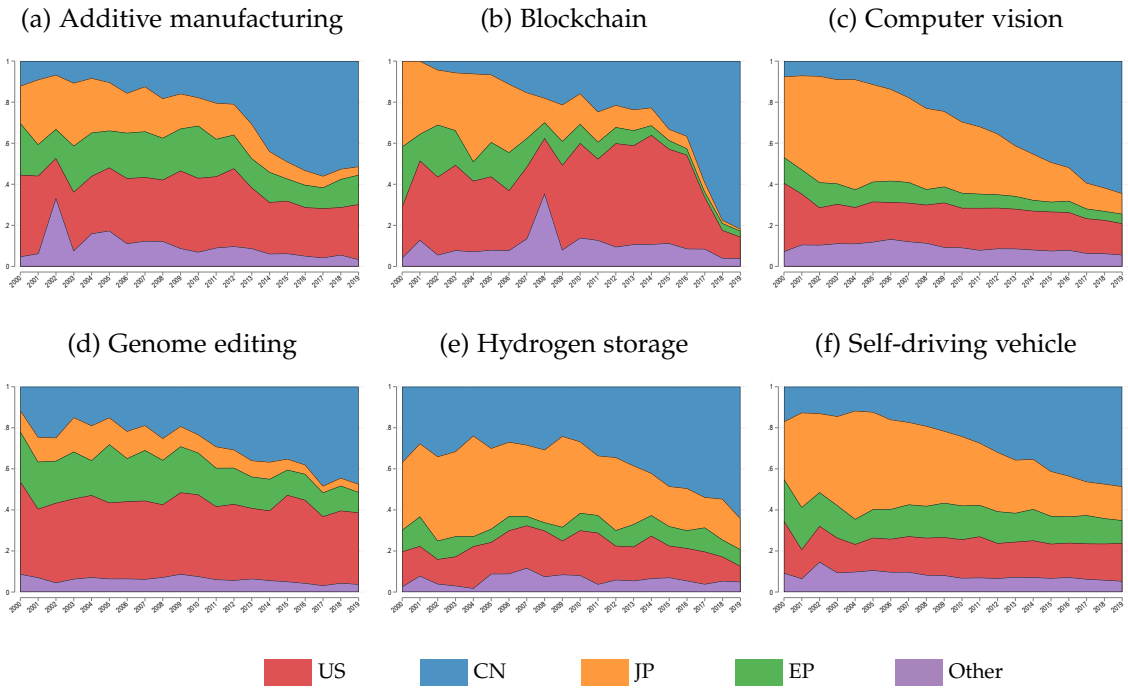
Notes: Yearly share of patents in national European patent offices. The share is calculated over the total of European patents, excluding patents filed at the EPO, and includes earliest publication of priority fillings in each family.

Figure D3: Relative contribution to frontier technologies - weighted by foreign citations



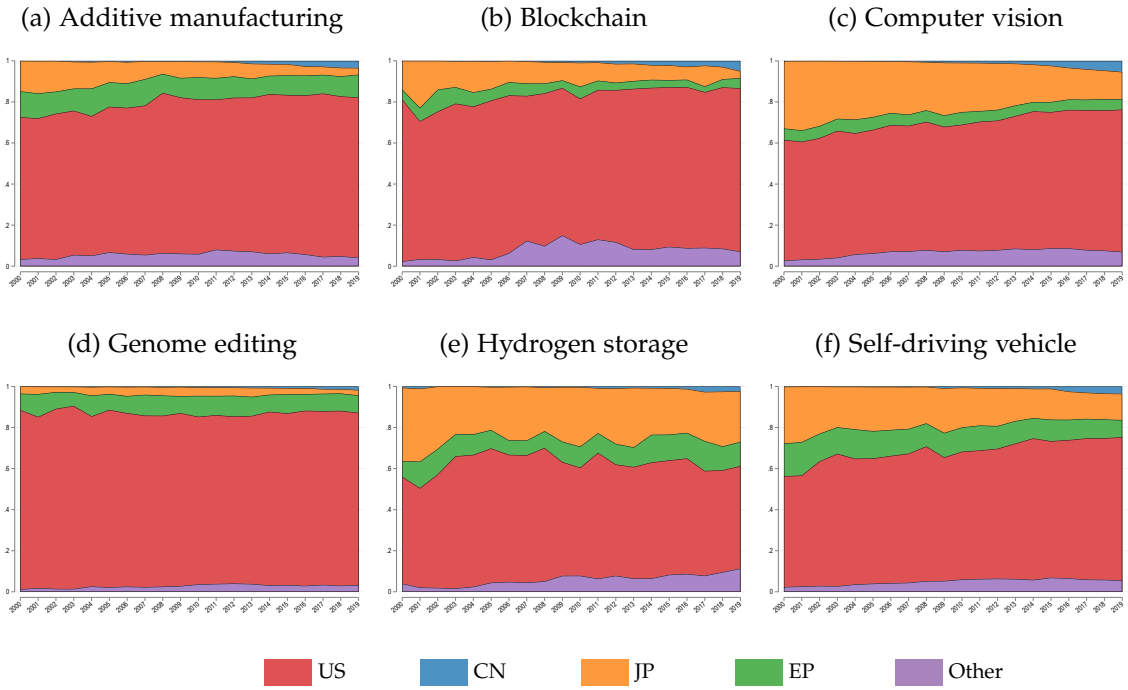
Notes: Patent counts in the four patent offices: USPTO (US), CNIPA (CN), EPO and European national patent offices (EP) and JPO (JP) as a share of the total patent count for each technology. Patents count is weighted by the number of forward citations received by patents in other patent offices. The year of publication is reported in x-axis. National European patent offices include all EU countries, UK, Norway and Switzerland.

Figure D4: Origin of priority filings for CNIPA patents



Notes: Yearly share of CNIPA patents with a priority filing in different patent offices (USPTO for the US, CNIPA for China, JPO for Japan, EPO and national European patent offices for Europe and European countries and a category for all other patent offices).

Figure D5: Origin of priority filings for USPTO patents



Notes: Yearly share of USPTO patents with a priority filing in different patent offices (USPTO for the US, CNIPA for China, JPO for Japan, EPO and national European patent offices for Europe and European countries and a category for all other patent offices).

Appendix References

Anish Mathews, Priya, Swati Koonisetty, Sanjay Bhardwaj, Papiya Biswas, Roy Johnson, and G Padmanabham. 2020. "Patent Trends in Additive Manufacturing of Ceramic Materials." *Handbook of Advanced Ceramics and Composites: Defense, Security, Aerospace and Energy Applications,* 319–354.

Baumann, Manuel, Tobias Domnik, Martina Haase, Christina Wulf, Philip Emmerich, Christine Rösch, Petra Zapp, Tobias Naegler, and Marcel Weil. 2021. "Comparative patent analysis for the identification of global research trends for the case of battery storage, hydrogen and bioenergy." *Technological forecasting and social change* 165:120505.

Bo, Zhang, Lyu Lucheng, Wang Yanpeng, Zhao Yajuan, and Qian Li. 2021. "Global Patent Analysis of Computer Vision." *Science Focus* 16 (2): 72–83.

Carson, Brant, Giulio Romanelli, Patricia Walsh, and Askhat Zhumaev. 2017. *Blockchain beyond the hype: What is the strategic business value?* McKinsey Insights.

Cho, Rico Lee-Ting, John S Liu, and Mei Hsiu-Ching Ho. 2021. "The development of autonomous driving technology: perspectives from patent citation analysis." *Transport Reviews* 41 (5): 685–711.

Clarke, Nigel S, Björn Jürgens, and Victor Herrero-Solana. 2020. "Blockchain patent landscaping: An expert based methodology and search query." *World Patent Information* 61:101964.

Cohen, Jon. 2017. "How the battle lines over CRISPR were drawn." *Science* 15.

Demush, Rostyslav. 2019. *A Brief History of Computer Vision (and Convolutional Neural Networks).* https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3. Accessed: 2021-05-26.

Energy Transitions Commission. 2021. *Making Clean Electrification Possible: 30 Years to Electrify the Global Economy.* Technical report, Making Mission Possible Series.

EPO. 2018. *Patents and self-driving vehicles. The inventions behind automated driving.* Report. EPO.

———. 2020. *Patents and additive manufacturing: Trends in 3D printing technologies.* Report. European Patent Office.

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg. 2005. "Market value and patent citations." *RAND Journal of economics,* 16–38.

Hydrogen Council. 2021. *A perspective on hydrogen investment, market development and cost Competitiveness.* Hydrogen Insight.

IIPRD. 2017. *Sample patent landscape study - blockchain.* Report. IIPRD.

IP Australia. 2018. *Blockchain Innovation: A Patent Analytics Report.* Report. IP Australia.

Isaacson, Thomas. 2020. *The Blockchain Patent Landscape Shows Accelerating Growth.* IPwatchdog articles. IPwatchdog.

Jefferson, Osmat Azzam, Simon Lang, Kenny Williams, Deniz Koellhofer, Aaron Ballagh, Ben Warren, Bernard Schellberg, Roshan Sharma, and Richard Jefferson. 2021. "Mapping CRISPR-Cas9 public and commercial innovation using The Lens institutional toolkit." *Transgenic Research* 30 (4): 585–599.

Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. "Large-scale video classification with convolutional neural networks." In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,* 1725–1732.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights* 3 (3): 303–20.

King, Sunny, and Scott Nadal. 2012. "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake." *self-published paper, August* 19 (1).

Ledford, Heidi. 2015. "CRISPR, the disruptor." *Nature News* 522 (7554): 20.

Marr, B. 2019. "Amazing Examples of Computer and Machine Vision in Practice." *Forbes.*

Office, Intellectual Patent. 2021. *Low-carbon hydrogen: A worldwide overview of patenting related to the UK's ten point plan for a Green Industrial Revolution.*

Rassenfosse, Gaétan de, and Cyril Verluise. 2020. *PatCit: A Comprehensive Dataset of Patent Citations.* V. 0.15, March.

Smyth, Stuart J, and Justus Wesseler. 2021. "The future of genome editing innovations in the EU." *Trends in Biotechnology.*

Squicciarini, Mariagrazia, Hélène Dernis, and Chiara Criscuolo. 2013. *Measuring Patent Quality: Indicators of Technological and Economic Value.* OECD Science, Technology and Industry Working Papers 2013/3. OECD Publishing, June.

Srebrovic, Rob. 2019. "Expanding your patent set with ML and BigQuery." Google Cloud Data Analytics https://cloud.google.com/blog/products/data-analytics/expanding-your-patent-set-with-ml-and-bigquery.

Travis, J. 2015. "CRISPR genome-editing technology shows its power." *Science* 350:1456–7.

van de Kuilen, Aalt. 2015. *Using PatBase for Patent Landscaping: a case study on 3D printing techniques.* Report. Minesoft.

Voulodimos, Athanasios, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. "Deep learning for computer vision: A brief review." *Computational intelligence and neuroscience* 2018.

WIPO. 2019a. *Artificial Intelligence.* Technical report, echnology Trends 2019: artificial intelligence.

———. 2019b. *Data collection method and clustering scheme.* Technical report, Technology Trends.

Zastrow, Mark. 2020. "3D printing gets bigger, faster and stronger." *Nature* 578 (7793): 20–24.

Zheng, Zibin, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. 2018. "Blockchain challenges and opportunities: A survey." *International journal of web and grid services* 14 (4): 352–375.

Zheng, Zibin, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang. 2017. "An overview of blockchain technology: Architecture, consensus, and future trends." In *2017 IEEE international congress on big data (BigData congress)*, 557–564. Ieee.