DOES TRAINING IN AI AFFECT PHD STUDENTS' CAREERS? EVIDENCE FROM FRANCE

Documents de travail GREDEG
GREDEG Working Papers Series

Pierre Boutros
Eliana Diodati
Michele Pezzoni
Fabiana Visentin

GREDEG WP No. 2025-29
https://ideas.repec.org/s/gre/wpaper.html

Does Training in AI Affect PhD Students' Careers? Evidence from France

Pierre Boutros
Université Côte d'Azur, CNRS, GREDEG, France
pierre.boutros@univ-cotedazur.fr

Eliana Diodati
University of Torino, Italy
Collegio Carlo Alberto, Turin, Italy
eliana.diodati@unito.it

Michele Pezzoni
Université Côte d'Azur, CNRS, GREDEG, France
Observatoire des Sciences et Techniques, HCERES, Paris, France
michele.pezzoni@univ-cotedazur.fr

Fabiana Visentin
UNU-MERIT, Maastricht University, the Netherlands
f.visentin@maastrichtuniversity.nl

GREDEG Working Paper No. 2025-29

Abstract: The rise of Artificial Intelligence (AI) urges us to better understand its impact on the labor market. This paper is the first to analyze the supply of individuals with AI training facing the labor market. We estimate the relationship between AI training and individuals' careers for 35,492 French PhD students in STEM who graduated between 2010 and 2018. To assess the unbiased effect of AI training, we compare the careers of PhD students trained in AI with those of a control sample of similar students with no AI training. We find that AI training is not associated with a higher probability of pursuing a research career after graduation. However, among students who have AI training during the PhD and pursue a research career after graduation, we observe a path dependence in continuing to publish on AI topics and a higher impact of their research. We also observe disciplinary heterogeneity. In Computer Science, AI-trained students are less likely to end up in private research organizations after graduation compared to their non-AI counterparts, while in disciplines other than Computer Science, AI training stimulates patenting activity and mobility abroad after graduation.

JEL codes: J24, O30

Keywords: Artificial Intelligence, Training, PhD students'careers,

1. Introduction

The growing importance of Artificial Intelligence (AI) has made the knowledge related to this technology a highly valuable asset for individuals competing for job positions (Squicciarini and Nachtigall 2021; Acemoglu et al. 2022; Lane et al. 2024). Extant studies have focused their attention on the demand side of the labor market, documenting companies' increasing interest in individuals with AI knowledge (Acemoglu and Restrepo 2018; Squicciarini and Nachtigall 2021; Acemoglu et al. 2022). However, little is known about the supply of individuals with AI knowledge.

Among multiple factors, educational attainment has been identified as a crucial factor in shaping the labor market supply (Goldin and Katz 2007; Furman 2018), especially in the AI context (Agrawal et al. 2019). Based on this assumption, countries have recently invested substantial resources in their higher education systems to develop AI education programs (OECD, 2024, Ch.4). Part of those resources has been devoted to training PhD students in AI at the frontier of science and technology.[1] PhD graduates are a relevant part of the high-skilled labour force and play a crucial role in the diffusion of knowledge, especially for up-to-date technologies like AI (Mangematin and Robin 2003; Cyranoski et al. 2011; Zolas et al. 2015; Arora et al. 2023).

This paper examines the relationships between students' AI training during their PhD period and their careers after graduation. Studying PhD graduates' careers allows us to tackle three fundamental issues that animate the current debate on the AI labor market. The first issue regards the effect of undertaking distinctive training, i.e., AI training, on the opportunity to take full advantage of graduates' qualifications by starting a research career (Cyranoski et al. 2011). The second issue regards the potential *brain drain* of AI-trained students from academia to the private sector due to the higher salaries and resources rewarding AI knowledge in the private sector (Jurowetzki et al. 2021; Ahmed et al. 2023). The third issue regards the impact of AI training on researchers' productivity after graduation (Bianchini et al. 2022).

In this study, we aim to reply to the following three questions: *Does AI training affect graduates' probability of starting a research career?* Conditional on starting a research career, *does AI*

[1] For instance, The *Agence Nationale de Recherche* (ANR) in France is co-financing 200 PhD contracts in AI at the national level in 2022: https://anr.fr/en/call-for-proposals-details/call/call-for-programmes-phd-contracts-in-artificial-intelligence/. Moreover, The *Interdisciplinary Institutes for Artificial Intelligence* (3IA) is securing support for research specialized centers in AI: https://instituts-3ia.fr/.

training affect graduates' probability of ending up in the private sector? *Does AI training affect graduates' scientific productivity?*

We investigate the three research questions across disciplines, distinguishing those where AI learning during the PhD program allows graduates to stand out, such as Biology, Chemistry, Engineering, Geology, Mathematics, Medicine, and Physics, from other disciplines where graduates' background may make them more likely to acquire AI knowledge even after completing the program.

We draw on a unique dataset covering the population of 35,492 French PhD students who graduated in STEM disciplines between 2010 and 2018 and identify students who acquired AI knowledge during their PhD training as those using AI-related keywords in their theses.

Studying the relationship between AI training and careers is challenging because students who acquire AI knowledge during their PhD training may differ from those who do not. The differences between the two groups may also impact their careers after graduation, potentially introducing biases in the estimated effect of AI training. We overcome this concern by implementing a Propensity Score Matching (PSM) approach. For each PhD student who undertakes AI training, we identify a 'control' student with similar characteristics but who does not undertake AI training.

We find that AI training is not significantly related to the probability of pursuing a research career after graduation, regardless of the discipline. However, AI-trained students in Computer Science who pursue a research career are less likely to end up in private research organizations after graduation than their non-AI counterparts. Conditional on starting a research career, we do not observe significant differences between AI- and non-AI-trained students in terms of their productivity after graduation, as measured by standard bibliometric indices such as the number of publications and co-authors. However, we observe that AI training, both in Computer Science and other disciplines, generates significant path dependence, as evidenced by the tendency to continue publishing on AI topics. Moreover, AI training is related to receiving a higher number of citations for the works published after graduation. Finally, AI training fosters patenting activity and graduates' mobility in disciplines other than Computer Science.

Our study contributes to the recent people-centric framework proposed to track the broader impact of AI on the economy and society (Lane et al. 2024; Shvadron et al. 2025). Specifically, we consider PhD students with AI training as "*people at the heart of [AI] investments*", and we investigate their careers to trace their "*economy-wide impact*" (Lane et al. 2024, 303). By doing so, we provide the first empirical evidence on the supply of AI workers in the labor market and, when investigating how training in AI affects PhDs' future careers, we embrace the idea that tracing the movement of highly trained individuals allows us to evaluate the impact of a substantial part of the public investments. Additionally, our empirical context benefits from the use of micro-level individual data from an entire country, enabling reliable estimates of the effects of undertaking AI training during the PhD.

2. Theoretical and empirical framework

Theoretical framework

Recent studies have shown that AI knowledge constitutes a highly valuable asset for the workforce. Using online job posting data, Squicciarini and Nachtigall (2021) and Acemoglu et al. (2022) show a relevant increase in AI-related job demand across all sectors since 2010. This trend encompasses not only AI-specific jobs, requiring advanced skills such as Natural Language Processing and Deep Learning skills, but also job positions based on soft and generic AI-related knowledge, such as communication, problem-solving, and creativity (Squicciarini and Nachtigall 2021).

While the literature has mainly focused on the demand for workers with AI knowledge by companies, the supply side has remained largely unexplored. Among the factors feeding the supply of workers with AI knowledge in the labor market, education plays a crucial role (Agrawal et al. 2019). Based on this assumption, governments worldwide have promoted AI education at all levels (Nestor et al. 2023; The AI Index 2023 Annual Report, Chapter 5). AI education evolved rapidly during our study period and had not yet integrated into the standard textbooks used in Bachelor's and Master's programs. Therefore, doctoral programs appear to be the most relevant educational programs for studying the impact of AI learning. Indeed, PhD students, trained as researchers, are expected to learn and develop frontier knowledge, such as AI, and disseminate it throughout society during their careers after graduation (Ahmed et al. 2023).

In this study, we investigate the probability that PhD graduates with AI training pursue a research career, which represents a key channel for the generation and diffusion of new knowledge in society. We expect that acquiring AI knowledge is positively related to the probability of pursuing a research career. This approach is aligned with recent policy evaluation work emphasizing the fact that "*Any attempt to describe the economy-wide impact of public investment in AI would involve identifying the people at the heart of these investments*" (Lane et al. 2024, 303).

While empirical evidence shows that almost 50% of graduates leave research careers within a decade (Naddaf 2024), the impact of PhD graduates' competencies in emerging and increasingly pervasive technologies like AI on this trend remains unknown.

In most advanced countries, PhD programs have traditionally been designed to train graduates to pursue careers as teachers and researchers at the university (OECD 2024, Ch. 1). Other career paths, such as the one in the private sector, have often been considered by graduates as secondary options (Sauerman and Stephan 2013). At the same time, universities have generated in the last decades an excess of PhD graduates (Cyranoski et al. 2011; Malloy et al. 2021; Sarrico 2022; Larson et al. 2024). The origin of this oversupply has been identified in the rise of the knowledge economy (Foray and Lundvall 1998) and the resulting public policies promoting higher education systems aimed at training a highly educated labor force (OECD 2021b).

These two forces, the graduates' willingness to pursue an academic career and the oversupply of graduates by universities, created intense competition for academic positions, exacerbating job insecurity and pushing many graduates to pursue careers outside academia (Bonnal and Giret 2010; Geuna and Shibayama 2015; Andalib et al. 2018; Sarrico 2022). Moreover, lower salaries in academia compared to the private sector make alternative careers more attractive (OECD 2021b; Sauermann and Roach 2014), causing many graduates to leave academia (Auriol et al. 2013; Geuna and Shibayama 2015; Waaijer et al. 2017). This is especially the case when the scientific and technological knowledge acquired during the PhD training is relevant for companies (Stokes 2011; Ahmed et al. 2023). Despite the role of acquiring specific knowledge relevant to companies on the type of career undertaken, the empirical evidence is limited, and no studies have assessed the effect of acquiring AI knowledge.

Therefore, in this study, we investigate how acquiring AI knowledge relates to the probability of starting a career as a researcher in the private sector. We expect that the saturation of academic positions and the demand from the private sector for AI knowledge will lead to a positive relationship between acquiring AI knowledge during the PhD and pursuing a research career in the private sector.

The emergence of a new research paradigm is the result of a process starting with the development of its theoretical foundations, followed by a second phase characterized by experimental work and the development of practical applications (Kuhn 1997). This process has characterized AI, in which theoretical foundations were developed since the 1950s, but experimental work and practical applications only became possible from the 1990s due to internet availability and the increase in hardware performance (Haenlein and Kaplan 2019). In recent years, AI applications in science and technology have played the role of a new "method of invention", reshaping the inventive process across all domains of science and technology (Cockburn et al. 2018; Bianchini et al. 2022; European Commission 2023). Applying AI to science and technology is expected to foster researchers' creativity and reduce the time needed to investigate new ideas (Chubb et al. 2022). Although an emerging literature is attempting to measure the effects of AI on science and technology production, empirical analyses remain limited.

Therefore, in this study, we investigate how acquiring AI knowledge relates to the researchers' productivity after graduation. We expect that the acquisition of AI knowledge during the PhD is positively related to the research outcome.

The careers of graduates who acquire AI knowledge during training are expected to vary substantially across disciplines. Indeed, the effect of learning about AI during the PhD depends on whether it becomes a distinctive factor for the graduate when competing in the labor market. In our study, we distinguish two groups of disciplines: Computer Science and all the other STEM disciplines. In Computer Science, the theoretical, methodological, and empirical knowledge developed during a PhD is expected to be similar to that required to easily absorb AI knowledge after graduation (Nelson and Winter 1982; Wesley M. Cohen and Levinthal 1990; Wesely M. Cohen and Levinthal 2000). Therefore, in Computer Science, not receiving formal AI training during the PhD does not hinder the ability to learn about AI afterward, making AI knowledge a less distinctive feature of the training. On the contrary, for PhD graduates in disciplines other than Computer Science, the theoretical, methodological, and empirical knowledge developed in their doctoral studies is often inadequate for easily assimilating AI knowledge post-graduation, making the PhD period a unique opportunity for acquiring that kind of knowledge. Therefore, in our analysis, we separate Computer Science from all the other disciplines, i.e., Biology, Chemistry, Engineering, Geology, Mathematics, Medicine, and Physics.

Empirical framework

We conduct our analysis for the period 2010-2018, including nine cohorts of French PhD graduates. This period is ideal for studying how the acquisition of AI knowledge influences graduates' careers. Indeed, it coincided with an oversupply of PhD graduates and the growing use of AI in science and technology, making AI a distinctive feature in a highly competitive labor market.

PhD students in France are selected through competitive application processes and are required to complete a high-level research thesis within three years (OECD, 2019). During the training period, students are expected to acquire a set of competencies that constitute essential components of their professional profile for entering the labor market. In the years covered by our study, career options for graduates have increasingly gone beyond academia to include research and non-research jobs in industry, public institutions, and self-employment (Cyranoski et al. 2011; Geuna and Shibayama 2015; Sarrico 2022; OECD 2023). In France, only 44% of PhD graduates in 2018 stayed in academia three years after graduation, while 17% moved to research careers in the private sector (Ministère de l'Enseignement Supérieur et de la Recherche 2024). This empirical evidence confirms that universities in France are producing more PhDs than the available academic jobs can absorb (Larivière 2011).

During the period covered by our analysis, a switch of graduates' careers from academia to industry has been observed across countries (OECD 2024). This is particularly true for graduates with specialized knowledge in high demand by the private sector. Empirical evidence of the *brain drain* from academia to the private sector of scientists with AI knowledge has been provided for the United States, confirming the industry's increasing interest in workers with this knowledge (Jurowetzki et al. 2021; Nestor et al. 2023).

The acquisition of AI knowledge during PhD training is a particularly important distinctive factor in the labor market during the period of our analysis, as AI began to be extensively applied in science and technology. Indeed, AI has experienced massive growth across disciplines since 2010, mainly due to advances in hardware development and data availability that have made it possible to give empirical ground to ideas and concepts theoretically conjectured by computer scientists for decades (WIPO 2019). For instance, the field of medicine observed a remarkable growth in the use of AI technologies that help physicians develop more precise prognoses (WIPO 2019, Ch. 1). In the same field, Deep Learning techniques have been rapidly expanding between 2013 and 2016 (WIPO 2019, Ch. 2), allowing physicians to analyze more effectively biological data and medical images.

The possibility of developing AI applications increased the industry's interest in the technology. Between 2013 and 2022, 338 AI startups were founded in France. Although this seems relevant, it remains limited when compared to other countries, such as the U.S. and China. In the U.S. and China, 4,643 and 1,337 startups were founded, respectively, and some of them became the AI giants that nowadays dominate the industry worldwide like OpenAI in the U.S and Sense Time in China (Nestor et al. 2023, The AI Index 2023 Annual Report, Ch. 4). In terms of public investments in AI, a large part of them took place in the last years of our study period, with limited impact on our analyses. The French government announced its national strategy for AI only in 2018, the last year of our study period, with an investment of $1.85 billion in AI research and development.[2] Similarly, the U.S. spent more than $2 billion in 2017 and 2018, and China plans to build a domestic AI industry worth $150 billion by 2030 (Shoham et al. 2018, The AI Index Annual Report 2018, p.58).

3. Data

3.1 Data sources

Our study sample merges four data sources to create a unique dataset on French PhD graduates in the STEM disciplines. We extract detailed information on doctoral dissertations from *L'Agence Bibliographique de l'Enseignement Supérieur* (ABES), the French public institute responsible for maintaining the bibliographic records of French universities. The extraction process allows us to gather a large collection of STEM theses from 2010 to 2018, including the student's name, graduation institution, defense year, supervisor name, thesis title and abstract, and scientific discipline.

To collect bibliometric information, we match the names of students and supervisors with the authors listed in the *OpenAlex*[3] database and assign the corresponding journal articles.[4] For each identified journal article, we retrieve its title and abstract, publication year, and author list. We collect information about research funding from the full list of individual grants awarded by the *Agence Nationale de la Recherche* (ANR)[5], the national funding agency in France.

For information on patents, we match the names of the PhD students and supervisors with the inventors' names listed on patent applications at the European Patent Office (EPO). For each patent, we retrieve the application filing year, the inventor's name and country, and the type of the applicant's institution.[6]

[2] https://www.inria.fr/en/french-national-artificial-intelligence-research-program

[3] OpenAlex is a free and open catalog of the global research system. It contains "metadata for 209M works (journal articles, books, etc.); 213M disambiguated authors; 124k venues (places that host works, such as journals and online repositories)" (Priem et al. 2022). Compared to other scientific databases like Elsevier's Scopus and Clarivate's Web of Sciences, OpenAlex has about twice the coverage, an open and fast API, and is completely free. Moreover, OpenAlex is updated weekly, ensuring relevant and up-to-date information.

[4] More details about the matching progress can be found in Appendix A.

[5] https://anr.fr/

[6] While bibliography and patent data do not capture individuals who do not publish or patent post-graduation, efforts to match a subset of PhD graduates in our dataset with LinkedIn profiles have demonstrated a low level of success, limiting the utility of such data beyond traditional publication and patent metrics.

After merging the four datasets mentioned above, we obtained a unique dataset including 35,492 PhD students who graduated from French institutions between 2010 and 2018. Our sample also includes 16,298 distinct supervisors.

3.2 Tracing AI knowledge

To identify the theses with AI content, we create a comprehensive list of AI keywords extracted from two recognized lists. The first list, containing 189 AI keywords, is taken from a recent OECD report aimed at developing a methodology for identifying AI-related publications and patents (Baruffaldi et al. 2020; Calvino et al. 2023). The second list, with 39 AI keywords, is extracted from a study aiming at identifying AI publications (Cockburn et al. 2018). After refining the list[7], we obtain 272 keywords. In our analysis, students trained in AI are the ones whose theses contain AI-related keywords. We calculate the corresponding variable, *AI student*, as equal to one if the student's thesis contains AI-related keywords in its title or abstract, zero otherwise. Overall, we identify 2,555 theses with AI content, and an equivalent number of *AI students*, which represent 7.2% of the 35,492 students in our study sample.

3.3 Tracing students' careers

To trace careers, we calculate the dummy variable *Research career,* which equals one if the student has at least one journal publication or patent application[8] within five years after the defense year, and zero otherwise.[9] We assume that students active in publishing or patenting are those who undertake a research career and are employed in universities, public or private research institutions, or companies' R&D departments. We consider the five years after the defense year (t) as the years elapsed between $t+2$ and $t+6$. By focusing on articles and patents from year $t+2$, we exclude outcomes from the thesis work that may be delayed in publication or filing, ensuring that we consider only post-graduation research outputs. . To assess the impact of AI training on PhD graduates who pursue a research career, we consider several outcomes measured within the same time frame, i.e., between $t+2$ and $t+6$. Specifically, we examine if they pursue a research career solely in private organizations. To do so, we define the variable *Private career* as a dummy variable that equals one if all their affiliations reported in the publication or patent documents published between $t+2$ and $t+6$ refer only to private

[7] We include the plural version of keywords that were initially presented in their singular form, and we drop those that are overly generic. More details can be found in Appendix B.
[8] We consider patent applications at the European Patent Office.
[9] In Appendix D, we validate our proxies for career outcomes by analyzing the information reported online for a subset of 100 PhD students randomly extracted from our original sample.

organizations. We also define the variable *Private career/collaboration* as a dummy variable that equals one if all the students' affiliations reported in the publication or patent documents published between *t+2* and *t+6* refer to at least one private organization. Finally, we measure the productivity of the students who undertake a research career using seven measures calculated between *t+2* and *t+6*: the total number of publications (*N. publications*), the number of AI-related publications (*N. AI publications*), if they hold at least one patent (*At least one patent)*, the number of coauthors they have (*N. of coauthors)*, if they have at least one international affiliations (*Abroad)*, if they have ties to institutions in the U.S. or China (*U.S. or China*), and the average number of citations received per paper (*Citations*).

3.4 Descriptive statistics

We distinguish the 35,492 PhD students who graduated in the STEM disciplines in France between 2010 and 2018 into two groups: graduates in Computer Science and graduates in other disciplines, i.e., Biology, Chemistry, Engineering, Geology, Mathematics, Medicine, and Physics. Table 1 shows a consistent annual increase in the share of AI theses in Computer Science. The percentage of AI theses reached 36.80% of total theses in 2018. During the same period, the share of AI theses in other STEM disciplines remains substantially lower but is gradually increasing, reflecting the growing use and integration of AI technologies across different disciplines (Jordan and Mitchell 2015). When comparing the two groups, graduates in Computer Science are less likely to pursue a research career compared to their peers from other STEM disciplines.

Table 1: Share of AI theses and share of students pursuing a research career by year

	Computer Science (3,999 theses)		*Other disciplines* (31,493 theses)	
Defense year	AI student	Research career	AI student	Research career
2010	23.34%	36.07%	4.23%	58.57%
2011	26.34%	40.05%	4.22%	56.38%
2012	29.78%	39.21%	3.95%	56.79%
2013	29.36%	37.75%	4.20%	56.25%
2014	24.50%	42.98%	4.52%	56.49%
2015	30.67%	38.66%	4.11%	56.75%
2016	29.91%	36.68%	4.57%	55.36%
2017	34.26%	35.45%	4.45%	55.53%
2018	36.80%	38.15%	4.61%	55.20%
All	29.81%	38.28%	4.33%	56.29%

To further explore these initial observations, Table 2 presents detailed descriptive statistics on career trajectories and productivity metrics for students who graduated with and without acquiring AI knowledge during the PhD. We observe that AI graduates and non-AI graduates do not show a significant difference in the probability of doing a research career in Computer Science, while AI graduates have a lower probability of pursuing a research career in other disciplines. In these other STEM disciplines, AI graduates also show a preference for collaboration with private research institutions or committing exclusively to private-sector research. We also observe that graduates with AI knowledge are likely to continue publishing in AI during their careers, both in Computer Science and other disciplines. Graduates with AI knowledge in other disciplines have a higher probability of patenting, but at the same time, have a smaller network. Finally, AI graduates in Computer Science appear to receive more citations than their non-AI counterparts. The differences observed between the two groups align with our hypothesis that the impact of acquiring AI knowledge differs across disciplines.

Table 2: Career type and productivity by AI thesis content and discipline

	Non-AI students	AI Students		Non-AI students	AI students
Computer Science (3,999 students)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.38	0.39	0.64	2,807	1,192
Conditional on pursuing a research career					
Private career/collaboration	0.20	0.16	0.10	1,068	463
Private career	0.10	0.08	0.12	1,068	463
Research career productivity					
N. of publications	3.96	3.92	0.90	1,068	463
At least one patent	0.11	0.10	0.34	1,068	463
N. AI publications	0.56	1.43	0.00	1,068	463
N. of coauthors	19.62	20.12	0.95	1,068	463
Abroad	0.48	0.47	0.72	1,068	463
U.S. and China	0.12	0.14	0.44	1,068	463
Citations	10.63	16.88	0.04	1,068	463
Other disciplines (31,493 students)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.57	0.51	0.00	30,130	1,363
Conditional on pursuing a research career					
Private career/collaboration	0.14	0.20	0.00	17,032	696
Private career	0.05	0.09	0.00	17,032	696
Research career productivity					
N. of publications	5.87	5.45	0.11	17,032	696
At least one patent	0.09	0.14	0.00	17,032	696
N. AI publications	0.15	1.21	0.00	17,032	696
N. of coauthors	53.00	28.53	0.00	17,032	696
Abroad	0.48	0.50	0.24	17,032	696
U.S. and China	0.15	0.17	0.26	17,032	696
Citations	16.08	15.22	0.34	17,032	696

NOTE: All career-related variables are calculated in *[t+2, t+6]*, where *t* is the thesis defense year.

4. Methodology

4.1 Estimation strategy

Although Table 2 informs us about the relationship between acquiring AI knowledge during the PhD and subsequent career outcomes, it does not allow us to identify an unbiased relationship between the two. Indeed, it is possible that students who acquire AI knowledge during their

PhD training have different characteristics compared to those who do not. These differences might also influence their careers. To mitigate potential biases, we implement a Propensity Score Matching (PSM) to estimate the treatment effect of acquiring AI knowledge by comparing treated individuals with untreated counterparts who are as similar as possible in their observable characteristics. We proceed in two steps.

First, for the entire population of French PhD students in STEM, we estimate a logit model predicting the probability of observing an AI student, based on observable characteristics (Equation 1).

$$\text{AI student}_i = \beta_0 + \beta_1 X_i + \beta_2 L_i + \beta_3 D_i + \varepsilon_i$$

Equation 1: Probability of observing an AI student

The predicted probability allows us to compare the career outcome of each student with AI training, i.e., the treated, with that of a non-AI student with a similar probability of showing AI training, i.e., the control.[10] We consider as relevant observable characteristics used to calculate the probability of receiving AI training the supervisor characteristics (X_i), location of the university (L_i), defense year (D_i), and discipline (F_i). Then, for this first analysis, we compare the probability of starting a research career (*Research career*) for AI students and the matched non-AI students.

Second, for the sample of PhD students who start a research career (*Research career* = 1), we estimate a logit model predicting the probability of observing an AI student, based on observable characteristics (Equation 2).

$$\text{AI student}_i|_{\text{Research career}_i=1} = \gamma_0 + \gamma_1 X_i + \gamma_2 L_i + \gamma_3 D_i + \tau_i$$

Equation 2: Probability of observing an AI student, conditional on starting a research career

Similarly to the previous matching exercise (Equation 1), the probability predicted using the estimates of Equation 2 allows us to compare the career outcomes of students with AI training with those of a control group of non-AI students with a similar probability of showing AI

[10] We use the *Nearest Neighbor* approach *without replacement* to match the treated and control groups. This technique ensures that each control unit is matched exclusively to one treated unit based on the closest propensity score, avoiding potential bias from reusing controls (Smith 1997). To improve our matching exercise, we perform an exact matching on the disciplines.

training based on a set of observable characteristics, i.e., X_i, L_i, D_i, and F_i as in Equation 1. As career outcomes, conditional on having started a research career, we consider the variables *Private career, Private career/collaboration, N. of publications*, *N. AI publications*, *At least one patent*, *N. of coauthors*, *Abroad*, *U.S. or China*, and *Citations*.

4.2 Student and supervisor characteristics

This section describes the set of observable characteristics used in Equations 1 and 2 to estimate the probability of observing an AI student, which allows us to match AI students with a similar control group of non-AI students.

As supervisor characteristics (X_i), we include *Mentorship experience*, which we calculate by counting the total number of the supervisor's students who successfully defended their thesis from 1990 until the focal student's enrollment. To measure the supervisor's academic output, we consider all their publications from the beginning of their career up to the enrollment of the focal student.[11] We then define a *High productive supervisor* as one who ranks in the top 25% for productivity. We also identify an *AI supervisor before student enrollment* as one who has at least one AI-related publication before the enrollment of the student. A *Supervisor with a private affiliation* is defined as a supervisor who has at least one affiliation with a private organization in their publications or patents. *PI supervisor* is a dummy variable that equals one if the supervisor is the principal investigator (PI) of an ANR[12] grant before the student's enrollment in the PhD program.

We also consider the location (L_i) of the university where both the student and the supervisor are based. To do so, we calculate dummy variables grouping universities from major French cities. For example, universities in Paris are aggregated into one group, represented by the *Paris* dummy variable. We create ten dummy variables for the ten cities that host the majority of the PhD students in France, with all other institutions that are not located in these cities falling under the residual category *Other*. In our analysis, we use the dummy variable *Paris* as our reference group. Regarding the *Defense year* (D_i), we introduce a set of dummy variables that denote the years when students defended their theses. The dummy variable *Defense year 2010*

11 Despite having comprehensive data on PhD graduates, information about the students' entry years into their PhD programs was missing. To address this, we estimate the entry year to be three years prior to their defense year. According to the national statistics for STEM disciplines in France (Pommier et al. 2022), the typical duration of PhD training is three years of research plus one additional year dedicated to the thesis defense. Therefore, we define the PhD training period as spanning from *t-3* to *t*, where *t* represents the defense year.

12 ANR is the main funding agency in France.

is our reference group. Lastly, we define eight dummy variables representing the major disciplines (F_i) of study in which the supervisor mentors the student. These disciplines include *Mathematics, Engineering, Computer Science, Physics, Geology, Medicine, Biology,* and *Chemistry.*

4.3 Finding a control sample

This section is in two parts. The first part aims to find a control sample of non-AI students in the entire population of French students. The second part aims to find a control sample of non-AI students among the students who started a research career after graduation. In both parts, we estimate the respective probability of being trained in AI during the PhD, i.e., propensity score, and assess the quality of our matching exercises to ensure that each AI student is matched with a similar non-AI student.

4.3.1 Matching AI students with a control sample of non-AI students in the entire population of French students

Table 3 reports the estimates of Equation 1 used to predict the propensity scores. The results suggest that students supervised by highly productive supervisors, supervisors with private affiliations, and principal investigators on research grants are less likely to write an AI thesis. On the other hand, students are more likely to develop an AI thesis if their supervisor has AI experience before their enrollment and has more supervision experience. We also control for the graduation location and year.

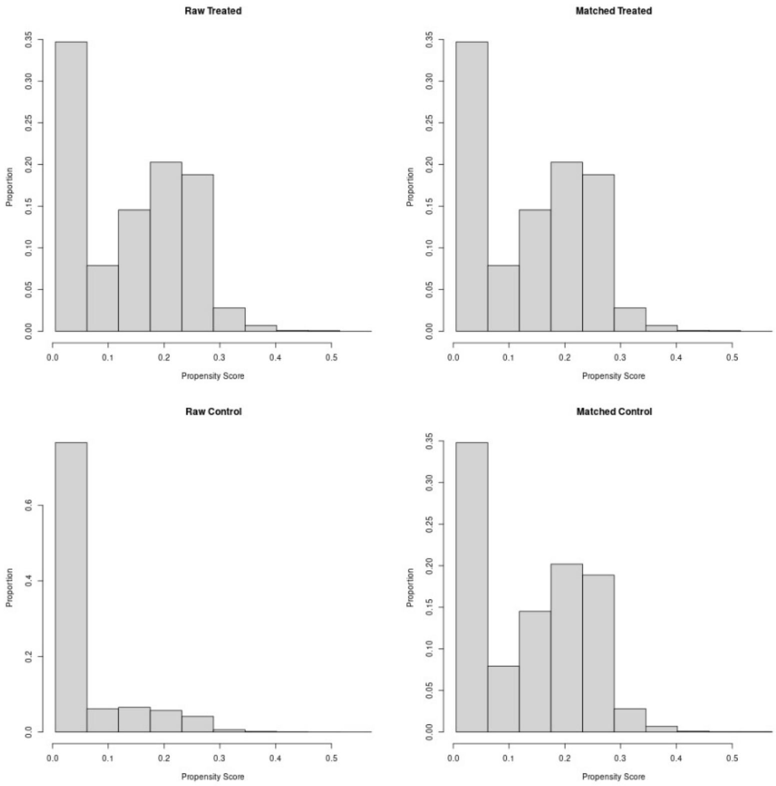
Table 3: PSM equation, probability of writing an AI thesis in the entire population of French students

Dependent variable: AI student	Estimate	Std. Error	Statistic	P-value
(Intercept)	-3.14***	0.09	-35.75	0.00
High productive supervisor	-0.63***	0.06	-11.23	0.00
AI supervisor before student enrollment	1.88***	0.04	42.13	0.00
Supervisor with a private affiliation	-0.37***	0.06	-6.51	0.00
PI supervisor	-0.28***	0.06	-4.61	0.00
Mentorship experience	0.09***	0.01	9.59	0.00
Bordeaux	-0.44***	0.14	-3.24	0.00
Lille	-0.03	0.12	-0.26	0.80
Lyon	-0.29***	0.10	-2.96	0.00
Marseille	-0.54***	0.15	-3.52	0.00
Montpellier	-0.32**	0.13	-2.40	0.02
Nantes	0.05	0.15	0.32	0.75
Others	0.18***	0.05	3.40	0.00
Rennes	-0.10	0.12	-0.81	0.42
Strasbourg	-0.77***	0.17	-4.47	0.00
Toulouse	0.07	0.09	0.75	0.45
Defense year 2011	-0.04	0.10	-0.34	0.74
Defense year 2012	-0.05	0.10	-0.49	0.63
Defense year 2013	-0.05	0.10	-0.55	0.58
Defense year 2014	-0.13	0.10	-1.32	0.19
Defense year 2015	-0.12	0.10	-1.24	0.22
Defense year 2016	-0.12	0.10	-1.17	0.24
Defense year 2017	-0.08	0.10	-0.79	0.43
Defense year 2018	-0.01	0.10	-0.05	0.96
Pseudo R-squared	0.12			
Number of Observations	35,492			

NOTE: The dummy variables *Paris* and *Defense year 2010* are the reference groups. The coefficients reported in the table are the coefficients of the Logit model. Discipline dummy variables are not included because we set an "exact matching" on the disciplines.

We use the predictions from the logit model estimated in Table 3 to calculate the probability of observing an AI student for both the the treated group (AI student), and the control group (non-AI students). As shown in Figure 1, the distribution of the propensity scores of the AI students and the non-AI students is different before the matching (*Figure 1, Raw treated* versus *Figure 1, Raw control*), but becomes very similar after the matching (*Figure 1, Matched treated* versus *Figure 1, Matched control*), indicating a high-quality matching exercise.

Figure 1: Propensity score matching distribution before and after the matching exercise in the entire population of French students



To further assess the quality of our matching exercise, Table 4 presents the average characteristics of the two groups of AI and non-AI-matched students after the matching exercise. Specifically, we compare the average characteristics of the 2,555 AI students with the average characteristics of the control group of 2,555 non-AI students. Table 4 reveals no statistically significant differences between the two groups after the matching exercise (the p-values of the tests of the difference between averages are always above the standard significance threshold of 0.1). The only exception is the variable *PI supervisor,* for which both groups show a weak (0.13 vs 0.15), but statistically significant difference (p-value 0.06).

Table 4: PSM balancing in the entire population of French students

	Non-AI students (2,555)	AI Students (2,555)	P-value
High productive supervisor	0.21	0.21	0.73
AI supervisor before student enrollment	0.63	0.63	0.82
Supervisor with a private affiliation	0.16	0.17	0.31
PI supervisor	0.13	0.15	0.06
Mentorship experience	1.79	1.89	0.20
Bordeaux	0.03	0.03	0.67
Lille	0.04	0.04	0.61
Lyon	0.05	0.05	0.28
Marseille	0.02	0.02	0.29
Montpellier	0.03	0.03	1.00
Nantes	0.02	0.02	0.85
Others	0.45	0.43	0.19
Paris	0.25	0.25	0.54
Rennes	0.04	0.04	0.51
Strasbourg	0.01	0.01	1.00
Toulouse	0.08	0.08	0.64
Defense year 2010	0.09	0.08	0.14
Defense year 2011	0.09	0.09	0.81
Defense year 2012	0.10	0.10	0.64
Defense year 2013	0.11	0.11	0.72
Defense year 2014	0.11	0.11	0.72
Defense year 2015	0.12	0.12	0.97
Defense year 2016	0.12	0.12	0.70
Defense year 2017	0.13	0.14	0.39
Defense year 2018	0.13	0.14	0.62

NOTE: Discipline dummy variables are not included because we set an “exact matching” on the disciplines.

4.3.2 Matching AI students with a control sample of non-AI students, conditional on pursuing a research career

Table 5 reports the estimates of Equation 2 used to predict the propensity scores. The sample size in this analysis decreased from 35,492 to 19,259 students because we selected only those students who started a research career. Although Equation 2 is estimated on a different sample from that of Equation 1, the findings in Table 5 are similar to those reported in Table 4. Students are less likely to write an AI thesis if they are supervised by highly productive supervisors, supervisors with affiliations to private institutions, or PIs on research grants. However, the probability increases for students whose supervisors have prior AI experience or more supervision experience. We also control for the graduation location and year.

Table 5: PSM equation. Probability of writing a thesis in AI, conditional on starting a research career

Dependent variable: AI thesis	Estimate	Std. Error	Statistic	P-value
(Intercept)	-3.44***	0.13	-26.19	0.00
High productive supervisor	-0.64***	0.08	-7.92	0.00
AI supervisor before student enrollment	1.99***	0.07	30.15	0.00
Supervisor with a private affiliation	-0.34***	0.08	-4.01	0.00
PI supervisor	-0.30***	0.09	-3.36	0.00
Mentorship experience	0.09***	0.01	6.65	0.00
Bordeaux	-0.41**	0.19	-2.14	0.03
Lille	0.11	0.17	0.65	0.52
Lyon	-0.18	0.14	-1.29	0.20
Marseille	-0.50**	0.22	-2.27	0.02
Montpellier	-0.39*	0.20	-1.97	0.05
Nantes	0.12	0.22	0.54	0.59
Others	0.24***	0.08	2.94	0.00
Rennes	-0.11	0.18	-0.63	0.53
Strasbourg	-0.75***	0.24	-3.06	0.00
Toulouse	0.20	0.13	1.57	0.12
Defense year 2011	0.03	0.15	0.21	0.83
Defense year 2012	0.04	0.15	0.29	0.77
Defense year 2013	-0.03	0.15	-0.18	0.86
Defense year 2014	-0.17	0.15	-1.14	0.26
Defense year 2015	-0.13	0.15	-0.86	0.39
Defense year 2016	-0.07	0.15	-0.47	0.64
Defense year 2017	0.06	0.14	0.42	0.67
Defense year 2018	-0.01	0.15	-0.08	0.94
Pseudo R-squared	0.13			
Number of Observations	19,259			

NOTE: The dummy variables *Paris* and *Defense year 2010* are the reference groups. The coefficients reported in the table are the coefficients of the Logit model. Discipline dummy variables are not included because we set an "exact matching" on the disciplines.

As in the previous matching exercise, Figure 2 shows similar propensity score distributions across groups, i.e., the treated group, and non-AI students, i.e., the control group, after the matching *(Figure 2, Matched treated* versus *Figure 2, Matched control)*. Moreover, Table 6 shows no statistically significant differences between the 1,159 AI students and the similar 1,159 non-AI students.

Figure 2: Propensity score matching distribution before and after the matching exercise, conditional on starting a research career after graduation

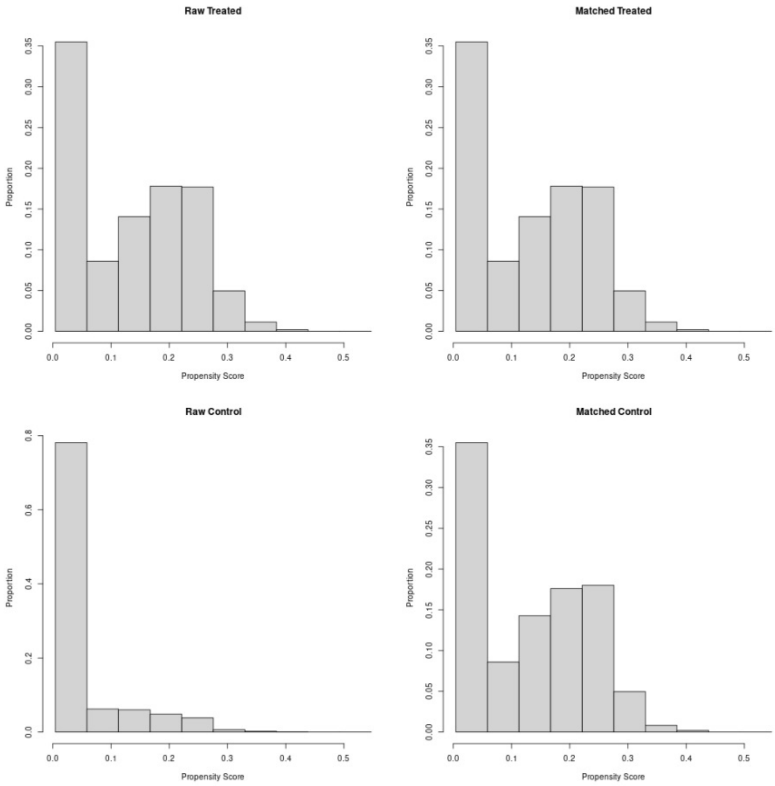


Table 6: PSM balancing, conditional on starting a research career

	Non-AI Students (1,159)	AI Students (1,159)	P-value
High productive supervisor	0.24	0.22	0.15
AI supervisor before student enrollment	0.64	0.64	0.80
Supervisor with a private affiliation	0.17	0.18	0.51
PI supervisor	0.13	0.16	0.12
Mentorship experience	1.90	1.77	0.23
Bordeaux	0.03	0.03	0.80
Lille	0.04	0.04	0.34
Lyon	0.06	0.06	0.66
Marseille	0.02	0.02	0.44
Montpellier	0.02	0.03	0.69
Nantes	0.02	0.02	1.00
Others	0.42	0.42	0.90
Paris	0.25	0.25	0.96
Rennes	0.04	0.03	0.28
Strasbourg	0.02	0.02	1.00
Toulouse	0.09	0.08	0.88
Defense year 2010	0.08	0.08	0.70
Defense year 2011	0.08	0.09	0.51
Defense year 2012	0.10	0.10	0.95
Defense year 2013	0.11	0.11	0.95
Defense year 2014	0.09	0.10	0.36
Defense year 2015	0.12	0.12	0.80
Defense year 2016	0.13	0.12	0.35
Defense year 2017	0.13	0.14	0.34
Defense year 2018	0.14	0.13	0.40

NOTE: Discipline dummy variables are not included because we set an "exact matching" on the disciplines.

5. Results

In this section, we present the results of our empirical analysis investigating the effects of acquiring AI knowledge on PhD graduates' career outcomes.

Table 7 shows that having AI knowledge does not significantly affect the probability of starting a research career (*Research career*), both in *Computer Science* and in *other disciplines*. However, conditional on starting a research career, AI-trained graduates in *Computer Science* only are 3 and 5 percentage points less likely to work in private organizations (*Private career*) or collaborate with private organizations (*Private career/collaboration*) than their non-AI counterparts, respectively.

When looking at productivity metrics, we find that acquiring AI knowledge during the PhD does not significantly affect the overall publications count or their network size. On the other hand, we observe that AI-trained graduates publish on average 0.78 more papers with AI content in *Computer Science*, and 0.94 more in *other disciplines,* compared to their non.AI counterparts. Moreover, their papers appear to attract 7.04 and 3.12 more citations per paper on average than those of their non-AI peers in *Computer Science* and *other disciplines,* respectively. Concerning the patenting activity, we find that only AI graduates in *other disciplines* have a 4-percentage point higher probability of patenting than their non-AI counterparts. For graduates in *other disciplines* only, having AI training during the PhD increases the probability of having a career abroad (*Abroad*) by 5 percentage points, especially with U.S. and Chinese affiliations (4 percentage points).

Table 7: Average career outcomes of AI and non-AI students after the Propensity Score Matching

	Non-AI students	AI Students		Non-AI students	AI students
Computer Science (2,384 students)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.38	0.39	0.56	1,192	1,192
Conditional on pursuing a research career					
Private career/collaboration	0.21	0.16	0.08	463	463
Private career	0.11	0.08	0.07	463	463
Research career productivity					
N. of publications	4.35	3.92	0.30	463	463
At least one patent	0.11	0.10	0.39	463	463
N. AI publications	0.65	1.43	0.00	463	463
N. of coauthors	28.18	20.12	0.58	463	463
Abroad	0.48	0.47	0.74	463	463
U.S. and China	0.11	0.14	0.16	463	463
Citations	9.84	16.88	0.02	463	463
Other disciplines (2,726 students)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.52	0.51	0.76	1,363	1,363
Conditional on pursuing a research career					
Private career/collaboration	0.19	0.20	0.64	696	696
Private career	0.08	0.09	0.39	696	696
Research career productivity					
N. of publications	5.22	5.45	0.54	696	696
At least one patent	0.10	0.14	0.05	696	696
N. AI publications	0.27	1.21	0.00	696	696
N. of coauthors	39.63	28.53	0.41	696	696
Abroad	0.45	0.50	0.05	696	696
U.S. and China	0.13	0.17	0.08	696	696
Citations	12.10	15.22	0.00	696	696

NOTE: All career-related variables are calculated in *[t+2, t+6]*, where *t* is the thesis defense year.

6. Discussion

Our analysis does not support the idea of a *brain drain* from the public to the private research sector discussed in the literature (Jurowetzki et al. 2021; Ahmed et al. 2023). One possible explanation is the relatively limited demand for AI talent in the French private sector during the period of our study, i.e., 2010-2018. Indeed, other countries such as the U.S. and China dominate the AI field worldwide in terms of tech giants and AI startups with 4,643 and 1,337 startups, respectively, compared to only 338 AI startups in France between 2013 and 2022 (Nestor et al. 2023, The AI Index 2023 Annual Report, Ch. 4). This disparity in private sector opportunities created a high demand for AI talents in the U.S. and China. Our data support this latter evidence by showing a positive correlation (0.18) between working in the private sector and pursuing a research career in the U.S. or China after graduation.

The lack of *brain drain* of AI students from the public to the private sector is particularly evident in Computer Science. Indeed, AI students in Computer Science show a significantly lower probability of working in the private sector after graduation than their non-AI counterparts. One possible explanation is the nature of the AI research in the period covered by our analysis, i.e., 2010-2018. During this period, computer scientists were probably focusing on designing the foundational architecture of AI, rather than producing AI applications. Indeed, hardware and data needed to develop applications were becoming available only during the study period. On the contrary, in disciplines other than computer science, PhD students' work was more likely to focus on applications in their discipline of AI techniques. Therefore, the lack of focus on developing applications has made AI students in Computer Science less valuable to companies seeking profitable technological applications, reducing their presence in the private sector. Our empirical evidence shows that AI students in Computer Science file fewer patents than their AI counterparts in other disciplines (10% vs. 14%, p-value < 5%), supporting the interpretation that AI graduates in Computer Science were less focused on technological applications of AI during our study period with respect to AI students in other disciplines.

In terms of research productivity, we observe a path dependency between the doctoral training in AI and their subsequent research focus, consistent with the literature showing that researchers tend to continue working in similar areas after their graduation (Shibayama 2019).

Contrary to our expectations and despite the growing interest in AI during the period of our analysis, we find no significant difference in the number of publications between AI and non-AI graduates. One would expect productivity advantages associated with AI due to its innovative potential (Cockburn et al. 2018). However, its novelty may cause resistance within the scientific community (Azoulay et al. 2019), potentially hurting productivity. These opposing effects likely cancel each other out, resulting in no significant differences in publication volume. The same mechanism applies to the graduates' network size. Consequently, AI graduates do not appear to have any relevant advantage over their non-AI peers, encountering equal opportunities to pursue a research career after graduation.

On the other hand, AI graduates in Computer Science and other disciplines receive significantly more citations, mainly due to the rapid global growth of the AI research community over the past decade (Bianchini et al. 2022). The expansion of the AI community creates a larger number of researchers working on AI-related topics, leading to higher citation rates for AI papers. This growth can also be seen in Table 1, columns 2 and 4, showing the growing number of PhD graduates with AI knowledge in France, especially in Computer Science.

7. Conclusion

In this paper, we study the career paths of 35,492 French PhD students who graduated in the STEM disciplines between 2010 and 2018. Specifically, we examine how acquiring AI knowledge during the PhD influences the likelihood of pursuing a research career after graduation. Moreover, conditional on pursuing a research career, we investigate how AI training affects the probability of working in private research organizations after graduation and how it affects the graduates' productivity, scientific networks, and mobility.

We start by identifying AI students as those who used AI-related keywords in their thesis abstracts or titles. Then, we trace the PhD graduates' career trajectories by analyzing their publications and patent documents.

Implementing a Propensity Score Matching, we find that AI training does not significantly affect the probability of starting a research career after graduation across all disciplines. However, AI-trained students in Computer Science are less likely to pursue a career in private research organizations compared to their non-AI counterparts. While standard productivity metrics show no significant differences between AI and non-AI trained graduates, those who acquire AI knowledge tend to continue publishing on AI topics and receive more citations regardless of the discipline. Additionally, AI training fosters patenting activity and promotes international mobility for non-Computer Science graduates.

Our paper also adds to the literature that explores the effects of AI on scientific research. Specifically, we provide new insights into how the acquisition of AI knowledge impacts early-stage scientists' productivity. Moreover, our contribution is methodological. We show how implementing an identification strategy that estimates the unbiased effect of AI training might change the evidence provided by descriptive statistics. For instance, according to our descriptive statistics, researchers in Computer Science with AI training who pursue a research career are as likely as their non-AI counterparts to work in private research organizations after graduation. This descriptive evidence is challenged when we apply a PSM approach. Indeed, the PSM approach shows that researchers in Computer Science with AI training are less likely than their non-AI counterparts to work in private research organizations after graduation. This latter result contradicts the idea of a brain drain of AI talents from public to private research organizations proposed in other works (Jurowetzki et al. 2021; Ahmed et al. 2023).

Our findings have important policy implications. The substantial public investments in training AI PhD students appear to have positive returns. First, we show that those graduates keep working on AI-related topics after graduation, serving as knowledge disseminators within their research communities. The impact of this knowledge transfer is evident in the significantly higher citation rate these graduates receive. This suggests that public funds were effectively used to diffuse AI knowledge. However, our results also show that other countries are attracting non-Computer Science AI graduates trained in French universities. In order to retain AI talent, policymakers in France should invest in creating more domestic job opportunities that align with the graduates' specialized training. Such investment is particularly urgent as computer scientists are likely to shift focus from theoretical contributions to more applied work in the near future.

Our paper is not without limitations. The first limitation concerns our proxies for career paths based on publication and patent information. Although our proxies seem to be in line with the career paths reconstructed using online sources, future studies might want to consider different proxies for career paths, such as conducting surveys on a subset of graduates. The second limitation regards the use of keywords to identify AI-related documents. Although this method is widely used in literature, a future extension of this work is to use machine learning techniques to identify AI content. The third limitation concerns the missing information about the careers of PhD graduates who do not publish or patent after graduation. Although we know from official statistics that 92% of PhD graduates in France are employed (Ministère de l'Enseignement Supérieur et de la Recherche 2024), we cannot trace the careers of those who leave research and stop publishing or patenting.

References

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American economic review*, *108*(6), 1488-1542.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293-S340.

Aghion, P., Dewatripont, M., & Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics*, 39(3), 617-635.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). Economic policy for artificial intelligence. *Innovation policy and the economy*, *19*(1), 139-159.

Andalib, M. A., Ghaffarzadegan, N., & Larson, R. C. (2018). The postdoc queue: A labour force in waiting. *Systems research and behavioral science*, *35*(6), 675-686.

Arora, A., Belenzon, S., Cioaca, L. C., Sheer, L., & Zhang, H. (2023). The effect of public science on corporate R&D. *NBER Working Paper Series*, no. w31899. Cambridge, Mass: National Bureau of Economic Research.

Auriol, L. (2010). Careers of Doctorate Holders: Employment and Mobility Patterns. *OECD Science, Technology and Industry Working Papers*, No. 2010/04, OECD Publishing, Paris. https://doi.org/10.1787/5kmh8phxvvf5-en.

Auriol, L., M. Misu and R. Freeman (2013), “Careers of Doctorate Holders: Analysis of Labour Market and Mobility Indicators”, OECD Science, Technology and Industry Working Papers, No. 2013/04, OECD Publishing, Paris, https://doi.org/10.1787/5k43nxgs289w-en.

Azoulay, P., Fons-Rosen, C., & Zivin, J. S. G. (2019). Does science advance one funeral at a time?. *American Economic Review*, *109*(8), 2889-2920.

Baruffaldi, S., van Beuzekom, B., Dernis, H., Harhoff, D., Rao, N., Rosenfeld, D., & Squicciarini, M. (2020). Identifying and measuring developments in artificial intelligence: Making the impossible possible. *OECD Science, Technology and Industry Working Papers*, No. 2020/05, OECD Publishing, Paris, https://doi.org/10.1787/5f65ff7e-en.

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

Bonnal, L., & Giret, J. F. (2010). Determinants of access to academic careers in France. *Economics of Innovation and New Technology*, *19*(5), 437-458.

Calvino, F., Criscuolo, C., Dernis, H., & Samek, L. (2023). What technologies are at the core of AI?: An exploration based on patent data. *OECD Artificial Intelligence Papers*, No. 6, OECD Publishing, Paris, https://doi.org/10.1787/32406765-en.

Chubb, J., Cowling, P., & Reed, D. (2022). Speeding up to keep up: exploring the use of AI in the research process. *AI & society*, 37(4), 1439-1457.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). The impact of artificial intelligence on innovation. *NBER Working Paper Series*, no. w24449. Cambridge, Mass: National Bureau of Economic Research.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, 35(1), 128-152.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., & Yahia, M. (2011). Education: the PhD factory. *Nature* 472 (7343): 276–80.

European Commission: Directorate-General for Research and Innovation, Arranz, D., Bianchini, S., Di Girolamo, V. and Ravet, J., Trends in the use of AI in science – A bibliometric analysis, Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2777/418191

Foray, D., & Lundvall, B. Å. (2009). The knowledge-based economy: from the economics of knowledge to the learning economy. In *The economic impact of knowledge* (pp. 115-121). Routledge.
Furman, J. (2018). Should we be reassured if automation in the future looks like automation in the past?. *The economics of artificial intelligence: An agenda*, *317*, 328.
Geuna, A., & Shibayama, S. (2015). Moving out of academic research: Why do scientists stop doing research?. In *Global mobility of research scientists* (pp. 271-303) , edited by Aldo Geuna, 271–303. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-801396-0.00010-7.
Goldin, C., & Katz, L. F. (2007). The race between education and technology: The evolution of US educational wage differentials, 1890 to 2005. Working Paper. *Working Paper Series. National Bureau of Economic Research*. https://doi.org/10.3386/w12984.
Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4), 5-14.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.
Jurowetzki, R., Hain, D., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). The Privatization of AI Research (-ers): Causes and Potential Consequences--From university-industry interaction to public research brain-drain?. *arXiv preprint arXiv:2102.01648*.
Kuhn, T. S. (1997). *The structure of scientific revolutions* (Vol. 962). Chicago: University of Chicago press.
Lane, J., Owen-Smith, J., & Weinberg, B. A. (2024). How to track the economic impact of public investments in AI. *Nature*, *630*(8016), 302-304.
Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, *90*(2), 463-481.
Larson, R. C., Ghaffarzadegan, N., & Xue, Y. (2014). Too many PhD graduates or too few academic job openings: The basic reproductive number R0 in academia. *Systems research and behavioral science*, 31(6), 745-750.
Malloy, J., Young, L., & Berdahl, L. (2021). PhD oversupply: The system is the problem. *Inside Higher Ed*, 22.
Mangematin, V., & Robin, S. (2003). The two faces of PhD students: Management of early careers of French PhDs in life sciences. *Science and public policy*, *30*(6), 405-414.
Ministère de l'Enseignement supérieur et de la Recherche. 2024. 'L'État de l'Enseignement Supérieur, de La Recherche et de l'Innovation En France'.
Naddaf, M. (2024). Nearly 50% of researchers quit science within a decade, huge study reveals. *Nature*. https://doi.org/10.1038/d41586-024-03222-7.
Nelson, R. R., & Winter, S. G. (1985). *An evolutionary theory of economic change*. Digitally reprinted. Cambridge, Mass.: The Belknap Press of Harvard Univ. Press.
Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J.C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.
OECD (2019). *Education at a Glance 2019: OECD Indicators*, OECD Publishing, Paris, https://doi.org/10.1787/f8d7880d-en.
OECD (2021a). *OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity.* OECD Publishing, Paris, https://doi.org/10.1787/75f79015-en.
OECD (2021b). Reducing the precarity of academic research careers. *OECD Science, Technology and Industry Policy Papers*, No. 113, OECD Publishing, Paris, https://doi.org/10.1787/0f8bd468-en.

OECD (2023). Promoting diverse career pathways for doctoral and postdoctoral researchers. *OECD Science, Technology and Industry Policy Papers*, No. 158, OECD Publishing, Paris, https://doi.org/10.1787/dc21227a-en.

OECD (2024). The state of academic careers in OECD countries: An evidence review. *OECD Education Policy Perspectives*, No. 91, OECD Publishing, Paris, https://doi.org/10.1787/ea9d3108-en.

OECD (2024). *Training Supply for the Green and AI Transitions: Equipping Workers with the Right Skills, Getting Skills Right*, OECD Publishing, Paris, https://doi.org/10.1787/7600d16d-en.

Pommier, S., Talby, M., Auffray-Seguette, M., Dalaut, M., Eijsberg, H., Elshawish, P., & Muller, H. (2022). Le doctorat en France Regards croisés sur la formation doctorale (Doctoral dissertation, Réseau National de Collèges Doctoraux).

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

Pyatt, G. (1966). Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. *The Economic Journal*, 76 (303): 635. https://doi.org/10.2307/2229541.

Sarrico, C. S. (2022). The expansion of doctoral education and the changing nature and purpose of the doctorate. *Higher Education*, 84(6), 1299-1315.

Sauermann, H., & Roach, M. (2014). Not all scientists pay to be scientists: PhDs' preferences for publishing in industrial employment. *Research Policy*, 43(1), 32-47.

Sauermann, H., & Stephan, P. (2013). Conflicting logics? A multidimensional view of industrial and academic science. *Organization science*, 24(3), 889-909.

Shibayama, S. (2019). Sustainable development of science and scientists: Academic training in life science labs. *Research Policy*, *48*(3), 676-692.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological methodology*, *27*(1), 325-353.

Squicciarini, M., & Nachtigall, H. (2021). Demand for AI skills in jobs: Evidence from online job postings. *OECD Science, Technology and Industry Working Papers*, *2021*(3), 1-74.

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Vannuccini, S., & Prytkova, E. (2024). Artificial Intelligence's new clothes? A system technology perspective. *Journal of Information Technology*, 39(2), 317-338.

Waaijer, C. J., Belder, R., Sonneveld, H., van Bochove, C. A., & van der Weijden, I. C. (2017). Temporary contracts: Effect on job satisfaction and personal lives of recent PhD graduates. *Higher education*, 74, 321-339.

WIPO (2019). WIPO Technology Trends 2019: Artificial Intelligence. Geneva: World Intellectual Property Organization.

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J.C., Lyons, T., Etchemendy, J., Grosz, B., and Bauer, Z. (2018). *The AI Index 2018 Annual Report. AI Index*. Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA, December 2018.

Shvadron, D., Zhang, H., Fleming, L., and Gross, Daniel P. Funding the U.S. Scientific Training Ecosystem: New Data, Methods, and Evidence (June 2025). NBER Working Paper No. w33944.

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Smith, J. O., Rosen, R. F., ... & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph. D. recipients. Science, 350(6266), 1367-1371.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., ... & Dinlersoz, E. (2021). *Advanced technologies adoption and use by us firms: Evidence from the annual business survey* (No. w28290). National Bureau of Economic Research.

Appendix

Appendix A: Publications and patents attribution to students and supervisors

In this appendix, we describe the matching procedure used to link OpenAlex authors and patent inventors with PhD students and supervisors.

Matching OpenAlex authors with PhD students and supervisors

We collect bibliometric information about PhD students and their supervisors using OpenAlex. We start by converting the students' names listed in the doctoral dissertations and the OpenAlex authors' names to lowercase, removing punctuation, accents, special characters, middle initials, standalone single letters, double spaces, and content within parentheses. Then, we match students' names with the OpenAlex authors' names. Among the 62,917 students in the doctoral dissertation dataset, we successfully matched 40,349 PhD students to unique OpenAlex author profiles, while 16,233 students were not matched with any profile, and 6,335 PhD students were matched with multiple profiles. For supervisors, we apply a similar matching strategy.

After matching, we filter the results according to multiple criteria. For students, we drop student-author pairs if the author's first publication occurs more than two years before the student's enrollment in the PhD program or more than ten years after her defense year. The rationale behind these filters is to ensure that the author's publications are realistically aligned with the student's academic timeline. For supervisors, we drop supervisor-author pairs if the author's first publication is older than five years before the supervisor's first PhD supervision, in line with the French requirement for all supervisors to hold an HDR (*Habilitation à Diriger des Recherches).*[13] After filtering the student-author pairs and the supervisor-author pairs according to the two preceding criteria, we obtain 36,884 student matches with a unique OpenAlex profile. We use this sample of students as the study sample for our analysis.

Matching patent inventors with PhD students and supervisors

The European Patent Office (EPO) patent data is extracted from the Patstat database. The EPO database contains inventors' and patent applicants' personal information, such as names, addresses, countries, and sectors (i.e., company, government, non-profit organization,

[13] The French academic system requires that all PhD superviosrs hold an HDR, a diploma that requires a significant publication record. Therefore, it is unlikely that a supervisor's first journal publication would appear five years after they began their first PhD supervision.

university, hospital, and unknown). We cleaned inventors' names by removing punctuation, accents, special characters, middle initials, standalone single letters, double spaces, and content within parentheses. Similarly, we cleaned students' names from the doctoral dissertation dataset. To avoid incorrect matches, matches are filtered based on the timeline of patenting activity relative to the student's thesis defense year. An inventor whose first patent occurred more than 6 years before the student's thesis defense is excluded. To avoid potential ambiguities, cases where a student is linked to multiple inventors are removed to retain unique matches. Similarly, we cleaned supervisors' names and matched them with inventors' names. In this case, we drop inventor-supervisor pairs in which the inventor has filed their last patent more than 10 years before the year of their first thesis supervision and inventor-supervisor pairs in which the inventor has filed their first patent more than 30 years after their first thesis supervision. As before, unique supervisor-inventor matches are kept. Sectoral information is considered to classify the applicants reported in the students' and supervisors' patents into public (government, nonprofit, university, and hospital) and private applicants (companies, individuals, and unknown). Country information is collected to identify U.S. or Chinese patent applicants. We exclude student-inventor pairs associated with multiple Patstat inventors' profiles by retaining only 1,790 students matched with a unique profile. Excluding students with multiple inventor profiles leads us to a study sample of 35,492 students.

Appendix B: AI theses identification strategy

In this appendix, we detail the methodology used to identify AI-related theses. The literature does not agree on a single definition of AI, as it includes a wide range of technologies (Agrawal et al. 2019; Vannuccini and Prytkova 2023). Sometimes, the definition can be quite narrow, focusing only on specific areas like symbolic systems, robotics, and machine learning (Cockburn et al. 2018), while others define AI more broadly, encompassing a wider set of technologies (Baruffaldi et al. 2020; Calvino et al. 2023). In this paper, we adopt a broad definition of AI, identifying documents with AI content through keywords present in titles or abstracts. We develop our list of keywords by merging two widely recognized sources: one from an OECD report with 189 AI keywords (Baruffaldi et al. 2020; Calvino et al. 2023), and another from an economic study with 39 AI keywords (Cockburn et al. 2018). After combining these lists and adding plural forms of the keywords, we refined our list to 281 unique AI keywords. The complete list of keywords is reported in Table B1.

We used these keywords to analyze 35,492 PhD theses in STEM disciplines in France between 2010 and 2018 and identified 2,555 theses (7.20%) as AI-related. According to Bianchini et al. (2022), some keywords used in the literature are too generic to accurately identify AI-related content. This is mainly why we see this high proportion of AI theses. Our broad approach resulted in classifying documents with marginal AI relevance, such as those containing "robotics".

To improve our keyword list, we evaluated each keyword by three metrics: the number and share of theses containing the keyword, its co-occurrence probability with other AI keywords, and its distribution across different disciplines. The rationale behind this approach is that if an AI keyword is used to identify a high number of theses, rarely co-occurs with other AI keywords in the same document, and is spread across various disciplines, then the keyword is considered to be too generic to identify AI-related documents, therefore, we drop the keyword from our list. We identified overly generic keywords using Table B2 after aggregating keywords that refer to the same concept. For instance, we group the keywords “robotics”, “robot”, “robots”, and “robotic” under the same concept “robot”. Five concepts—"algorithm," "network," "mapping," "robot," and "decision making"—were deemed too generic due to high frequency (columns 1-2), low co-occurrence with other AI keywords (column 3), and widespread use across various disciplines (columns 4-11). Keywords associated with these concepts were also removed from our list.

After this refinement, we ended up with 272 keywords. Using this revised list, we identified 2,555 AI theses, representing 7.20% of the total theses analyzed. This approach ensures a more precise identification of AI documents by excluding overly generic terms.

Table B1: Full list of keywords

action recognition	cognitive computing	feature extraction	image matching	machine learning	particle swarm optimisation	similarity learning
activity recognition	cognitive modelling	feature learning	image processing	machine translation	pattern analysis	simultaneous localisation
adaboost	collaborative filtering	feature selection	image recognition	machine vision	pattern recognition	social robot
adaptive boosting	collaborative systems	fuzzy c	image retrieval	mapping	pedestrian detection	social robots
adversarial network	collision avoidance	fuzzy logic	image segmentation	mapreduce	physical symbol system	sparse representation
adversarial networks	community detection	fuzzy number	independent component analyses	markovian	q learning	spectral clustering
algorithm	computational intelligence	fuzzy numbers	independent component analysis	memetic algorithm	random field	speech recognition
algorithms	computer vision	fuzzy set	inductive monitoring	memetic algorithms	random fields	speech recognitions
ambient intelligence	convolutional neural network	fuzzy sets	industrial robot	meta learning	random forest	speech to text
ant colony	convolutional neural networks	gaussian mixture model	industrial robots	motion planning	random forests	stacked generalization
ant colony optimisation	crowdsourcing and human computation	gaussian mixture models	instance based learning	multi agent system	rankboost	statistical relational learning
artificial bee colony algorithm	cyber physical system	gaussian process	intelligence augmentation	multi agent systems	recommender system	stochastic gradient
artificial intelligence	cyber physical systems	generative adversarial network	intelligent agent	multi label classification	recommender systems	supervised learning
artificial neural network	data mining	generative adversarial networks	intelligent agents	multi label classifications	recurrent neural network	support vector machine
artificial neural networks	decision making	genetic algorithm	intelligent classifier	multi layer perceptron	recurrent neural networks	support vector machines
association rule	decision tree	genetic algorithms	intelligent classifiers	multi objective evolutionary	regression tree	support vector regression
autoencoder	decision trees	genetic programming	intelligent infrastructure	multi objective optimisation	regression trees	support vector regressions
autoencoders	deep belief network	gesture recognition	intelligent infrastructures	multi objective optimisations	reinforcement learning	swarm intelligence
autonomic computing	deep belief networks	gradient boosting	intelligent software agent	multi sensor fusion	relational learning	swarm optimisation
autonomous vehicle	deep convolutional neural	gradient tree boosting	intelligent software agents	multi sensor fusions	robot	symbol processing
autonomous vehicles	deep learning	graphical model	k means	multi task learning	robot systems	symbolic error analyses
backpropagation	deep neural network	graphical models	kernel learning	naive bayes classifier	robotic	symbolic error analysis
bayesian belief network	deep neural networks	hebbian learning	knowledge representation and reasoning	naive bayes classifiers	robotics	symbolic reasoning
bayesian belief networks	dictionary learning	hidden markov model	latent dirichlet allocation	natural gradient	robots	systems and control theory
bayesian learning	differential evolution	hidden markov models	latent semantic analyses	natural language generation	rough set	temporal difference learning
bayesian network	differential evolution algorithm	hierarchical clustering	latent semantic analysis	natural language processing	rule based learning	text mining
bayesian networks	dimensionality reduction	high dimensional data	latent variable	natural language understanding	rule learning	text to speech
biped robot	dynamic time warping	high dimensional feature	latent variables	natural languages	self organising map	topic model
biped robots	emotion recognition	high dimensional input	layered control systems	nearest neighbour algorithm	semantic web	trajectory planning
blind signal separation	ensemble learning	high dimensional space	learning automata	nearest neighbour algorithms	semi supervised learning	trajectory tracking
bootstrap aggregation	evolutionary algorithm	human action recognition	legged robot	network	sensor data fusion	transfer learning
brain computer interface	evolutionary algorithms	human activity recognition	legged robots	neural network	sensor data fusions	unmanned aerial vehicle
brain computer interfaces	evolutionary computation	human robot interaction	link prediction	neural networks	sensor fusion	unmanned aerial vehicles
chatbot	extreme learning machine	human robot interactions	link predictions	neural turing	sensor fusions	unsupervised learning
chatbots	extreme learning machines	humanoid robot	logic theorist	neuromorphic computing	sensor network	variational inference
classification tree	face recognition	humanoid robotics	logitboost	non negative matrix factorisation	sensor networks	vector machine
classification trees	facial expression recognition	humanoid robots	long short term memory	object detection	sentiment analyses	vector machines
cluster analyses	factorisation machine	image alignment	long short term memory lstm	object recognition	sentiment analysis	virtual assistant
cluster analysis	factorisation machines	image classification	lstm	obstacle avoidance	service robot	virtual assistants
cognitive automation	feature engineering	image grammars	machine intelligence	optimal search	service robots	visual servoing
						xgboost

Table B2: AI concept statistics

Concepts*	(1) N. of theses**	(2) Share	(3) Keyword co-occurrence	(4) Biology	(5) Chemistry	(6) Computer science	(7) Engineering	(8) Geology	(9) Mathematics	(10) Medicine	(11) Physics
algorithm	4098	33.82%	41.83%	2.19%	0.00%	40.43%	28.86%	2.34%	13.25%	2.49%	9.80%
network	2442	20.15%	30.59%	17.44%	0.01%	27.81%	23.38%	5.73%	3.19%	8.27%	7.99%
robot	746	6.16%	67.61%	2.58%	0.01%	38.03%	49.30%	0.00%	1.17%	1.88%	6.57%
mapping	611	5.04%	36.50%	17.84%	0.03%	21.93%	21.93%	8.02%	6.38%	7.86%	12.11%
decision making	347	2.86%	45.82%	14.12%	0.04%	36.31%	26.80%	2.02%	2.88%	14.70%	2.88%
machine learning	275	2.27%	85.82%	4.73%	0.02%	61.82%	16.00%	1.09%	8.73%	3.27%	3.27%
sensor network	258	2.13%	74.29%	0.48%	0.00%	56.67%	38.10%	0.48%	1.43%	0.48%	2.38%
image processing	207	1.71%	68.60%	3.38%	0.02%	32.37%	32.37%	0.48%	12.08%	2.90%	16.43%
neural network	179	1.48%	71.43%	11.80%	0.07%	26.09%	27.95%	1.86%	5.59%	15.53%	11.18%
computer vision	135	1.11%	83.70%	0.74%	0.01%	57.04%	28.89%	0.74%	7.41%	0.74%	4.44%
genetic algorithm	134	1.11%	58.78%	3.05%	0.02%	22.90%	50.38%	3.05%	3.05%	1.53%	13.74%
data mining	101	0.83%	77.23%	2.97%	0.03%	69.31%	12.87%	3.96%	3.96%	4.95%	1.98%
markovian	77	0.64%	44.16%	1.30%	0.02%	18.18%	12.99%	0.00%	51.95%	2.60%	11.69%
artificial intelligence	76	0.63%	80.26%	2.63%	0.03%	59.21%	26.32%	1.32%	3.95%	5.26%	1.32%
support vector machine	73	0.60%	88.89%	1.39%	0.02%	40.28%	37.50%	4.17%	2.78%	5.56%	6.94%
multi agent system	71	0.59%	54.84%	0.00%	0.00%	62.90%	32.26%	0.00%	1.61%	1.61%	1.61%
random field	71	0.59%	62.50%	0.00%	0.00%	18.75%	48.44%	6.25%	18.75%	3.13%	4.69%
fuzzy logic	65	0.54%	76.92%	1.54%	0.02%	43.08%	46.15%	0.00%	3.08%	3.08%	3.08%
semantic web	65	0.54%	52.31%	0.00%	0.00%	95.38%	1.54%	0.00%	0.00%	3.08%	0.00%
supervised learning	58	0.48%	86.21%	3.45%	0.06%	68.97%	13.79%	3.45%	6.90%	1.72%	1.72%
pattern recognition	56	0.46%	55.36%	17.86%	0.32%	41.07%	12.50%	3.57%	5.36%	8.93%	5.36%
bayesian network	52	0.43%	83.33%	2.38%	0.06%	50.00%	40.48%	0.00%	7.14%	0.00%	0.00%
decision tree	51	0.42%	75.00%	10.42%	0.22%	45.83%	16.67%	8.33%	0.00%	10.42%	4.17%
natural language processing	49	0.40%	69.39%	0.00%	0.00%	83.67%	10.20%	0.00%	0.00%	6.12%	0.00%
artificial neural network	45	0.37%	76.19%	2.38%	0.06%	23.81%	47.62%	7.14%	0.00%	9.52%	9.52%
evolutionary algorithm	45	0.37%	70.73%	4.88%	0.12%	43.90%	36.59%	0.00%	2.44%	0.00%	7.32%
feature extraction	44	0.36%	95.45%	0.00%	0.00%	52.27%	31.82%	0.00%	4.55%	9.09%	2.27%
image segmentation	44	0.36%	79.55%	2.27%	0.05%	38.64%	27.27%	2.27%	11.36%	6.82%	11.36%
random forest	44	0.36%	90.00%	5.00%	0.13%	37.50%	27.50%	17.50%	10.00%	0.00%	2.50%
recommender system	44	0.36%	83.78%	0.00%	0.00%	94.59%	2.70%	0.00%	2.70%	0.00%	0.00%
graphical model	43	0.35%	79.49%	0.00%	0.00%	61.54%	17.95%	0.00%	17.95%	0.00%	2.56%
feature selection	42	0.35%	97.62%	2.38%	0.06%	47.62%	26.19%	2.38%	16.67%	4.76%	0.00%
humanoid robot	42	0.35%	87.88%	3.03%	0.09%	36.36%	48.48%	0.00%	3.03%	3.03%	6.06%
reinforcement learning	41	0.34%	90.24%	12.20%	0.30%	58.54%	14.63%	0.00%	2.44%	9.76%	2.44%
autonomous vehicle	37	0.31%	82.35%	0.00%	0.00%	47.06%	44.12%	0.00%	5.88%	0.00%	2.94%
deep learning	35	0.29%	97.14%	2.86%	0.08%	54.29%	28.57%	0.00%	2.86%	11.43%	0.00%
human robot interaction	34	0.28%	96.77%	0.00%	0.00%	67.74%	29.03%	0.00%	0.00%	0.00%	3.23%
object detection	34	0.28%	79.41%	0.00%	0.00%	52.94%	29.41%	0.00%	11.76%	0.00%	5.88%
unsupervised learning	34	0.28%	94.12%	2.94%	0.09%	58.82%	14.71%	0.00%	20.59%	0.00%	2.94%
dimensionality reduction	33	0.27%	72.73%	12.12%	0.37%	42.42%	24.24%	0.00%	15.15%	3.03%	3.03%
object recognition	33	0.27%	54.55%	18.18%	0.55%	45.45%	12.12%	0.00%	15.15%	9.09%	0.00%
convolutional neural network	32	0.26%	96.43%	3.57%	0.13%	53.57%	32.14%	0.00%	7.14%	3.57%	0.00%
gaussian mixture model	31	0.26%	74.19%	0.00%	0.00%	48.39%	29.03%	0.00%	9.68%	6.45%	6.45%
image classification	30	0.25%	90.00%	0.00%	0.00%	63.33%	13.33%	10.00%	6.67%	0.00%	6.67%
k means	29	0.24%	86.21%	3.45%	0.12%	41.38%	17.24%	3.45%	17.24%	10.34%	6.90%
speech recognition	29	0.24%	58.62%	0.00%	0.00%	82.76%	13.79%	0.00%	0.00%	0.00%	3.45%
hidden markov model	27	0.22%	74.07%	7.41%	0.27%	51.85%	22.22%	0.00%	18.52%	0.00%	0.00%
trajectory tracking	27	0.22%	81.48%	0.00%	0.00%	33.33%	55.56%	0.00%	0.00%	0.00%	11.11%
high dimensional data	26	0.21%	96.15%	0.00%	0.00%	42.31%	11.54%	0.00%	34.62%	7.69%	0.00%
visual servoing	26	0.21%	80.77%	0.00%	0.00%	26.92%	61.54%	0.00%	3.85%	0.00%	7.69%
community detection	25	0.21%	100.00%	0.00%	0.00%	80.00%	0.00%	0.00%	8.00%	0.00%	12.00%
hidden markov models	25	0.21%	96.00%	0.00%	0.00%	40.00%	32.00%	0.00%	28.00%	0.00%	0.00%
independent component analysis	24	0.20%	70.83%	20.83%	0.87%	8.33%	37.50%	4.17%	12.50%	8.33%	8.33%
brain computer interface	23	0.19%	55.56%	5.56%	0.31%	50.00%	22.22%	0.00%	0.00%	16.67%	5.56%
unmanned aerial vehicle	23	0.19%	77.27%	0.00%	0.00%	36.36%	50.00%	9.09%	4.55%	0.00%	0.00%
gaussian process	22	0.18%	36.36%	0.00%	0.00%	13.64%	31.82%	0.00%	54.55%	0.00%	0.00%
ambient intelligence	20	0.17%	75.00%	0.00%	0.00%	75.00%	20.00%	0.00%	5.00%	0.00%	0.00%
face recognition	20	0.17%	75.00%	0.00%	0.00%	55.00%	30.00%	0.00%	0.00%	15.00%	0.00%
hierarchical clustering	20	0.17%	55.00%	10.00%	0.50%	30.00%	15.00%	10.00%	10.00%	15.00%	5.00%
latent variable	20	0.17%	64.71%	0.00%	0.00%	41.18%	5.88%	0.00%	41.18%	11.76%	0.00%
machine translation	20	0.17%	55.00%	0.00%	0.00%	95.00%	0.00%	0.00%	5.00%	0.00%	0.00%
motion planning	20	0.17%	95.00%	0.00%	0.00%	35.00%	55.00%	0.00%	10.00%	0.00%	0.00%
image retrieval	19	0.16%	89.47%	0.00%	0.00%	84.21%	10.53%	0.00%	0.00%	0.00%	5.26%
industrial robot	19	0.16%	100.00%	0.00%	0.00%	13.33%	73.33%	0.00%	6.67%	0.00%	6.67%
sparse representation	19	0.16%	84.21%	0.00%	0.00%	31.58%	57.89%	0.00%	10.53%	0.00%	0.00%
cyber physical system	18	0.15%	52.94%	0.00%	0.00%	82.35%	17.65%	0.00%	0.00%	0.00%	0.00%
trajectory planning	18	0.15%	88.89%	0.00%	0.00%	22.22%	50.00%	0.00%	22.22%	0.00%	5.56%
obstacle avoidance	17	0.14%	88.24%	0.00%	0.00%	35.29%	52.94%	0.00%	0.00%	5.88%	5.88%
cluster analysis	15	0.12%	53.33%	26.67%	1.78%	6.67%	20.00%	13.33%	13.33%	20.00%	0.00%
collision avoidance	14	0.12%	78.57%	0.00%	0.00%	35.71%	57.14%	0.00%	0.00%	0.00%	7.14%
dictionary learning	14	0.12%	92.86%	0.00%	0.00%	35.71%	28.57%	0.00%	21.43%	14.29%	0.00%
collaborative filtering	13	0.11%	84.62%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
mapreduce	13	0.11%	69.23%	0.00%	0.00%	92.31%	7.69%	0.00%	0.00%	0.00%	0.00%
ant colony	12	0.10%	83.33%	0.00%	0.00%	66.67%	8.33%	0.00%	16.67%	0.00%	8.33%
dynamic time warping	12	0.10%	83.33%	0.00%	0.00%	41.67%	41.67%	8.33%	0.00%	8.33%	0.00%
emotion recognition	12	0.10%	75.00%	0.00%	0.00%	50.00%	25.00%	0.00%	0.00%	16.67%	8.33%
fuzzy set	12	0.10%	70.00%	0.00%	0.00%	70.00%	20.00%	0.00%	10.00%	0.00%	0.00%
recurrent neural network	12	0.10%	83.33%	0.00%	0.00%	41.67%	16.67%	0.00%	16.67%	8.33%	16.67%
gesture recognition	11	0.09%	63.64%	0.00%	0.00%	72.73%	18.18%	0.00%	9.09%	0.00%	0.00%
text mining	10	0.08%	70.00%	0.00%	0.00%	80.00%	10.00%	0.00%	0.00%	10.00%	0.00%
adaboost	9	0.07%	100.00%	0.00%	0.00%	55.56%	44.44%	0.00%	0.00%	0.00%	0.00%
autonomic computing	9	0.07%	66.67%	0.00%	0.00%	88.89%	11.11%	0.00%	0.00%	0.00%	0.00%
differential evolution	9	0.07%	55.56%	22.22%	2.47%	11.11%	44.44%	11.11%	11.11%	0.00%	0.00%
image recognition	9	0.07%	100.00%	11.11%	1.23%	55.56%	22.22%	0.00%	11.11%	0.00%	0.00%
kernel learning	9	0.07%	77.78%	11.11%	1.23%	66.67%	11.11%	0.00%	11.11%	0.00%	0.00%
memetic algorithm	9	0.07%	66.67%	11.11%	1.23%	88.89%	0.00%	0.00%	0.00%	0.00%	0.00%
pattern analysis	9	0.07%	22.22%	44.44%	4.94%	0.00%	0.00%	11.11%	0.00%	11.11%	33.33%
semi supervised learning	9	0.07%	100.00%	11.11%	1.23%	44.44%	22.22%	0.00%	11.11%	0.00%	11.11%
transfer learning	9	0.07%	100.00%	0.00%	0.00%	55.56%	33.33%	0.00%	11.11%	0.00%	0.00%
action recognition	8	0.07%	87.50%	0.00%	0.00%	50.00%	25.00%	0.00%	25.00%	0.00%	0.00%
biped robot	8	0.07%	100.00%	0.00%	0.00%	40.00%	60.00%	0.00%	0.00%	0.00%	0.00%
ensemble learning	8	0.07%	100.00%	0.00%	0.00%	75.00%	12.50%	0.00%	0.00%	12.50%	0.00%
image matching	8	0.07%	87.50%	0.00%	0.00%	50.00%	25.00%	0.00%	25.00%	0.00%	0.00%
spectral clustering	8	0.07%	75.00%	0.00%	0.00%	37.50%	25.00%	0.00%	25.00%	0.00%	0.00%

stochastic gradient	8	0.07%	100.00%	0.00%	0.00%	50.00%	12.50%	0.00%	37.50%	0.00%	0.00%
link prediction	7	0.06%	100.00%	14.29%	2.04%	71.43%	14.29%	0.00%	0.00%	0.00%	0.00%
pedestrian detection	7	0.06%	85.71%	0.00%	0.00%	57.14%	28.57%	0.00%	0.00%	14.29%	0.00%
sensor fusion	7	0.06%	85.71%	0.00%	0.00%	57.14%	42.86%	0.00%	0.00%	0.00%	0.00%
service robot	7	0.06%	100.00%	0.00%	0.00%	85.71%	14.29%	0.00%	0.00%	0.00%	0.00%
support vector regression	7	0.06%	71.43%	0.00%	0.00%	14.29%	71.43%	0.00%	0.00%	0.00%	14.29%
activity recognition	6	0.05%	83.33%	0.00%	0.00%	50.00%	16.67%	0.00%	0.00%	0.00%	33.33%
deep convolutional neural	6	0.05%	100.00%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
deep neural network	6	0.05%	100.00%	16.67%	2.78%	50.00%	33.33%	0.00%	0.00%	0.00%	0.00%
fuzzy c	6	0.05%	100.00%	0.00%	0.00%	50.00%	33.33%	16.67%	0.00%	0.00%	0.00%
genetic programming	6	0.05%	83.33%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
humanoid robotics	6	0.05%	100.00%	16.67%	2.78%	16.67%	66.67%	0.00%	0.00%	0.00%	0.00%
intelligent agent	6	0.05%	83.33%	0.00%	0.00%	83.33%	16.67%	0.00%	0.00%	0.00%	0.00%
regression tree	6	0.05%	50.00%	16.67%	2.78%	0.00%	0.00%	33.33%	50.00%	0.00%	0.00%
collaborative systems	5	0.04%	80.00%	0.00%	0.00%	80.00%	0.00%	0.00%	0.00%	0.00%	20.00%
generative adversarial network	5	0.04%	100.00%	20.00%	4.00%	0.00%	60.00%	0.00%	20.00%	0.00%	0.00%
high dimensional space	5	0.04%	100.00%	0.00%	0.00%	60.00%	0.00%	0.00%	20.00%	20.00%	0.00%
multi label classification	5	0.04%	60.00%	0.00%	0.00%	80.00%	20.00%	0.00%	0.00%	0.00%	0.00%
natural language generation	5	0.04%	60.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
sentiment analysis	5	0.04%	60.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
association rule	4	0.03%	100.00%	0.00%	0.00%	75.00%	0.00%	25.00%	0.00%	0.00%	0.00%
backpropagation	4	0.03%	75.00%	0.00%	0.00%	50.00%	25.00%	0.00%	0.00%	0.00%	25.00%
image alignment	4	0.03%	75.00%	0.00%	0.00%	0.00%	25.00%	0.00%	25.00%	50.00%	0.00%
lstm	4	0.03%	75.00%	25.00%	6.25%	50.00%	25.00%	0.00%	0.00%	0.00%	0.00%
machine vision	4	0.03%	100.00%	0.00%	0.00%	50.00%	25.00%	0.00%	0.00%	0.00%	25.00%
meta learning	4	0.03%	50.00%	0.00%	0.00%	75.00%	0.00%	0.00%	0.00%	25.00%	0.00%
multi layer perceptron	4	0.03%	100.00%	0.00%	0.00%	50.00%	25.00%	0.00%	0.00%	25.00%	0.00%
multi sensor fusion	4	0.03%	50.00%	0.00%	0.00%	75.00%	25.00%	0.00%	0.00%	0.00%	0.00%
natural languages	4	0.03%	75.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
simultaneous localisation	4	0.03%	100.00%	0.00%	0.00%	25.00%	75.00%	0.00%	0.00%	0.00%	0.00%
text to speech	4	0.03%	75.00%	0.00%	0.00%	50.00%	25.00%	0.00%	0.00%	0.00%	25.00%
topic model	4	0.03%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
autoencoder	3	0.02%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
classification tree	3	0.02%	66.67%	0.00%	0.00%	33.33%	0.00%	33.33%	0.00%	33.33%	0.00%
cognitive computing	3	0.02%	100.00%	0.00%	0.00%	0.00%	33.33%	0.00%	0.00%	33.33%	33.33%
facial expression recognition	3	0.02%	100.00%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
high dimensional feature	3	0.02%	100.00%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
knowledge representation and reasoning	3	0.02%	66.67%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
long short term memory	3	0.02%	100.00%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
multi task learning	3	0.02%	66.67%	0.00%	0.00%	66.67%	0.00%	0.00%	33.33%	0.00%	0.00%
natural gradient	3	0.02%	100.00%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
natural language understanding	3	0.02%	33.33%	0.00%	0.00%	66.67%	33.33%	0.00%	0.00%	0.00%	0.00%
sensor data fusion	3	0.02%	100.00%	0.00%	0.00%	33.33%	33.33%	0.00%	33.33%	0.00%	0.00%
virtual assistant	3	0.02%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
bayesian learning	2	0.02%	50.00%	50.00%	25.00%	0.00%	50.00%	0.00%	0.00%	0.00%	0.00%
fuzzy number	2	0.02%	100.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%
latent dirichlet allocation	2	0.02%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
latent semantic analysis	2	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%	50.00%
multi objective optimisation	2	0.02%	100.00%	0.00%	0.00%	50.00%	50.00%	0.00%	0.00%	0.00%	0.00%
naive bayes classifier	2	0.02%	100.00%	0.00%	0.00%	50.00%	50.00%	0.00%	0.00%	0.00%	0.00%
neuromorphic computing	2	0.02%	100.00%	0.00%	0.00%	50.00%	0.00%	0.00%	0.00%	0.00%	50.00%
optimal search	2	0.02%	50.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
statistical relational learning	2	0.02%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
swarm intelligence	2	0.02%	100.00%	0.00%	0.00%	50.00%	0.00%	0.00%	0.00%	0.00%	50.00%
symbolic reasoning	2	0.02%	50.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
variational inference	2	0.02%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
adversarial network	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
cognitive automation	1	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%
cognitive modelling	1	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%
deep belief network	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
evolutionary computation	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
extreme learning machine	1	0.01%	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
gradient boosting	1	0.01%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%
hebbian learning	1	0.01%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
high dimensional input	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
human action recognition	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
particle swarm optimisation	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
q learning	1	0.01%	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
relational learning	1	0.01%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%
rough set	1	0.01%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
similarity learning	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
social robot	1	0.01%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
vector machine	1	0.01%	100.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%

NOTE: *Concepts are sets of similar keywords. For instance, the terms "robotics", "robot", "robots", and "robotic" refer to the same concept "robot". **The keyword may co-occur with other keywords in the same document. Therefore, the same theses might be counted for different keywords introducing double counting for the variable "*N. of theses*".

Appendix C: Empirical analysis conducted without separating Computer Science from Other disciplines

In this appendix, we provide additional descriptive statistics for our data, along with the descriptive statistics and results presented in the main text, but applied to all disciplines. Table C1 shows that the share of AI students has been increasing over the years, while the share of PhD graduates pursuing research careers has declined.

Table C2: AI students and research careers

		All disciplines (35,492 theses)	
Defense year	AI students	Research career	N. of students
2010	6.50%	55.89%	3,167
2011	6.67%	54.57%	3,359
2012	6.67%	54.93%	3,821
2013	7.05%	54.15%	3,998
2014	6.74%	54.98%	4,034
2015	7.05%	54.75%	4,183
2016	7.21%	53.41%	4,394
2017	7.81%	53.27%	4,481
2018	8.73%	53.02%	4,055
All	7.20%	54.26%	35,492

Table C2 presents descriptive statistics on the subsample of PhD graduates who pursue a research career, in a time frame between *t+2* and *t+6*. The table shows that productivity levels, measured by the number of publications, patents, and citations, are decreasing. Additionally, more graduates are publishing AI-related papers, likely due to the rise in the number of graduates with AI knowledge, while fewer of them have foreign affiliations.

Table C2: Conditional on starting a research career, type of career and productivity by year

Defense year	Private career /collaboration	Private career	N. publications	At least one patent	N. AI publications	N. of coauthors	Abroad	U.S. or China	N. of citations	N. of students
2010	14.24%	5.37%	5.82	10.17%	0.18	42.25	49.27%	14.41%	16.10	1,770
2011	14.08%	4.91%	6.05	9.71%	0.19	46.03	49.05%	14.08%	15.04	1,833
2012	13.72%	4.67%	6.03	7.86%	0.18	43.68	51.41%	15.86%	17.21	2,099
2013	14.41%	4.99%	5.68	9.84%	0.21	45.13	50.95%	16.03%	17.74	2,165
2014	15.33%	5.55%	5.82	9.65%	0.20	60.70	48.06%	15.06%	20.19	2,218
2015	16.42%	5.90%	5.84	10.96%	0.30	59.62	48.52%	16.11%	17.83	2,290
2016	15.64%	5.33%	5.67	10.35%	0.29	55.30	46.83%	14.66%	15.81	2,347
2017	14.91%	5.32%	5.43	8.29%	0.31	45.20	46.50%	13.45%	12.43	2,387
2018	14.84%	6.23%	5.07	6.88%	0.30	44.41	44.09%	13.35%	9.64	2,150
All	14.89%	5.37%	5.70	9.29%	0.24	49.48	48.22%	14.79%	15.77	19,259

Table C3 presents detailed descriptive statistics on career trajectories and productivity metrics for students who graduated with and without AI knowledge across all disciplines. We observe that graduates with AI knowledge are significantly less likely to pursue a research career, in line with the findings of Ahmed et al. (2023) who argue that AI graduates tend to choose industry roles over research careers. Among those who do pursue a research career, the data show that graduates with AI knowledge are more likely to collaborate with private research institutions or engage exclusively in private-sector research without any public affiliations. Regarding productivity measures, the descriptive statistics indicate that graduates with AI knowledge are more likely to patent and publish papers with AI content. However, they tend to publish less, collaborate with fewer coauthors, and receive fewer citations than their non-AI peers.

Table C3: Career type and productivity by AI thesis content and discipline

	Non-AI students	AI Students		Non-AI students	AI students
All disciplines (35,492)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.55	0.45	0.00	32,937	2,555
Conditional on pursuing a research career					
Private career/collaboration	0.15	0.18	0.00	18,100	1,159
Private career	0.05	0.09	0.00	18,100	1,159
Research career productivity					
N. of publications	5.76	4.84	0.00	18,100	1,159
At least one patent	0.09	0.12	0.00	18,100	1,159
N. AI publications	0.18	1.30	0.00	18,100	1,159
N. of coauthors	51.03	25.17	0.00	18,100	1,159
Abroad	0.48	0.49	0.50	18,100	1,159
U.S. and China	0.15	0.15	0.52	24,020	1,667
Citations	15.76	15.88	0.92	24,020	1,667

In Table C4, we present the results of the PSM for all disciplines. The results are similar to the ones presented in Section 5 and show that acquiring AI knowledge during the PhD training period does not affect the probability of starting a research career after graduation or working in private research institutions. Regarding productivity measures, we find that AI-trained graduates are significantly more likely to keep publishing in AI and receive more citations. Moreover, students with AI knowledge are more likely to collaborate with research institutions based in the U.S. and China after their graduation.

Table C4: PSM results for all disciplines

	Non-AI students	AI Students		Non-AI students	AI students
All disciplines (5,110)	Mean	Mean	P-value	N. of stud.	N. of stud.
Research career	0.45	0.45	0.87	2,555	2,555
Conditional on pursuing a research career					
Private career/collaboration	0.20	0.18	0.46	1,159	1,159
Private career	0.09	0.09	0.61	1,159	1,159
Research career productivity					
N. of publications	4.88	4.84	0.89	1,159	1,159
At least one patent	0.11	0.12	0.30	1,159	1,159
N. AI publications	0.42	1.30	0.00	1,159	1,159
N. of coauthors	35.06	25.17	0.32	1,159	1,159
Abroad	0.47	0.49	0.20	1,159	1,159
U.S. and China	0.12	0.15	0.03	1,159	1,159
Citations	11.20	15.88	0.00	1,159	1,159

Appendix D: Manual verification of career outcomes using online data

We construct careers and scientific outcomes using patent and publication data. PhD graduates who pursue a research career are those who publish at least one paper or file a patent in the 5 years after graduation. In this appendix, we conduct manual checks for 100 PhD students randomly extracted from our original sample, and using LinkedIn, ResearchGate, and personal websites, we validated our proxies for students' careers.

According to LinkedIn, ResearchGate, and personal websites, 58 PhD graduates pursued research careers and 42 had non-research careers, including 7 graduates with no online information, whom we consider as non-researchers due to the lack of online presence. Our career proxy calculated with patent and publication data predicted 50 graduates as having research careers and 50 as having non-research careers. Table D1 presents the results of our manual checks.

Table D1: Classification of PhD graduates into Research vs. Non-Research Careers

		Classification based on LinkedIn, ResearchGate, and personal websites	
		Research Careers (58 students)	non-Research Careers (42 students)
Classification based on the career proxy obtained from patent and publication data	Research Careers (50 students)	TP = 48	FP = 2
	non-Research Careers (50 students)	FN = 10	TN = 40

NOTE: TP stands for True Positives, TN stands for True Negatives, FN stands for False Negatives, and FP stands for False Positives.

To assess the reliability of our career proxy, we calculate two performance measures: precision and recall. To do so, we used the values reported in Table D1 that include the researchers having a research career correctly identified by our career proxy (True positives, TP), the one not having a research career correctly identified by our career proxy (True negatives, TN), and the errors made: False negatives (FN) and False positives (FP). We calculate precision and recall as follows:

$$Precision = \frac{TP}{TP + FP} = \frac{48}{48 + 2} = 96\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{48}{48 + 10} = 82.76\%$$

We found that 96% of the students identified as having a research career by our career proxy were correct, while 82.76% of the students who actually had a research career according to LinkedIn, ResearchGate, and personal websites were correctly identified.

Appendix E: Content distance between AI and non-AI theses, by discipline

In our main analysis, we split the sample by distinguishing Computer Science graduates from those in the other STEM disciplines, hypothesizing that Computer Science graduates who developed AI knowledge during the PhD are similar to those who did not, whereas AI graduates in non-Computer Science disciplines differ significantly from their non-AI counterparts. To support this claim, we train a Word2Vec model on a random sample of 1,000,000 pairs of theses drawn from our dataset. The goal is to assess the textual similarity based on the titles and abstracts of each pair of theses. Table E1 presents the standardized average similarity scores. AI and non-AI theses in Computer Science show a high textual proximity of 0.72 standard deviations, meaning that their content largely overlaps. On the other hand, the similarity between AI and non-AI theses in non-Computer Science disciplines is negligible, often negative, showing that the content of these documents is very different.

Table E1: Textual similarity of AI and non-AI theses across disciplines

Discipline	N. of AI theses	N. of non-AI theses	Average similarity (Standardized)
Computer science	1976	4598	0.72
Mathematics	366	3177	0.02
Engineering	1366	12597	-0.10
Geology	87	2733	-0.43
Physics	297	8688	-0.46
Medicine	203	6820	-0.74
Chemistry	34	4619	-0.80
Biology	210	12192	-0.88

NOTE: This table compares the similarity of the content of a random sample of 1,000,000 pairs of theses drawn from our study sample. Specifically, it assesses the similarity of the texts of the titles and abstracts of each pair of theses by using a neural network algorithm for text analysis.

Acknowledgments: We thank participants of the Atlanta Conference on Science and Innovation Policy, Atlanta, the U.S.; the KID summer school, Nice, France; the PhD Workshop in Economics of Innovation, Complexity, and Knowledge (WICK), Collegio Carlo Alberto, Turin, Italy; the Workshop on the Organization, Economics, and Policy of Scientific Research (WOEPS), Strasbourg, France; the International Conference on the Science of Science and Innovation, Copenhagen, Denmark. We also extend our gratitude to our colleagues at Maastricht University, Maastricht, Netherlands, for their valuable comments and suggestions. We are also grateful to the French government for its financial support of this work through the grant ANR-24-CE26-1154.u

Competing interest: The authors disclose no relevant financial interests related to the research in this paper.

DOCUMENTS DE TRAVAIL GREDEG PARUS EN 2025

GREDEG Working Papers Released in 2025

2025-01	Bruno Deffains & Frédéric Marty *Generative Artificial Intelligence and Revolution of Market for Legal Services*
2025-02	Annie L. Cot & Muriel Dal Pont Legrand *"Making war to war" or How to Train Elites about European Economic Ideas: Keynes's Articles Published in L'Europe Nouvelle during the Interwar Period*
2025-03	Thierry Kirat & Frédéric Marty *Political Capitalism and Constitutional Doctrine. Originalism in the U.S. Federal Courts*
2025-04	Laurent Bailly, Rania Belgaied, Thomas Jobert & Benjamin Montmartin *The Socioeconomic Determinants of Pandemics: A Spatial Methodological Approach with Evidence from COVID-19 in Nice, France*
2025-05	Samuel De La Cruz Solal *Co-design of Behavioural Public Policies: Epistemic Promises and Challenges*
2025-06	Jérôme Ballet, Damien Bazin, Frédéric Thomas & François-Régis Mahieu *Social Justice: The Missing Link in Sustainable Development*
2025-07	Jérôme Ballet & Damien Bazin *Revoir notre manière d'habiter le monde. Pour un croisement de trois mouvements de pensée: capabilités, services écosystémiques, communs*
2025-08	Frédéric Marty *Application des DMA et DSA en France : analyse de l'architecture institutionnelle et des dynamiques de régulation*
2025-09	Éléonore Dodivers & Ismaël Rafaï *Uncovering the Fairness of AI: Exploring Focal Point, Inequality Aversion, and Altruism in ChatGPT's Dictator Game Decisions*
2025-10	Hayk Sargsyan, Aleksandr Grigoryan & Olivier Bruno *Deposits Market Exclusion and the Emergence of Premium Banks*
2025-11	Lubica Stiblarova & Anna Tykhonenko *Talent vs. Hard Work: On the Heterogeneous Role of Human Capital in FDI Across EU Member States*
2025-12	Mathieu Chevrier *Social Reputation as one of the Key Driver of AI Over-Reliance: An Experimental Test with ChatGPT-3.5*
2025-13	Rania Belgaied *L'accessibilité aux médecins généralistes libéraux pour la ville de Nice*
2025-14	Lorenzo Corno, Giovanni Dosi & Luigi Marengo *Behaviours and Learning in Complex Evolving Economies*
2025-15	Grégory Donnat *Real Exchange Rate and External Public Debt in Emerging and Developing Countries*
2025-16	Michela Chessa *Politics as A (Very) Complex System: A New Methodological Approach to Studying Fragmentation within a Council*
2025-17	Benjamin Montmartin & Mathieu Lambotte *Competition, Conformism and the Low Adoption of a Generous Pricing Scheme Offered to Physicians*

2025-18 Leonardo Ciambezi & Alessandro Pietropaoli
Relative Price Shocks and Inequality: Evidence from Italy

2025-19 Matteo Orlandini, Sebastiano Michele Zema, Mauro Napoletano & Giorgio Fagiolo
A Network Approach to Volatility Diffusion and Forecasting in Global Financial Markets

2025-20 Daniele Colombo & Francesco Toni
Understanding Gas Price Shocks: Elasticities, Volatility and Macroeconomic Transmission

2025-21 Simone Vannuccini
Move Fast and Integrate Things: The Making of a European Industrial Policy for Artificial Intelligence

2025-22 Patrice Bougette, Oliver Budzinski & Frédéric Marty
Revisiting Behavioral Merger Remedies in Turbulent Markets: A Framework for Dynamic Competition

2025-23 Amal Ben Khaled, Rami Haj Kacem & Nathalie Lazaric
Assessing the Role of Governance and Environmental Taxes in Driving Energy Transitions: Evidence from High-income Countries

2025-24 Sandye Gloria
Emergence. Another Look at the Mengerian Theory of Money

2025-25 Barbara Brixová, Anna Tykhonenko, Ľubica Štiblárová & Marianna Siničáková
Middle Income Convergence Trap Phenomenon in CEE Countries

2025-26 Grégory Donnat, Maxime Menuet, Alexandru Minea & Patrick Villieu
Does Public Debt Impair Total Factor Productivity?

2025-27 Nathalie Lazaric, Loubna Echajari & Dorota Leszczyńska
The Multiplicity of Paths to Sustainability, Grand Challenges and Routine Changes: The Long Road for Bordeaux Winemakers

2025-28 Olivier Brette & Nathalie Lazaric
Some Milestones for an Evolutionary-Institutional Approach to the Circular Economy Transition

2025-29 Pierre Boutros, Eliana Diodati, Michele Pezzoni & Fabiana Visentin
Does Training in AI Affect PhD Students' Careers? Evidence from France