# I, Google: Estimating the Impact of Corporate Involvement on AI Research

Daniel Yue
Harvard Business School

This draft: October 16, 2023
Click here for the latest version.

**Abstract.** While corporate involvement in modern scientific research is an indisputable fact, the impact of corporate involvement on scientific progress is controversial. Corporate interests can lead to constraints that redirect research activities into applied problems in a way that benefits the company but reduces scientific impact. However, corporations also provide resources such as funding, data sets, collaborators, engineers, and technical problems that researchers may otherwise be unable to access or know about, spurring knowledge creation. This paper empirically assesses the impact of corporate involvement on scientific research by focusing on *dual-affiliated* artificial intelligence researchers located at the intersection of academia and industry. After controlling for the researcher's quality and topic preferences, I find that corporate involvement leads to up to a 44% increase in field-weighted citations received by a paper. I document evidence that this effect arises because the average benefit of a firm's scientific resources exceeds the cost of that firm's scientific constraints. Specifically, I show that corporate involvement significantly increases the likelihood of a breakthrough paper and that these effects are magnified by the involvement of firms with greater resources. However, corporate involvement also alters the direction of the dual-affiliate author's research to be more aligned with the firm's commercial interests. This is the first large-scale quantitative study of any field of science to demonstrate a direct positive effect of corporate involvement on science or to describe the underlying mechanism.

Whether due to the warm idealism of technological innovation or the cold logic of profit maximization, firms have long been a source of fundamental breakthroughs in basic science. In the 20th century, firms were the birthplace of scientific developments like the transistor and cosmic microwave background radiation (AT&T), high-temperature superconductivity (IBM), and polymerase chain reaction (Cetus Corporation) (Mullis 2000; Gertner 2013; Kernighan 2019). Corporate involvement has continued more recently across a broad set of scientific topics, including the mapping of the human genome (Celera Corporation) and breakthroughs in artificial intelligence (Alphabet / Meta). Yet these examples stand out as exceptions to the widespread view that profit orientation implies that firms are *inefficient* at producing public goods like basic science research (Bush 1945; Arrow 1962). Instead, firms redirect researchers to pursue problems that are more specific, feasible, and commercially valuable, but less generally applicable (Aghion et al. 2008; Lacetera and Zirulia 2012). Nevertheless, these firm constraints are seen as *good* for welfare, as part of an economically beneficial division of labor between firms and universities. This perspective underlies much of modern innovation policy (e.g. rationalizing public support for academic science (Bush 1945; Nelson 1959))) and firm innovation strategy (with firms focusing more on external sourcing of knowledge through licenses, alliances, and acquisitions via markets for technology (Arora et al. 2001, 2018)). Motivated by the tension in these perspectives, I ask: When can corporate involvement have a positive effect on basic science research?

A recent story illustrates the core tradeoff of corporate involvement. Famously, in 2012, Geoffrey Hinton took leave from the University of Toronto to join Google at the start of the AI 'brain drain' (Metz 2021). Due to the subsequent exodus of university AI researchers into companies, researchers, journalists, and policy-makers responded with concern about the negative effects of corporate involvement on the AI field[1], with arguments in line with the aforementioned literature (Simonite 2020; Ho et al. 2022; Gofman and Jin 2023). Less well known (outside of computer scientists) is that at Google, Hinton and his collaborators published an arXiv pre-print entitled "Distilling the Knowledge in a Neural Network" (Hinton et al. 2015), which focused on reducing the runtime memory footprint and computational requirement of models while maintaining their performance. Model distillation was not a new idea, having been developed at Cornell nearly a decade prior (Buciluǎ et al. 2006). But with access to resources like Google's proprietary speech and image datasets and computational infrastructure, Hinton's team was able to extend the methodology to neural networks and show its effectiveness at scale. Hinton's work on model distillation was clearly motivated with firm value in mind (with an AI algorithm's efficiency affecting whether it could be used in mobile applications), yet also was a direct continuation of his prior academic research on neural networks. The work went on to become one of the seminal papers in the field of artificial intelligence, a fundamental breakthrough with over 14k citations on Google Scholar as of May 2023[2].

This example illustrates the resource-constraint tradeoff that I will argue characterizes corporate involvement in research more generally. On the one hand, corporate involvement

[1]The title of this paper, "I, Google", references the unintended negative effects of AI depicted in Isaac Asimov's short story series "I, Robot" (and the more broadly known 2004 movie of the same name starring Will Smith). Like the company manufacturing the robots in those stories, could Google, despite the best of intentions, be having a negative effect on AI research (e.g. Hinton) in a way that leads us into catastrophe (e.g. a socially inefficient outcome)?

[2]Interestingly, this paper was rejected by its originally intended conference and was never officially published, possibly due to a perceived lack of general relevance by the academic reviewers. It remains an arXiv pre-print to this day. Nevertheless, illustrative of Hinton's eminence in AI research, despite the large number of citations, this paper is only the 12th most cited of his works according to Google Scholar.

brings constraints intended to align research agendas toward problems that are relevant to the firm — in this case, research on the precursors to today's large language models — and away from questions that drift from the firm's core business model. On the other hand, firm involvement provides access to resources that enable scientific research that couldn't have been done otherwise — like the computing power and datasets necessary to train large machine learning (ML) models. When firm resources are available, but the firm's constraints do not bind — like when the research is already aligned with firm value — then corporate involvement can lead to significant breakthroughs in science.

But is this story an idiosyncratic case or an example of a more general mechanism? To answer this, I assemble a novel dataset demonstrating the association between corporate involvement and scientific outcomes in AI research. I report a surprising fact: that in top AI conferences, while the average paper without corporate involvement receives 38 citations, the average paper with corporate involvement receives 115 (202% more)[3]. Motivated by this large difference, I seek to disentangle how much of this is driven by differences in firm resources and how much is driven by differences in quality and topic preferences between researchers at companies and universities. To do this, I focus on a special set of researchers located at the intersection of academia and industry: *dual-affiliated* researchers, who are employed by and work with teams at universities and firms simultaneously. I identify 3,965 such AI researchers in the publication record, authoring 77,847 papers. By focusing on the difference between papers written by the same researcher with their university and industry teams, I am able to better isolate a causal effect of corporate involvement on citations[4]. Of the 202% difference in citations, my estimates attribute most of it to the positive selection of high-quality AI researchers into firms. But selection alone does not explain all this effect: within-researcher, corporate involvement is still conditionally associated with a *positive* average effect of between 12% and 44% on field weighted citations (depending on the specification).

The rest of the paper shows that the resource-constraint tradeoff is the primary mechanism driving the main effect that I observe here. In particular, I present three findings that match the empirical signature of the resource-constraint mechanism. First, I show a direct implication of the tradeoff: because corporate constraints do not always bind (e.g. if the topic is valuable to both the firm and science) but resources are always available, corporate involvement has a larger effect on the right tail of the citation distribution than the mean. As evidence, I find that the effect of corporate involvement on citations increases at higher quantiles of the citations distribution: the effect of corporate involvement on citations at the 90th citation quantile is 16% (compared to a 10% effect at the median and a 7% effect at the 10th citation quantile), and corporate involvement increases the likelihood of being a top 95 citation quantile paper by 54%. Second, I document that the effect is *heavily* driven by firms with greater resources: a one-standard-deviation increase in my preferred resource measure (126 distinct other papers written by the same focal company in the same year) increases the main effect by an additional 19%. Finally, I show that corporate involvement changes the topics that researchers work on to ones that are more relevant for firms, a smoking gun for the operation of firm constraints. I find that corporate involvement increases the likelihood of working on topics aligned with firm values like Computer Science, Engineering, and

[3]In this paper, I operationalize scientific quality by using citations, a traditional measure of scientific impact. See Section 2.3 for details behind the numbers given here and my measure of corporate involvement.

[4]This empirical strategy is similar to empirical designs in the literature on university-industry relations, which use similar within-researcher (Stern 2004) or within-idea (Bikard 2020; Bikard and Marx 2020; Marx and Hsu 2022) estimators to address first-order endogeneity concerns.

Natural Language Processing but decreases the likelihood of working on interdisciplinary efforts focused on problems from basic disciplines like Medicine, Biochemistry, and Physics. Beyond ruling in the resource-constraint mechanism through these three additional facts, I rule out the possibility that these results are driven by alternative explanations or other operational choices by conducting a series of analysis extensions and robustness tests.

Does this mechanism generalize to other fields of science, or is it unique to AI research? In the discussion, I argue for the generality of this mechanism in three ways. First, I argue that while this mechanism is general, the effect has not been shown before because of the empirical challenge of disentangling a causal effect of corporate involvement from a selection effect. Second, I show that across a wide variety of other fields beyond computer science, corporate involvement in papers from top journals is positively associated with citations, a previously unreported fact. This provides suggestive evidence that the resource-constraint mechanism demonstrated here may well generalize as a first-order mechanism operating in other fields of science but was previously unreported due to a lack of systematic data. Finally, my conceptual framework provides boundary conditions for the mechanism. It highlights that the effect of firm involvement on a given field of science depends on the amount of unique resources that firms can provide for basic research relative to the constraints imposed by the firm. Therefore, the effect may be smaller (and therefore harder to statistically isolate) in more commonly studied fields like biology, chemistry, and pharmaceuticals[5] due to firms having less relevant resources. For example, the ability to push candidate drugs through clinical trials may not lead to as many fundamental insights in drug development research compared to the ability to access large datasets in artificial intelligence research.

In summary, this study develops new theory and statistical evidence to argue that *when* the resource benefits outweigh the constraint costs of corporate involvement, *then* corporate involvement can have a positive impact on basic science research. In other words, the Hinton story is not an isolated example within an idiosyncratic field of research but rather a more general mechanism. I believe this to be the first large-scale quantitative study of any field of science to demonstrate a direct positive effect of corporate involvement on scientific quality or to describe the underlying mechanism. These results stand in contrast to a long history of research demonstrating the negative effects of corporate involvement on basic science research (see Foray and Lissoni (2010) or Perkmann et al. (2013) for a summary), which provides the basis for present thinking on the university-firm relationship in innovation (Arora et al. 2001, 2018). In doing so, it joins a small, recent group of papers calling for a more nuanced understanding of the difference between university and firm environments for scientists and the development of scientific ideas (Sauermann and Stephan 2013; Bikard 2020; Bikard and Marx 2020; Nagle and Teodoridis 2020; Marx and Hsu 2022), and an even smaller group asking for conditions in which firms can provide a helpful, complementary organizational environment to universities in the pursuit of basic knowledge (Azoulay et al. 2009; Bikard et al. 2019; Hartmann and Henkel 2020).

The paper proceeds as follows. First, I formalize the resource-constraint tradeoff into a model that allows me to characterize the empirical signature of the mechanism. Second, I introduce AI research as an empirical setting and describe my data and empirical strategy. In Section 3, I present my main results and mechanism tests, which provide empirical support for the hypotheses derived from the model. I conclude with a discussion of the generality of my results, contributions to the innovation literature on the university-firm

[5]Previous research in the university-firm relations literature and the science of science focuses largely on these fields due to their long history, large size, and economic importance.

relationship, and implications of these findings for managers at scientifically oriented firms and AI policymakers.

# 1 A Theory of Corporate Involvement and Science

This section develops and formalizes the idea that the effect of corporate involvement on science depends on a tradeoff between the resources and constraints that come with corporate involvement. While there is separate evidence from the prior literature for each of resources and constraints, this section conceptualizes them in tension with each other and models the consequences. The model guides subsequent empirical analysis.

## 1.1 Corporate Involvement as Constraints and Resources

The idea that corporate involvement *constrains* researchers by changing the projects that they pursue has extensive support in the research literature. Broadly speaking, there are two (compatible) viewpoints on how and why constraints arise from economics and sociology. From an economics perspective, the ability of firms to redirect researcher's focus toward topics of greater value to the firm is seen as the primary reason that firms would bother to hire researchers in the first place. Whereas researchers may have intrinsic preferences seeking scientific credit through publishing high scientific value ideas (Dasgupta and David 1994; Stephan 1996), the goal of the firm is to create and then capture value from scientific activities. As such, researchers trade away their right to choose project topics in exchange for higher wages (Aghion et al. 2008), or possibly take a pay cut in order to maintain the right to publish (Stern 2004). In practice, these types of explicit constraints may manifest through explicit employment contracts, publication oversight committees, obligations to collaborate with product teams, and pressure to convert discoveries into protected intellectual property.

From a sociology perspective, firms redirect researchers by bringing them into an organizational structure that operates under different *norms* or *institutional logic* (Merton 1973; Murray 2010). In particular, the literature shows that firms are less open than universities in terms of creating intellectual property protection and preventing public disclosure in a way that harms science (Perkmann et al. 2013). For example, Murray et al. (2016) found that intellectual property rights over genetically engineered mice were a deterrent to exploratory research. This literature has particularly focused on the effects of the commercialization of academic ideas via industry sponsorship and academic entrepreneurship, wherein there is evidence that for-profit norms have spread into academia. For example, Czarnitzki et al. (2015) shows that industry sponsorship limits public disclosure of research. Surveys of academic scientists suggest that patenting causes researchers to shift their research towards more commercial ideas (Blumenthal et al. 1997; Krimsky 2004). I interpret differences in norms as a form of implicit constraint, where firms design research environments conducive to specific directions of research through the design of teams and the availability of research tools. For example, by assembling a team of diverse researchers around a single area of shared overlap of interest to the firm, firms increase the likelihood that the team will work in that area even in the absence of formal constraints. Thus, even absent explicit organizational control mechanisms, firms may still constrain researchers through control of the research environment. Together, both the economics and sociology literature provide a basis for my assumption of firm involvement as constraining project selection.

By contrast, evidence that firms provide unique resources that benefit science is more limited in the literature. One prominent historical example comes from Bell Labs, AT&T's

corporate research lab. Bell Labs was the site of many famous scientific discoveries, not limited to information theory, the transistor, the Unix operating system, the cosmic microwave background radiation (CMBR), and satellite communications (Gertner 2013). Several economists of note have argued that Bell Labs is exemplary of a broader class of corporate research labs that are uniquely situated to contribute to science due to their unique access to corporate resources and technical problems (Rosenberg 1982; Arora et al. 2020), although their arguments are based on casual observation rather than quantitative analysis. In the context of AI research, several recent papers have shown that firms have greater compute and data assets than universities (Hartmann and Henkel 2020; Thompson et al. 2023), and have argued that this explains the prevalence of firm publishing in that field. However, their arguments focus on demonstrating firm publishing through correlation analysis rather than understanding firms' impact on science through causal analysis. More systematically, Bikard et al. (2019) demonstrate that collaborations between academia and firms enable researchers to focus on research implications (and ignore commercial implications) of research in a way that promotes both more follow-on research.

While direct evidence on the benefits of firm resources is scarce, from a more first-principles perspective, there is far more evidence that resources are an essential input for the production of scientific research. For example, the literature has shown that access to funding can lift constraints and enable more and more creative research (Azoulay et al. 2011; Ganguli 2017), stimulating both follow-on research and private-sector innovation (Azoulay et al. 2019). Beyond funding, other resources that benefit science may include cloud computing clusters much larger than those found at a typical university, usage of large private data sets, dedicated research time protected from administrative responsibilities, and familiarity with novel problems. Interpreting this through the lens of firm involvement, firms may use profits to directly fund research (rather than, say, rely on a grant-funding organization like the NIH or NSF) or generate these alternative resources as a byproduct of other commercial activities. Notably, some of these resources may be scarce/unavailable on well-functioning factor markets, creating the possibility that firm researchers may have access to *unique* resources unavailable in university environments. Overall, the key theoretical takeaway is that resources enable scientists to effectively execute their projects — whether through greater data access, engineering help, or other channels.

### 1.2 A Simple Model of Corporate Involvement and Science

I build on this literature by formalizing these concepts into a simple model of the effects of corporate involvement on science. I am interested in comparing the distribution of scientific quality of published papers produced by identical researchers given an exogenous choice of Institution (Firm or University). To do this, I model the scientific quality of an observed published paper $Y_{p^*}$ as arising from a three-stage process[6]. First, the researcher generates a set of project ideas; then, she chooses a project to pursue; finally, she executes the project given available resources. In particular, let a set of $n > 1$ project proposals (indexed by $p$) be identically distributed random vectors of baseline scientific and firm value $(S_p, F_p) \sim G$ (where $G$ is an arbitrary joint distribution associated with the researcher and

[6]I chose the letter $Y$ to emphasize that I think of scientific quality as observable, e.g. by measuring it through citations or other measures of scientific influence. The separation of this production process into idea generation, project choice, and project execution is standard in the innovation literature (see Girotra et al. (2010); Keum and See (2017)). However, I note that this choice is done for theoretical clarity, though there are probably other ways of modeling this process.

$S_p, F_p > 0$). The researcher chooses one of those projects to pursue $p^*$ and then executes it, leading to a published paper with scientific quality $Y_{p^*} \equiv f(S_{p^*}, R_{\text{Institution}})$, where $R_{\text{Institution}}$ is the amount of resources available given the institutional setting.

Motivated by the literature, I model the effect of firm involvement as arising from constraints and resources. I formalize the concept of constraints as changing which projects are chosen:

$$p^* = \text{SELECT}(\{(S_p, F_p)\}, \text{Institution}_r)$$

where $\text{Institution}_r \in \{\text{Firm}, \text{University}\}$. I further assume a particularly simple selection function for each institutional environment. For university researchers ($\text{Institution}_r = \text{University}$), I assume that researchers seek to maximize the scientific value of the projects that they pursue. I do this to express the idea in the literature that scientists seek scientific credit rather than pecuniary rewards and that this search for credit can explain a wide range of scientists' behaviors. Formally,

$$p^*_{\text{Univ}} = \text{SELECT}(\{(S_p, F_p)\}, \text{University}) \equiv \text{argmax}_p S_p$$

For firm scientists, I assume that researchers seek to maximize the firm value of the projects that they pursue. This assumption reflects the idea in the literature that the first-order effect of firms on researchers comes from their ability to steer the direction of research towards profitable project choices, whether through explicit contracts or implicit norms[7].

$$p^*_{\text{Firm}} = \text{SELECT}(\{(S_p, F_p)\}, \text{Firm}) \equiv \text{argmax}_p F_p$$

While simple, this formulation neatly captures the idea in the literature that firms do not prevent publishing but rather influence it through specific constraints imposed by the firm. In particular, these constraints are encoded within the joint distribution $G$.

I model resources as affecting how projects are executed. Specifically, I assume that resources affect scientific quality as a constant multiplicative factor on scientific value: $Y_p \equiv R \times S_{p^*}$. Further, I assume that $R$ is determined solely by the institutional environment and, therefore, that the effect of resources is constant across all project ideas. Without loss of generality, I set $R_{\text{Univ}} = 1$, and denote $R \equiv R_{\text{Firm}}$. Under these assumptions, the random variables representing the scientific quality of papers for university and firm researchers are given by $S_{p^*_{\text{Univ}}}$ and $R \times S_{p^*_{\text{Firm}}}$ respectively.

Beyond these assumptions, I impose a constraint on the joint distribution $G$ for technical reasons: I require that $S_p$ and $F_p$ are related linearly and share the same marginal distribution, with their slope given by the correlation between these variables $\rho \in [-1, 1]$. For Hypothesis 2, I further require that $G$ is a multi-variable normal distribution, although specific distributional assumption can likely be relaxed. A discussion of the assumptions underlying this model and the extent to which they can be relaxed is in Appendix A.2.

### 1.3 Empirical Implications

Given this formulation, I arrive at the following results (proofs in Appendix A.1).

**Hypothesis 1.** $R > R^*_{Avg} \iff E[RS_{p^*_{Firm}}] > E[S_{p^*_{Univ}}]$. *If firm resources are sufficiently large, then on average, the scientific quality of papers produced by a firm researcher will be greater than that of an (otherwise identical) university researcher.*

[7] See prior subsection for a literature review of these concepts.

Hypothesis 1 is the main effect of interest; given that firm involvement is a tradeoff of resources and constraints, when resources are sufficiently large, then the effect of firm involvement on scientific quality will be positive.

**Hypothesis 2.** $R > R^*_{QuantileEffect} \iff (Q_{95}[RS_{p^*_{Firm}}]) - Q_{95}[S_{p^*_{Univ}}]) > (Q_{50}[RS_{p^*_{Firm}}] - Q_{50}[S_{p^*_{Univ}}])$. *If firm resources are sufficiently large, then the average effect of firm involvement on scientific quality will be greater at the right tail (e.g. the 95th quantile) of the distribution than at the median.*

Hypothesis 2 provides a more surprising and distinctive empirical implication of this model, justifying the formal approach. The intuition for this quantile effect prediction is nevertheless straightforward: firm researcher selection reduces the variance of firm value but not scientific value. Therefore, some projects selected for high firm value will *happen* to have high scientific value; that is, they fall in 'Pasteur's Quadrant' for this firm (Stokes 1997). This can occur even when the correlation of firm and scientific value is negative (though it is less likely for more negative $\rho$). For those projects, the constraints of firm involvement do not 'bind', but the resources still benefit. When firm involvement constraints do not bind, firms can have a disproportionately positive effect on the right tail of the scientific quality distribution.

**Hypothesis 3.** $\partial_R(E[RS_{p^*_{Firm}}] - E[S_{p^*_{Univ}}]) > 0$. *The average effect of firm involvement on scientific quality will be greater when a researcher is involved at a firm with more resources.*

Hypothesis 3 is hardly surprising given my assumptions about how resources affect scientific quality. I highlight it here not for its theoretical insightfulness but rather because it serves as a second, independent empirical implication that is distinctive of this specific mechanism for the effect of corporate involvement.

Finally, let us interpret $\theta_p \equiv \arctan(F_p/S_p)$ as a measure of how much research aligns with the firm's commercial interests. Then,

**Hypothesis 4.** $E[\theta_{p^*_{Firm}}] > E[\theta_{p^*_{Univ}}]$. *The papers produced by firm researchers will, on average, be more aligned with their firm's commercial interests than the research produced by an (otherwise identical) university researcher.*

Hypothesis 4 is similarly unsurprising given my assumptions about the university and firm researcher selection functions but serves as a useful third distinctive empirical test for this mechanism. In particular, it is a direct consequence of the assumption that firm constraints are operating.

These hypotheses collectively delineate the distinctive empirical signature of a resources-based mechanism driving the effect of firm involvement. I now turn to empirical examination of these hypotheses in the context of AI research.

# 2 Empirical Setting and Data

## 2.1 Institutional Context

My empirical analysis centers on academic computer science research, particularly the subfield of artificial intelligence. I do this for three primary reasons.

First, AI research has first-order importance both economically and scientifically. Prior back-of-the-envelope consulting firm estimates of AI impact suggest that trillions of dollars

have been added to the global economy by AI in 2022 alone – largely driven by advances from the past decade and AI's broad applicability as a general purpose technology (McKinsey 2018; PWC 2018). More rigorously, the scientific literature has begun to demonstrate the effectiveness of AI applications in business contexts and its impact on the broader economy (Ferreira et al. 2022; Senoner et al. 2022; Babina et al. 2023; Rock 2021). AI is also thought to be a potential source of scientific breakthroughs, as exemplified by recent breakthroughs in protein folding (Jumper et al. 2021). As a result, there are many important, pressing policy debates about how to best fund and develop AI research, mostly proceeding without the guidance of quality empirical estimates.

Second, many prominent companies are heavily involved in AI research. To illustrate this, I use the SCOPUS publications database to examine firm involvement in top AI conference publications in Appendix Figure B2. Firm involvement is (unsurprisingly) led by the dominant technology companies in the USA (Google, Microsoft, and IBM). Still, there is a long tail of companies that are regularly involved, such as Adobe, Intel, and Salesforce. Nevertheless, universities are still responsible for the majority of conference publications, especially the right tail of non-leading research universities. For example, the 30th-ranked university (Technion - Israel Institute of Technology) publishes 22.4% of the output of the 1st-ranked university (Carnegie Mellon), but the 30th-ranked firm (Raytheon BBN) only publishes 2.2% of the research output of the leading firm (Google). Overall, variation in corporate involvement across research papers powers my ability to draw general, quantitative conclusions.

Finally, as an academic research setting, AI research has left an extensive paper trail of affiliations, collaborations, and citation linkages. Two particular features stand out. First, despite the relatively recent explosion of AI research, there is sufficient sample size to power a quantitative study because AI research is primarily published through a small set of conferences rather than traditional academic journals[8]. Conferences differ from journal publications in that they have a much faster peer-review process, and the unit of work may be much smaller (with projects on the order of months rather than years). The work published in CS conference proceedings tends to be presented much closer to the date of discovery than in fields like economics or management, where it may take more than three years to publish findings. As a result, many authors publish many papers each year; it's not unusual for established researchers to publish dozens of conference papers a year. Despite the regularity of publication, AI research conferences like NeurIPS or ICML have been the source of some of the most influential papers of the past decade. Second, affiliations for individual researchers are recorded on a per-paper level. This provides a nuanced, dynamic measure of researcher affiliation over the course of their career – a necessary first step in any empirical test of the effect of firms on workers. What's more, unlike in disciplines like the social sciences, corporate involvement is typically seen as a mark of prestige rather than one of potential bias, reducing concerns around the under-measurement of corporate involvement.

The benefits of a narrower study scope outlined above come at the cost of generalizability. While AI research spans many research subfields and approaches, such research is naturally resource-intensive in a way that skews towards digital resources and fast publication timelines. As a result, results should be extrapolated to other fields of science with caution, as the specific relationship between firm resources and scientific needs will

[8]AI research is also published in traditional journal publications, but this process is longer and is not a requirement for being a researcher in the field. However, I do include these traditional publications in addition to conference publications in my empirical analysis.

likely vary by field[9]. Nevertheless, AI research is similar along many important dimensions to other fields of science, such as large heterogeneity in team size, skills, and experience, global representation, and broad representation of cross-disciplinary measures.

Over the past decade-plus of AI's rapid growth, practicing AI researchers and journalists have observed the special role of firms in the field and have speculated as to the function and dangers of corporate involvement (Gertner 2013; MacroPolo 2020; Smith 2021; AAAI 2020). This study presents a unique opportunity to test their insights in a way that is more systematic and can yield generalizable insights into the role that companies play in science.

### 2.2 Analysis Sample — Dual-Affiliated Researchers

My empirical approach focuses on *dual affiliated* researchers who are simultaneously employed by and work with research teams from both a university and a firm. While not a new employment arrangement, AI researcher dual affiliations have risen in frequency over the past decade due to a sharp increase in firm interest in hiring senior AI researchers to facilitate the development of in-house AI research teams and the growth of AI research on firm-specific problems. Due to the relatively fixed supply of senior AI researchers, such researchers have the power to negotiate working arrangements that suit their preferences – in particular, they retain the ability to teach, advise students, and direct their own research freely at a university while simultaneously accessing the research resources available at firms.

To illustrate how dual affiliations work, consider the example of Joelle Pineau, a Professor of Computer Science at McGill University and VP of Research at Facebook AI Research, "She spends half her time conducting university research, but she has given up teaching entirely. At McGill, she tends to work on healthcare projects, for which she receives research funding from Google, Samsung, and Huawei." When reflecting on why she decided to work for Facebook, she recounts, "So much of AI research is done in industry, that for me to be a leading AI scientist I felt I had to cross over... Facebook, like many other well-run tech firms, has put a lot of forethought into how to structure these teams" (Murghia 2019).

In order to identify dual-affiliated researchers at scale, I use publications data from Elsevier's SCOPUS Database. I chose SCOPUS over other publications databases due to its strong coverage of both CS journals and publications and its extensive coding efforts related to author affiliations, including affiliation disambiguation/deduplication and labeling whether an affiliation is a university, firm, or some other type. Given this data, identifying dual affiliates is straightforward – SCOPUS provides a list of all of their relevant affiliations for each author on each paper. I define the relevant set of dual-affiliated authors as follows. First, I create a database of papers from top CS conferences using a list of top AI conferences (see Appendix B.1 for details). Second, I expand this dataset to include any publication by any of these authors before 2017[10]. Using this dataset, I identify any papers on which an author is dual-affiliated — that is, they have both a company and a university affiliation on the same paper. I call this a *dual paper*, the corresponding author a *dual author*, and the year of publication a *dual year* (for that author).

Using this method, I identify 3965 dual-affiliated researchers. Appendix Table B2 provides descriptive statistics about these researchers. Dual-affiliated researchers tend to be senior, well-established researchers; for example, the median year of first publication is 2002, the median identified researcher publishes an average of 3.4 papers per year in their observed career, and the median (across authors) of the most cited paper (within five years

[9] An extended discussion of the generalizability of these results is provided in Section 4.1.

[10] I impose this constraint because I require 5 years of lag time for citations to materialize.

of publication) in a career is 107 citations. As expected, given my methodology for identifying them, dual-affiliated researchers tend to primarily publish in computer science-related publications, although there is peripheral involvement in other disciplines like medicine. Finally, most dual-affiliated researchers only have dual-affiliation for a single year. However, this distribution varies a lot: while the 75th percentile author has 2 dual-years, the max has 24 registered dual-years, so many of the dual-years represented in my sample come from a limited number of dual authors.

I then derive a paper-researcher-level sample based on these dual-affiliated researchers. Specifically, I form a dataset comprising any paper in SCOPUS (either conference or journal publication) authored by these dual-affiliated researchers, filtering out papers with greater than 10 authors or with more than two dual-affiliated authors on them[11]. Because my interest is in comparing papers across institutional affiliations holding the researcher fixed, I further filter this data set to include only paper-author published in the same year and same researcher as a dual paper-researcher[12]. I call this set of 77,847 paper-researchers "in dual year" and use it as my primary analysis sample[13].

### 2.3 Empirical Strategy

An average difference in citations between university and company-backed research is not evidence of a causal effect of corporate involvement on scientific quality because it may simply reflect that the researchers who choose to work in firms are different from the researchers who choose to work in universities. Prior studies have shown researcher prominence is a leading driver of moving to industry in the AI setting (Jurowetzki et al. 2021), suggesting that raw differences in citations overstate the overall effect[14]. Further, I suspect that the researchers that do affiliate with companies may do so at strategic times — for example, firms may deliberately seek out hot researchers, inducing spurious correlation into my different estimates.

Even without statistical controls, a simple visual comparison of the citation distributions of papers from top AI conferences and papers by dual-affiliated authors in dual years confirms the dramatic effect of researcher selection into industry in this sample. In Figure 1, I plot the raw citation distribution (x-axis on log 10 scale) for the sample of papers from top CS conferences (top) and papers by dual-affiliated authors in dual-affiliated years (bottom). While in both cases, the right-tail of the citation distribution is greater for papers with a large amount of corporate involvement, the difference is much more pronounced for the top conferences sample. A comparison of differences in means of these distributions quantifies

[11] I remove papers with greater than 10 authors because it becomes harder to attribute credit to the focal researcher as team size grows. I remove papers with more than two dual-affiliated authors because it's unclear how to leverage researcher-fixed effects with multiple dual-affiliated researchers on the same paper. To handle this two-author case, I run a paper-researcher-level analysis, including a paper as two observations if it has two dual-authors and clustering along the paper dimension in addition to other fixed effects. Two-dual-researcher papers comprise 15,811 papers in my sample; there are 46,255 papers with only one dual-author in my sample. The results are robust to the exclusion of these multiple-dual-author papers or arbitrary selection of one of the authors.

[12] Some dual-affiliated authors publish within a dual-year using only their university affiliation or (more rarely) only their firm affiliation. While the precise reason for this behavior varies, I nevertheless include these papers in my sample because they provide a snapshot of the researcher's activities within the year.

[13] A full description of the data cleaning pipeline is described in Appendix B; I delay presenting the descriptive statistics of this sample until Section 2.4.

[14] This selection effect appears to differ depending on the field of study; see Roach and Sauermann (2010) for further discussion.

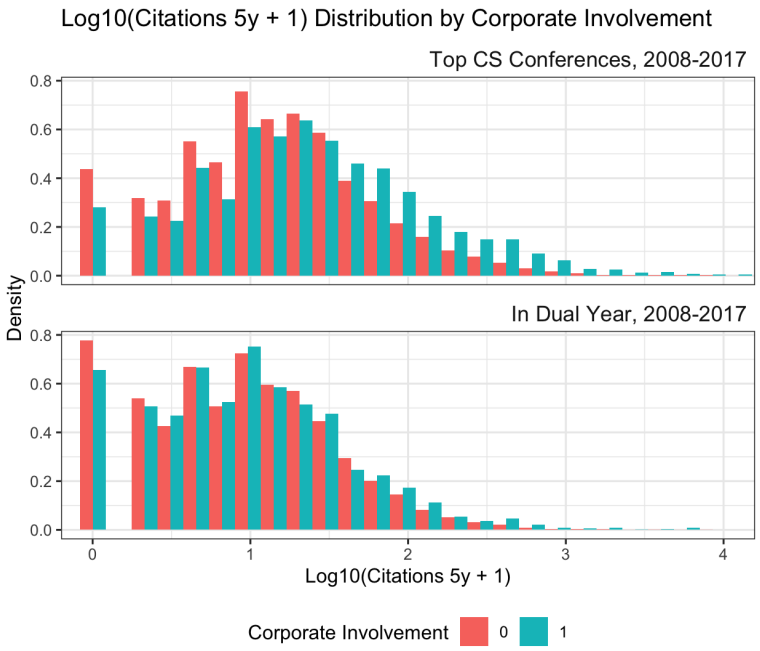


Figure 1. Distribution of (Logged) Citations in Top CS Conferences and In Dual Year (Analysis) Samples.

this visual intuition: whereas top conference papers with corporate involvement average 77 more citations in 5 years (relative to a baseline average of 38, a 202% increase), in dual year papers with corporate involvement average only 17 more citations in 5 years (relative to a baseline average of 24, a 70.8% increase).

This endogeneity concern mirrors a broader challenge in the university-industry relations literature in separating selection concerns from causal effects due to a lack of convincing natural experiments in varying institutional settings. That literature typically uses empirical designs focused on *idea-twins*, which focuses on differences within pairs of frequently co-cited papers that differ in their source environments. By 'holding the idea fixed', these designs enable attribution of differences in outcomes (like use in patents) to differences between the university and industry environments (Bikard 2020; Bikard and Marx 2020; Marx and Hsu 2022). However, given that I theorize constraints as affecting the idea-selection process, idea-twin designs do not work here because they control away variation of interest. Instead, I introduce a novel empirical strategy: controlling for researcher selection by directly comparing papers by the same dual-affiliated author in the same year but produced with different teams (some from universities, some from firms). In other words, the researcher's university-backed papers from the same year function as a control group for their firm-backed papers. I aim to interpret the difference as arising from differences between these institutional environments.

As an example, consider publications by Prof. Joelle Pineau from 2018. In that year, in collaboration with colleagues from industry, she published papers like "Deep Reinforcement Learning that Matters" (Henderson et al. 2018a) in top CS Conferences (AAAI). In that same year, with only university co-authors (primarily from McGill, her home university),

Prof. Pineau published papers[15] like "Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis" (Durand et al. 2018), "Ethical Challenges in Data-Driven Dialogue Systems" (Henderson et al. 2018b), and "A decision-theoretic approach for the collaborative control of a smart wheelchair" (Ghorbel et al. 2018), also in top AI conferences. I focus on estimating the difference in citations between Prof. Pineau's papers done in collaboration with industry and with universities from that same year.

Specifically, all estimating equations regress a paper-researcher $p$'s outcome (such as citations or subject dummies) on a measure of corporate involvement within researcher-year $rt$:

$$E[Y_p \mid X_{prt}] = f(\beta \cdot \text{Percent Private}_p + \gamma_{rt} + \delta \cdot \mathbf{X}_{prt})$$

where $Y$ is a paper outcome, Percent Private is a measure of corporate involvement between 0 and 1, $\gamma$ is the researcher-year fixed-effect, and $\mathbf{X}$ contains a collection of control variables including fixed effects for a paper's university and teammate experience measures.

The researcher-year fixed-effect controls for many time-invariant and time-variant characteristics that could influence paper outcomes by effectively comparing papers only to other papers by the same author in the same year[16]. For example, $\gamma_{rt}$ accounts for differences in researcher quality or propensity to work on topics of interest to a company, as well as changes to that researcher's scientific agenda at the yearly level. The university fixed effect and teammate experience measures similarly exclude differences in university quality or teammate quality from my estimator. In a sense, my specification 'holds university resources fixed' while allowing firm resources to vary.

$f(\cdot)$ represents the various functional forms I use to estimate my regressions. The primary dependent variable of interest, five-year citation count[17], is skewed and non-negative. I deal with this by estimating a series of alternative specifications: first, ignoring this issue and using a linear model, second, by taking the natural logarithm of the outcome and using a linear model, Third, following the literature on the science of science, I estimate the conditional fixed-effects Poisson model (Hausman et al. 1984; Azoulay et al. 2019). Finally, I use a linear model predicting the percentile cites, the citation quantile of the paper relative to all other papers from the same year and the same field. I cluster standard errors at the researcher-year and paper level[18]. I also use dummy variables as dependent variables throughout the analysis. In that case, I estimate linear probability models, though my results are robust to the use of logistic regression.

While this empirical strategy eliminates the primary selection concern from my estimates (e.g. that researchers may strategically choose institutional affiliation at particular times), it nevertheless deviates from the idealized scenario described by the theoretical model in Section 1. In the ideal econometric scenario for estimating the model, researchers would be randomly allocated to work in corporate or academic institutions, and their subsequent publication output could be compared between researchers to infer the causal effect of corporate involvement[19]. Instead, the proposed strategy examines a scenario that deviates in

[15]She also published an influential short textbook "An Introduction to Deep Reinforcement Learning" (Francois-Lavet et al 2018), though in my analysis I only consider journal and conference publications.

[16]As noted in the introduction, this follows similar designs in the literature on the university-industry relationship (Stern 2004; Bikard 2020; Bikard and Marx 2020; Marx and Hsu 2022).

[17]As explained later, I actually prefer to use a field-weighted version of citation counts (FWCI) rather than raw citations in my main specification.

[18]I cluster along the paper-level to handle the case when papers have two dual-authors on the same paper. As noted before, I explicitly exclude papers with more than two dual authors from my sample altogether.

[19]In this ideal scenario, one could also examine the effect of corporate affiliation on researcher-level out-

several key ways from the ideal. Perhaps the most apparent deviation is that because the researcher operates in both institutional environments simultaneously, there's a possibility of between-institution resource spillover. For example, what if corporate involvement on one project gave a researcher ideas that they pursue with their university team on another? While this mechanism is plausible, we argue that the only plausible effect of this spillover would be to minimize differences between the university and firm research environments, positively affecting the quality of university papers (assuming a positive effect of firm resources). In other words, this spillover would negatively bias any estimate of an effect of firm involvement on citations, implying that any observed positive effect is an underestimate of a 'true' effect.

Two other deviations between the proposed empirical strategy and theorized model present a more substantive challenge to analysis. First, a researcher may *match* ideas to a particular institutional environment at various points in the research pipeline, leading to residual endogeneity concerns. Second, my empirical strategy cannot rule out alternative causal mechanisms such as selective publication (e.g. firms suppressing the publication of certain papers). Both of these types of deviations are present in and have the potential to bias my estimates of a main effect[20]. Therefore the estimated main effect should be interpreted as a conditional correlation rather than an average treatment effect.

Nevertheless, in Section 3, I provide a collection of evidence that strongly suggests a specific causal interpretation of the estimated main effect, namely that it arises primarily from the tradeoff of resources and constraints that come with firm involvement. I establish this evidence in three stages. First (Section 3.1), using the regression framework described above, I show that corporate involvement is still positively conditionally associated with a significant citation premium even after controlling for researcher topic and preferences. Second (Section 3.2), I present a collection of test results based on alternative outcomes and heterogeneity in the main effect that match the empirical signature of my theory, strengthening my interpretation of this effect as arising from the resources-constraint tradeoff. Finally (Section 3.3), I consider the ways that my empirical strategy deviates from the idealized scenario described in my model and show through additional tests that these deviations cannot explain the observed corporate citations premium.

### 2.4 Variable Definition and Descriptive Statistics

To implement this empirical strategy, I operationalize theoretical concepts with specific variables from the SCOPUS sample defined above. I form the following groups of variables in my paper-author-level dataset.

1. **Corporate Involvement** (Percent Private). I measure corporate involvement on a paper using the affiliation labels provided by the authors. To measure the extent of corporate involvement in a paper, I first expand my dataset to the paper-author-affiliation level and code each affiliation by whether it is "private" or "academic" (or

comes like productivity (count of publications). However, in this paper, we restrict the analysis to paper-level outcomes. We do this because the amount of time spent by dual-affiliated researchers within each institutional setting is endogenous, but a publication is a comparable unit of work in either institution. Therefore, conditional on publication, comparing papers produced in different institutional environments effectively controls for differences in researcher effort between institutions

[20]These concepts are visually organized in Figure 4, which appears later in the manuscript when I directly address the topic of alternative explanations and discuss how they affect the interpretation of results.

other)[21]. I then aggregate to the paper-author level, computing Percent Private at that level. Finally, I aggregate to the paper level, computing the average Percent Private across all authors on the paper. The result is a paper-level measure of private involvement that is driven mostly by the teammates of the dual author. This variable is intended to capture the difference between working with one's university team vs. their firm team. To ensure robustness, I also explore creating binary measures based on this continuous measure, such as binary dummies indicating whether Percent Private is greater than certain thresholds. These corporate involvement measures serve as the primary predictor variables in my regression analysis.

2. **Scientific Quality** (Citations 5y, FWCI 5y). I use citations to measure scientific quality outcomes[22]. In order to eliminate differences in citations to papers that have been around for longer, I limit my citation counts to those accrued within 5 years of publication. Importantly, my preferred citation measure, Field Weighted Citation Index (FWCI), accounts for differences in citation rates across academic fields by dividing out the average citation amount for papers in that field in the same year. A FWCI (5y) of 1 is interpreted as a paper receiving the average number of citations in 5 years for papers in that field. Most directly, I use a measure of FWCI directly computed and calculated by the SCOPUS database. Beyond this direct usage (in standard or log10 form), I also form dummy indicators of citation quantiles by calculating FWCI quantiles for each year of publication and then marking papers based on their FWCI relative to quantiles from the same year of publication. These dummies enable me to capture changes in the citation tail to test the various aspects of Hypothesis 2.

3. **Company Resources** (Company Papers, Authors, Credits). I proxy for the scientific research resources available at a firm by the amount of other publication activity going on at the same firm in the same year as a focal publication. To do this, I assign each paper a firm affiliation based on company affiliations on that paper (and take the largest firm if there are multiple authors with company affiliations). I then aggregate my dataset to the affiliation-year level and compute the distinct number of papers and authors with that firm affiliation in that year. I also form a measure called paper 'credit', which weighs authors less if the co-author team on a paper is larger (so a paper with one university researcher and one firm researcher on it would count as a half-credit for the university and a half-credit for the firm). If a paper has no firm involvement, I code these variables as 0. I use these measures to test Hypothesis 3.

4. **Alignment with Firm Commercial Interest** (Academic Subject Labels: Computer Sciences, Medicine, etc.). To form a broad, high-level outcome measure of a project's alignment with firm commercial interests, I leverage the SCOPUS subject labels. SCOPUS codes paper subjects using the All Science Journal Classification (ASJC) scheme, manually assigned based on the paper's source publication[23]. The ASJC scheme has three levels; in my analysis, I focus on the middle level (the 'ASJC Group'), as the values at these levels are interpretable based on my theory. For ex-

[21] See Appendix B.4 for an extended discussion of cleaning and validation of the affiliation labels used in this analysis.

[22] While citations are noisy measures of quality, they remain the standard measure for scientific quality across a broad set of disciplines studying innovation and science.

[23] For details on the ASJC scheme, see Elsevier's documentation.

ample, given that I am focusing on computer scientists, I interpret the ASJC Group 'Medicine' as less aligned with the firm's commercial interests than the ASJC Group 'Computer Science'[24]. Importantly, papers may have multiple ASJC codes, which means they may (for example) be both Medicine and Computer Science (although this is rare). I form dummy variables based on these labels to capture high-level differences in disciplinary focus, which allows me to test Hypothesis 4.

5. **Journal / Conference Rankings** (SJR, SNIP). As an alternative measure of scientific quality and as a way to control for publishing ability (independent of scientific quality), I measure the journal ranking of each paper based on its publication source in its year of publication. I use two standard measures of journal rank: the Scimango Journal Ranking (SJR) and the Source-Normalized Impact per Paper (SNIP).

6. **Team Information** (Number of Authors, etc.). Finally, I develop an array of control variables based on the teammates of the focal dual-affiliated author on a given paper. Beyond the count of authors on the paper, I reconstruct the publication records of each teammate of the focal dual-author on each paper and aggregate across teammates to form summary measures of teammate experience on each paper. In particular, I take the median across the following measure: Teammate's Publications in the Prior and Prior 5 Years, Citations in the Prior 5 Years, Distinct ASJC Codes across all teammates in the Prior 5 Years (a measure of team diversity), and Years Since Teammate's first academic publication. I also explore operationalizations involving the maximum across teammates.

| Variable | Mean | Std. Dev. | Min | Q25 | Q50 | Q75 | Max | Obs. |
|---|---|---|---|---|---|---|---|---|
| Year | 2013.02 | 7.65 | 1963.0 | 2008.00 | 2015.00 | 2019.00 | 2022.00 | 77847 |
| Conference Publication | 0.61 | 0.49 | 0.0 | 0.00 | 1.00 | 1.00 | 1.00 | 77847 |
| Percent Private | 0.20 | 0.30 | 0.0 | 0.00 | 0.00 | 0.33 | 1.00 | 77847 |
| Percent Private = 0 | 0.56 | 0.50 | 0.0 | 0.00 | 1.00 | 1.00 | 1.00 | 77847 |
| Percent Private = 1.0 | 0.05 | 0.22 | 0.0 | 0.00 | 0.00 | 0.00 | 1.00 | 77847 |
| Citations 5y | 21.16 | 123.57 | 0.0 | 2.00 | 6.00 | 17.00 | 7837.00 | 47411 |
| Field-Weighted Citation Index (FWCI) 5y | 2.99 | 11.95 | 0.0 | 0.30 | 1.04 | 2.74 | 899.43 | 47411 |
| Company Papers in Year | 68.68 | 174.25 | 0.0 | 0.00 | 0.00 | 38.00 | 1193.00 | 77847 |
| Company Authors in Year | 76.39 | 195.51 | 0.0 | 0.00 | 0.00 | 41.00 | 1297.00 | 77847 |
| Company Credits in Year | 31.12 | 81.20 | 0.0 | 0.00 | 0.00 | 17.32 | 582.69 | 77847 |
| Computer Science (ASJC Group) | 0.74 | 0.44 | 0.0 | 0.00 | 1.00 | 1.00 | 1.00 | 77847 |
| Medicine (ASJC Group) | 0.05 | 0.21 | 0.0 | 0.00 | 0.00 | 0.00 | 1.00 | 77847 |
| Scimago Journal Ranking (SJR) | 1.85 | 3.35 | 0.1 | 0.34 | 0.79 | 1.75 | 28.50 | 43981 |
| Source-Normalized Impact per Paper (SNIP) | 2.07 | 2.22 | 0.0 | 0.72 | 1.33 | 2.48 | 16.10 | 43667 |
| Number of Authors | 4.33 | 1.85 | 1.0 | 3.00 | 4.00 | 5.00 | 10.00 | 77847 |
| Number of Co-author Publications in Prior Year (Median) | 4.56 | 5.82 | 0.0 | 1.00 | 3.00 | 6.00 | 345.00 | 77847 |
| Number of Co-author Publications in Prior 5 Years (Median) | 15.99 | 21.79 | 0.0 | 2.00 | 10.00 | 21.00 | 1217.00 | 77847 |
| Number of Co-author Citations 5y in Prior 5 Years (Median) | 17.78 | 55.40 | 0.0 | 2.00 | 8.00 | 18.06 | 4349.73 | 77847 |
| Number of Distinct ASJC Codes in Prior 5 Years (Union) | 70.68 | 88.03 | 0.0 | 8.00 | 44.00 | 100.00 | 1921.00 | 77847 |
| Number of Co-author's Years Since First Publication (Median) | 4.56 | 5.82 | 0.0 | 1.00 | 3.00 | 6.00 | 345.00 | 77847 |

Table 1. Summary Statistics of Analysis Sample for Selected Variables.

The descriptive statistics of my dataset for selected variables are presented in Table 1. The full process for creating and cleaning my analysis sample, including the creation and validation of distinct affiliation entities, is detailed in Appendix B. A list of the top firm affiliations present in my sample is available in Table B3. I also visualize the Year and Percent Private in Figure B3 to provide deeper intuition for this sample. A few key details merit

[24] I provide more details of this logic in Section 3.2.3.

further discussion. First, over half of the papers were published after 2015; this confirms the large increase in dual-affiliations and overall publications in AI over the past decade. Second, over half (56%) of the papers in my sample have no firm involvement, indicative of publishing arrangements where dual-authors do not publish with firm affiliations if they are working with their university team. Third, I only have citation data for papers published before 2018 (61% of my sample); this reflects my choice to only count citation data if the paper has had five years for citations to materialize. The rest of the sample published on or after 2018 is kept because it's useful for testing subject outcomes associated with Hypothesis 4. Fourth, and as one might expect given the quality of researchers in my sample, these papers are disproportionately influential in their field; the average FWCI 5y for these papers is 3x, meaning they are three times as cited as the average paper in their field. Fifth, almost three-quarters of papers (74%) are in the Computer Science ASJC Group, reflecting the heavy AI focus of the researchers of interest. Nevertheless, significant ancillary attention is given to other disciplines; over 5% of papers are from Medicine.

# 3 Results

I organize my econometric results as follows. First, I present results showing the main effect of corporate involvement on citation measures. Next, I collect evidence supporting the interpretation of this result as an outcome of a tradeoff between firm resources and constraints. I do this by conducting a series of tests that aim to match my data to the empirical signature of my theory from Section 1. Finally, I rule out alternative explanations of my main effect through an extensive series of arguments and robustness tests.

## 3.1 Main Effect of Corporate Involvement on Scientific Quality

I present the test of my main effect (Hypothesis 1) in Table 2. In Panel A, I show results from four specifications using 5-year Field Weighted Citation Index (FWCI) as the dependent variable. Model (1) shows the results of using a linear regression model, which constitutes the main effect of interest. It shows that, on average, a paper with full corporate involvement will have 0.8128 FWCI *greater* than a similar paper with only university involvement, an effect size that constitutes 27.2% (= 0.8128/2.9938) over the average amount. However, because the citation distribution is right-skewed, this effect could be driven by a smaller number of outlier papers; my next two tests aim at ruling this out[25]. Model (2) does this by taking the natural logarithm of FWCI 5y, estimating an effect of $\approx 12.53\%$. Model (3) uses the conditional fixed-effects Poisson model, estimating an effect of 44.8% ($= \exp(0.3701) - 1$). Finally, Model (4) uses the FWCI quantile relative to all publications in SCOPUS; it indicates that papers with corporate involvement are, on average, 3% higher in the FWCI distribution. Overall, these specifications tell the same story: that corporate involvement has a *positive*, statistically significant effect on the citations received by a paper, even after adding controls.

To aid with interpretability, in Panel B, I present the same specifications using raw Citations 5y as the dependent variable instead of FWCI. Panel B, Model (1) shows a similar effect to Panel A, Model (1), equivalent to a 24% average, although not statistically

[25]The differences in the estimates from Models (1) to (3) highlight that the effect of corporate involvement does not have a constant effect on the FWCI distribution, and that (in particular) the average effect is strongly driven by the right-tail of the distribution. I explore effect heterogeneity in Section 3.2.1

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Specification: | $Y$ (Linear) | Ln($1+Y$) | Poisson | Percentile($Y$) |
| *Panel A. $Y$ = Field-Weighted Citation Index (FWCI) 5y* | | | | |
| Percent Private | 0.8128** | 0.1253*** | 0.3702*** | 0.0345*** |
| | (0.3352) | (0.0224) | (0.1022) | (0.0067) |
| Mean(DV) | 2.9938 | 0.8897 | 3.0233 | 0.6907 |
| *Panel B. $Y$ = Citations 5y* | | | | |
| Percent Private | 5.0859 | 0.1935*** | 0.4469*** | 0.0289*** |
| | (3.5697) | (0.0354) | (0.1523) | (0.0060) |
| Mean(DV) | 21.1591 | 1.9699 | 21.3673 | 0.6780 |
| Author-Year FE | Y | Y | Y | Y |
| University FE | Y | Y | Y | Y |
| Team Controls | Y | Y | Y | Y |
| Observations | 47411 | 47411 | 46949 | 47411 |

*Notes* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The data is at the paper-author level. Robust standard errors clustered at the author-year and paper level are shown in parentheses. Panel A shows the estimation for the primary variable of interest: field-weighted citation index (FWCI). For interpretability, Panel B shows the same estimate, but done using Citations to the paper. All citation measures only include references from the first five years after publication.

Table 2. Main Effect Estimation.

significant due to the increased noise. However, I am unconcerned with the lack of statistical significance because Panel B, Models (2) through (4) corroborate the same qualitative conclusions as Panel A, including statistical significance.

Beyond the fact that they are positive, the effects between 12% and 44% estimated in Table 2 are interesting because they are large. For effects of a comparable magnitude, the literature estimates average effects of about 30% increase in citations from the status glow of winning the prestigious Howards Hughes Medical Investigator Award (Azoulay et al. 2014), a 5-10% decrease in citations on papers related to a paper that was retracted (Azoulay et al. 2015), and a 20% decrease in citations for a paper where the underlying research was scooped (Hill and Stein 2020). Each of these other events studied in the literature is a newsworthy, notable event in the life of a scientist; corporate involvement is a much more common phenomenon, yet I estimate a similar magnitude of impact on citations.

Although striking in both sign and magnitude, due to limitations in the empirical strategy followed here (outlined in Section 2.3), the effect found here cannot be interpreted as any form of treatment effect, but rather a conditional correlation. The rest of this manuscript is devoted to arguing that this effect comes about due to unique *resources* that firms are able to provide researchers that enable higher quality scientific work; however, this argument hinges on further analysis rather than the estimates of this effect on their own. Finally, I observe that the effects found here, while large relative to the literature, are small in comparison to the 70.8% raw difference in average citations in this sample (or the 200% difference found in the top AI conferences sample in Figure 1), indicative of a massive amount

of researcher selection. I further explore this selection with further analysis in Appendix C, which estimates alternative specifications with different combinations of control variables to probe the source of this selection.

## 3.2 Testing the Resources-Constraint Mechanism

Why do I see the effect of Section 3.1? I argue that this result is best explained as the result of firm involvement in providing a package of resources and constraints, where (for the case of scientific quality) the benefits of resources outweigh the loss of scientific quality due to firm constraints. To argue this, I estimate alternative tests focusing on heterogeneity in this effect and effect on other outcomes, seeking evidence that matches the empirical signature of this resource mechanism described in Section 1.

### 3.2.1 Larger Effect for Higher Baseline Scientific Quality

First, I seek to develop evidence for Hypothesis 2. Recall that this hypothesis formalizes the intuition that the *constraints* of firm involvement do not always affect scientific quality because they do not always bind; this occurs when firm and scientific value happen to be correlated (even if they are, on average, negatively correlated). Because of this, when constraints do not bind, but the resources of the firm are still available, researchers can produce papers that are of higher quality than would have been possible in the university environment alone.

Even simple descriptive statistics from my sample support the hypothesis: Only 19.8% of papers have greater than half of authorship affiliated with corporate involvement, but they comprise 42% of the top 100 FWCI papers. Similarly, 43.5% of papers have any corporate involvement, but they comprise 63% of the top 100 FWCI papers. To formalize and generalize these observations into statistical tests, I take two approaches: first, quantifying the effect at each quantile using quantile regression, and second, quantifying the percentage of papers in each quantile using FWCI quantile dummies.

In the first test of Hypothesis 2, I estimate quantile regression with fixed effects (Machado and Santos Silva 2019), with estimates visualized in Figure 2. The figure reveals a striking upward-sloping pattern in the effect of corporate involvement on Ln(FWCI 5 Year + 1), indicating that the effect increases for higher quantile FWCI papers. (A constant effect that did not vary throughout the distribution would show up as a flat line in this test.) Specifically, while I previously estimated an average effect of 12.53% using OLS regression, quantile regression reveals that the effect increases to 16.1% at the 90th quantile and decreases to 6.9% at the 10th quantile. Further, as the logarithm applied the the dependent variable compresses the right tail of the distribution, I expect this effect is understated compared to estimating quantile regressions directly on the FWCI distribution.

While the quantile regression analysis above gives a sense of the magnitude of the effect at each quantile, it does not quantify the shift in FWCI distribution associated with corporate involvement. The second test, presented in Table 3, uses (as outcomes) dummy indicators of whether each paper exceeds various FWCI quantiles relative to the analysis sample. Specifically, Panel A, Model (1) estimates an effect using the outcome Q99, a dummy variable indicating whether each paper has a FWCI above the 99th percentile of FWCI in the sample. Model (1) shows that while the Q99 outcome has an expected 1% occurrence in the sample (technically 1.21%[26]), corporate involvement is associated with an

[26]The Mean(DV) shown in the sample for quantiles does not exactly match the quantile number because

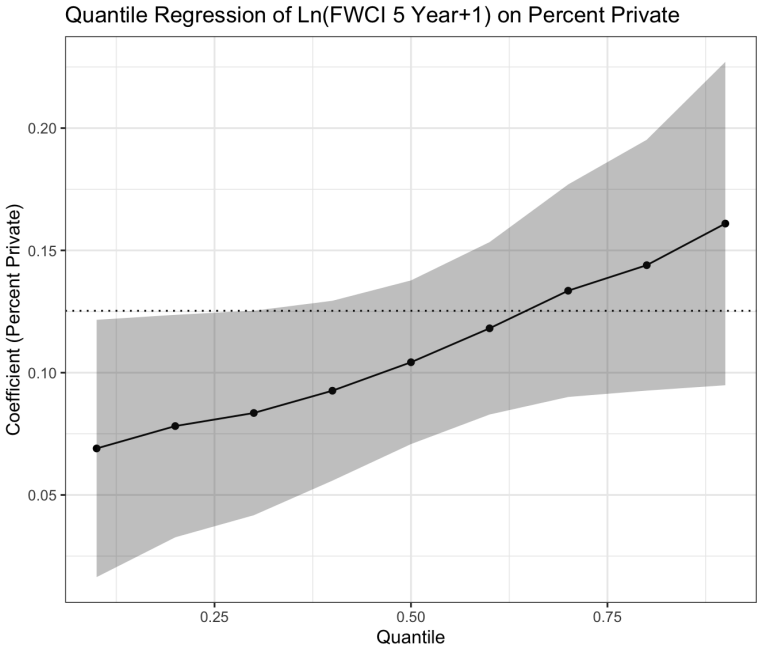


Figure 2. Quantile Regression Results

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| DV: | Q99 | Q95 | Q90 | Q75 | Q50 | Q25 |
| *Panel A. Y = Field-Weighted Citation Index (FWCI) 5y* | | | | | | |
| Percent Private | 0.0077** | 0.0312*** | 0.0341*** | 0.0472*** | 0.0586*** | 0.0499*** |
| | (0.0036) | (0.0072) | (0.0091) | (0.0125) | (0.0142) | (0.0123) |
| Mean(DV) | 0.0121 | 0.0579 | 0.1126 | 0.2715 | 0.5242 | 0.7673 |
| Ratio | 0.6364 | 0.5389 | 0.3028 | 0.1738 | 0.1118 | 0.0660 |
| Benchmark Ratio | 0.4100 | 0.2180 | 0.2850 | 0.2048 | 0.1238 | 0.0668 |
| *Panel B. Y = Citations 5y* | | | | | | |
| Percent Private | 0.0058* | 0.0286*** | 0.0275*** | 0.0421*** | 0.0447*** | 0.0514*** |
| | (0.0031) | (0.0066) | (0.0087) | (0.0119) | (0.0138) | (0.0125) |
| Mean(DV) | 0.0123 | 0.0582 | 0.1116 | 0.2653 | 0.5148 | 0.7564 |
| Author-Year FE | Y | Y | Y | Y | Y | Y |
| University FE | Y | Y | Y | Y | Y | Y |
| Team Controls | Y | Y | Y | Y | Y | Y |
| Observations | 47411 | 47411 | 47411 | 47411 | 47411 | 47411 |

*Notes* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The data is at the paper-author level. Robust standard errors clustered at the author-year and paper level are shown in parentheses. Panel A shows the estimation for the primary variable of interest: field-weighted citation index (FWCI). For interpretability, Panel B shows the same estimate, but done using Citations to the paper.

Table 3. Test of Hypothesis 2 with Quantile Dummy Indicator Outcomes.

of two factors. First, my main sample includes papers twice if there are two dual-affiliated authors on the paper (see Section 2.2 for discussion). Second, citation count is a discrete quantity, so there can be ties when assigning quantiles.

0.77% increase in absolute likelihood, equal to a 63.64% = 0.0077/0.0121 relative increase over baseline rate, as shown in the calculated Ratio row. In other words, papers with Percent Private = 1 are, on average, over 60% more likely to be a top 99% FWCI 'breakthrough' paper than papers with Percent Private = 0!

Beyond the striking effect magnitude described here, Table 3 show that the effect of Percent Private is more pronounced at the right-tail of the distribution than the middle. To see this, consider models (2) through (6), which repeat the same specification of Model (1) but use other FWCI quantile dummy indicators as outcomes. These estimates reveal a relative-to-baseline ('Ratio') effect of 63.64%, 53.89%, 30.28%, 17.38%, 11.18%, and 6.8% at (respectively) Q99, Q95, Q90, Q75, Q50, and Q25. We can compare these results to a simple benchmark ratio that I calculate based on if the effect shown here is constant. Specifically, I numerically simulate the equivalent regression specification here (quantile dummy on another dummy variable) for two normal distributions offset by a mean effect of 0.1557 of a standard deviation (the equivalent log10(FWCI+1) effect found in Table 2, Panel A, Model (2)). I then compute the ratio of coefficient to baseline rate and list it in the table as 'Benchmark Ratio'. This makes it easy to see that, relative to a constant effect, the effects here are much more pronounced at the right tail of the distribution (between Q95 and Q99) and are otherwise similar throughout the distribution.

Panel B, Models (1)-(6) repeat these estimates but use raw Citations rather than FWCI, showing that the qualitative story does not shift. Overall, Table 3 shows that while corporate involvement has a positive effect throughout the FWCI distribution, the effect is most pronounced at the right tail of the distribution.

These two tests provide strong supportive evidence of Hypothesis 2. I emphasize here that both effects presented here are very large in magnitude — corporate involvement likelihood of being a top 95 FWCI quantile paper by 53.89%, and increases the magnitude of the effect at Q90 by 28.8% (= 16.1%/12.53% - 1) relative to the mean effect. I conclude that, even beyond an average effect on citations, corporate involvement significantly increases the likelihood of a researcher producing a paper that is a breakthrough piece of research. This is particularly important to know if we think of science as being driven by the most cited papers rather than by incremental advances.

#### 3.2.2 Larger Effect for Involvement by Firms with More Resources

In order to test for heterogeneity with respect to firm resources (Hypothesis 3), I augment my primary specification by interacting my corporate involvement measure with various measures of firm resources in the year of publication. Recall from Section 2.4 that I measure firm resources by looking at variants of the number of papers produced by the same company in the same year of publication of a focal paper. In Table 4, Models (1) through (3), I present the same specification using three different measures of firm resources: distinct papers, distinct authors, and paper credits[27], where these measures have been standardized to aid interpretation (with standard deviations of 126 papers, 195 authors, and 56 credits). All three models demonstrate the same striking degree of heterogeneity of the corporate involvement effect: whereas the average effect remains similar (14.44% here compared to 12.53% from before), the average effect for a firm that is one up by standard-deviation of resources rises by over 19%. Given that the main effect is large and significant, it's obvious that the heterogeneity effects are even larger and more economically significant.

[27]Recall that a paper credit is weighted authorship, such that if one paper has two authors with one from a university U and the other from firm F, then U and F would each get half a credit for that paper.

| DV: ($Y$ = FWCI 5y) | (1) Ln($1+Y$) | (2) Q99 | (3) Q95 | (4) Q50 |
|---|---|---|---|---|
| Percent Private | 0.1440*** | 0.0101** | 0.0339*** | 0.0724*** |
| | (0.0262) | (0.0046) | (0.0087) | (0.0162) |
| Percent Private x Company Credits in Year | 0.1996*** | 0.0313*** | 0.0374* | 0.0963*** |
| | (0.0559) | (0.0115) | (0.0194) | (0.0293) |
| Company Credits in Year | −0.0316 | −0.0039 | −0.0043 | −0.0250 |
| | (0.0281) | (0.0053) | (0.0097) | (0.0168) |
| Team Controls | Y | Y | Y | Y |
| University FE | Y | Y | Y | Y |
| Author-Year FE | Y | Y | Y | Y |
| Mean(DV) | 0.8897 | 0.0121 | 0.0579 | 0.5242 |
| Observations | 47411 | 47411 | 47411 | 47411 |

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 4. Test for Resource Heterogeneity.

Further, this effect occurs throughout the distribution, including (most notably) at the right tail of the distribution: while the likelihood of being in the top 99th FWCI quantile increases on average by absolute 1.01% for papers with corporate involvement, it increases by an absolute 3.13% (!) for papers with involvement by firms with one standard deviation of credits; this constitutes a 258% *additional* likelihood of being a citation outlier for each standard deviation of credits.

### 3.2.3 Greater Alignment with Firm Commercial Interests

Finally, I test whether researchers choose projects that align more with a firm's commercial interests when working for a firm (Hypothesis 4). I test this by using the same specification but using dummy indicators for the academic subject of a paper. Though not very granular, the academic subject label is useful for testing a paper's alignment with firm commercial interests because of the narrow population focus of my analysis (dual-affiliated AI researchers). Specifically, because these researchers tend to work only with information technology, telecommunications, and semiconductor firm affiliations rather than pharmaceutical or biomedical firms (see Table B3), I can interpret subjects more easily in terms of alignment with a firm's commercial interests. Specifically, I interpret engineering disciplines as more aligned with a firm's commercial interests and physical science or social science disciplines as less aligned[28]. For example, I interpret a paper in a physics journal as

[28]For concreteness, I interpret the disciplinary labels with respect to alignment with a firm's commercial interests as follows: More Aligned/Engineering (Arts & Humanities (see footnote Footnote 29), Chemical Engineering, Computer Science, Energy, Engineering, Material Science), Less Aligned/Physical Sciences (Biochemistry, Chemistry, Environmental Sciences, Mathematics, Medicine, Neuroscience, Physics, Planetary Science), and Less Aligned/Social Sciences (Decision Sciences, Management, Social Sciences). Of course, this is only true in tendency rather than as an absolute fact. There are (of course) engineering papers that are *not* commercially aligned with firms and physical science papers that *are* commercially aligned with the firm. The argument here hinges on the assumption that subject dummies capture broad directional changes in that commercial alignment; therefore, if a dual-affiliated researcher collaborating with a firm-based team

an interdisciplinary effort to solve core problems in particle physics using machine learning techniques, a paper that is likely not aligned with the commercial interests of the firm.

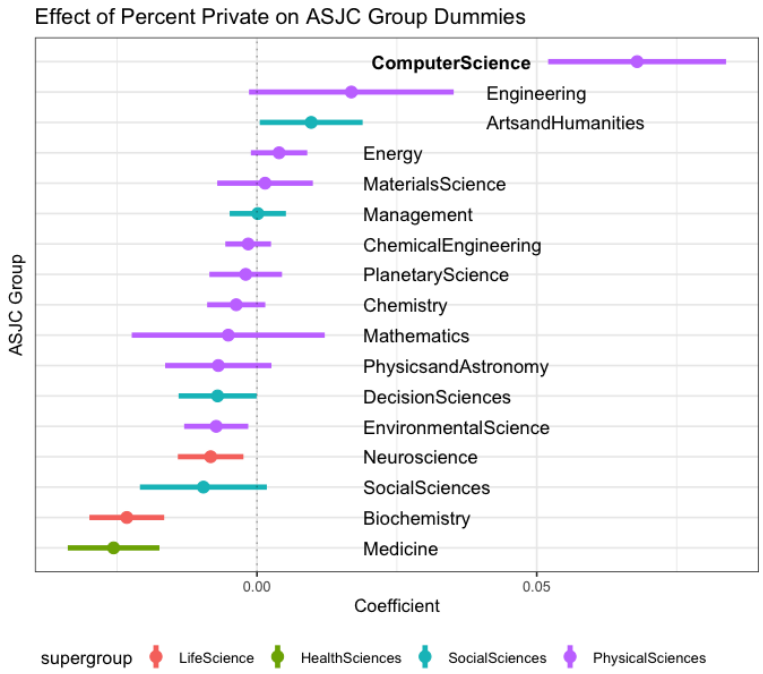


Figure 3. Effect of Corporate Involvement on ASJC Group.

I visualize the regression coefficients of my test in Figure 3, with each line corresponding to a different regression/outcome variable (an ASJC group subject dummy). Because the outcome variable is a dummy variable, the coefficients can be interpreted as the average absolute percent increase in the likelihood of working on that particular subject, given corporate involvement and controlling for researcher-year, university, and team experience. Even without presenting the baseline frequency values, a striking pattern immediately emerges: corporate involvement is associated with a much greater likelihood of working on engineering subjects like Computer Science (6.79%), Engineering (1.68%), and Humanities[29] (0.97%), and much less likely to work on physical science topics like Medicine (-2.56%), Biochemistry (-2.32%), or Physics (0.69%). Factoring in the baseline values, these numbers become more extreme: Medicine, Biochemistry, and Physics drop by a relative -53.2%, -74.0%, and -10.3%, respectively. I conclude that corporate involvement strongly affects the academic subject of research papers, causing them to shift away from interdisciplinary efforts and towards more core technical research that aligns with the firm's commercial interests.

Overall, the Section 3.2 presents strong evidence in favor of interpreting my main effect as driven by a firm's involvement in introducing both unique resources and constraints to researchers.

works more on (say) computer science than medicine, then we can say that corporate involvement associates with working on firm-aligned topics. However, if firm-affiliated researchers are shifting research fields from (say) computer science to medicine but remain working on problems of equal (dis-)interest to the firm in either case, then this measure doesn't actually test the hypothesis of interest.

[29]Note that this is due to the fact that many Natural Language Processing publication venues are coded as Arts and Humanities.

### 3.3 Selection, Robustness, and Other Stories

Could this collection of results be explained in some other way? As described in Section 2.3, the empirical strategy followed here deviates from the idealized theoretical model used to characterize the resource-constraint tradeoff described in Section 1, admitting the possibility that one of these deviations may be an explanation for the results presented to this point. To help with organizing this section, in Figure 4, I visualize the theorized mechanism (panel A) and the ways that my empirical strategy deviates from it. Panel B shows the primary endogeneity concern addressed by the empirical strategy pursued here — it eliminates the concern that differences in papers come from differences in researchers across universities and firms.

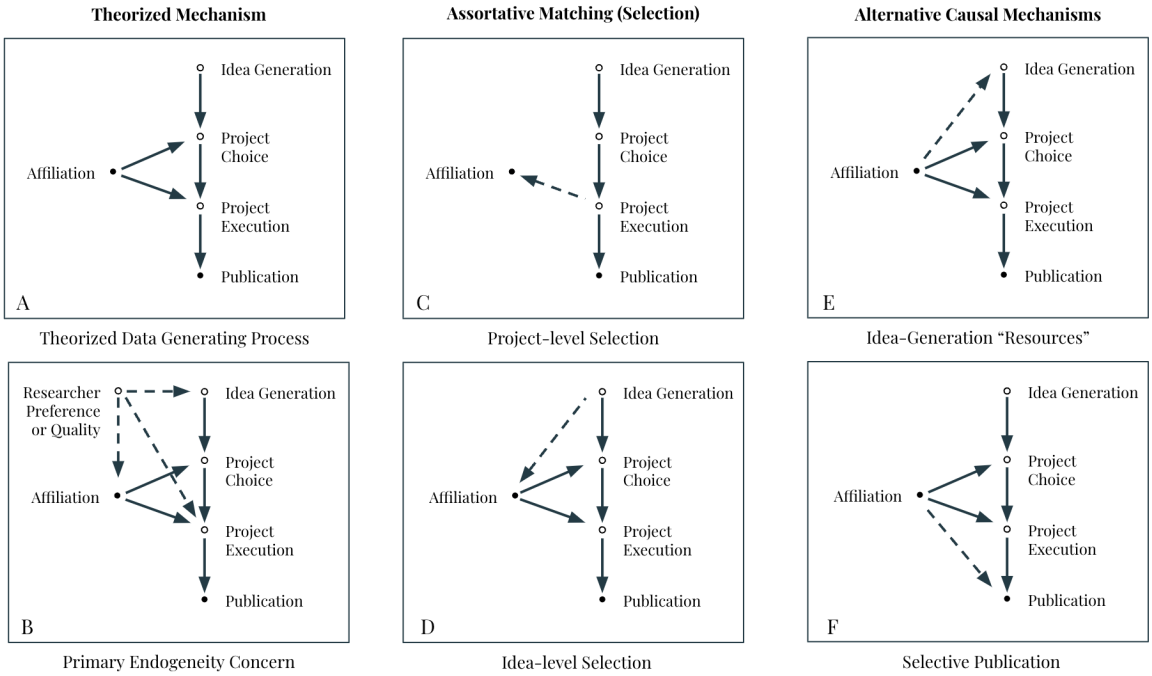


Figure 4. Directed Acyclic Graph representation of the theorized mechanism and possible deviations of this paper's empirical strategy from the idealized theoretical mechanism.

However, two types of deviations remain. In this section, I consider and rule out each of both types of deviations as possible alternative explanations of the corporate citation premium and the other supporting evidence presented before. I start by considering the possibility that a (dual-affiliated) researcher *matches* ideas to institutional environments (Figure 4, panels C and D). Second, I explore the possibility that these results are driven by an alternative causal mechanism apart from the resource-constraint tradeoff (Figure 4, panels E and F). Finally, I conclude with a series of robustness tests to ensure my estimates are not driven by the specifics of how I constructed my analysis sample or operationalized my variables.

#### 3.3.1 Assortative Matching (Selection)

One concern is that in the prior analysis, dual-affiliated researchers choose which institutions to work with at various points in the research pipeline and may do so strategically.

Could these choices explain the corporate citation premium? We consider three levels at which assortative matching may occur and their implications for interpreting results.

*Researcher-level Selection* — Researchers that select into dual-affiliation are different than the average AI researcher (see Section 2.2). This fact doesn't threaten the internal validity of the estimates on its own because the estimates are within-researcher-year rather than between researchers. Nevertheless, this selection does affect how we should think about the generalizability of the results to other researchers. In particular, authors that select into dual affiliation are more likely to have ideas that would be of interest to firms naturally and, therefore, would benefit more from firm involvement.

*Project-level Selection* — Could researchers execute a project and then strategically choose which institution to publish it under after the results are realized (Figure 4, panel C)? I find this unlikely for two reasons. The first is that research in AI is typically done in teams, and teams are typically formed within a particular institutional setting. For this reason, we expect that affiliation cannot be changed after a project is completed for the vast majority of publications due to the implicit contract formed between co-author teams that limits flexibility in which authors to include on the paper. Second, discussions with AI researchers suggest that dual-affiliated researchers carefully manage the boundaries of which projects belong to which environments due to concerns over corporate interference with ideas pursued in universities, again suggesting that affiliation is determined far before the completion of a project.

*Idea-level Selection* — The most plausible selection mechanism for explaining the corporate citation premium is the possibility that researchers generate ideas and then *choose* the institutional environment that they develop them within. I call this "idea-level selection" because, theoretically, it can be understood as a causal effect of generated ideas on the affiliations in which they are pursued, inducing spurious correlation between affiliation and citations (Figure 4, panel D).

Can idea-level selection explain the observed corporate citation premium (and other results)? For this to be the case, one would need to expect some form of *positive selection* of ideas into firms: that dual-affiliated researchers prefer to work on more scientifically valuable ideas in the firm relative to the university. To explore this idea more carefully, in Appendix C, I develop a simple selection model where a dual-affiliated researcher matches ideas to institutional environments based on various selection functions. The model shows that the estimated main effect can be decomposed into an "ATT" resource effect and a selection effect whose sign depends on the correlation of the selection function and the scientific value of ideas. I argue that the only realistic way to justify a *positive* selection effect is to assume that researchers prefer working on ideas in a firm environment *because* of access to resources that make those ideas scientifically more valuable. In other words, to justify positive selection, one has to assume the resource mechanism that I am arguing for in the first place.

Nevertheless, it is useful to know how much of the estimated main effect is driven by resources and how much is driven by idea-level selection. To this end, I develop two additional arguments supporting that idea-level selection does not significantly drive the observed main effect. First, I emphasize that my estimates already control for a key 'resource' that may lead to idea-level selection: teammate experience. Of course, this control may remove some causal variation of interest from my estimates: after all, the ability to collaborate with high-quality or interdisciplinary teammates may be part of the causal resource benefit of firms. But this control also helps alleviate the concern that dual-affiliated researchers may match ideas to teammates, who systematically vary across settings (Ph.D. trainees at uni-

versities versus the other PhDs and seasoned researchers at firms). Table C2 shows the direct effect of including teammate controls in the regression specification. By comparing Model (3) to Model (4), I see that controlling for all teammate experience variables, the effect of corporate involvement on my FWCI outcomes only moderately falls (e.g. 15.71% to 12.53% for the main outcome in Panel A). This relative stability of the coefficient between specifications provides confidence that one of the more salient channels of idea-level selection does not drive significant changes in the overall estimate.

As a second argument, I re-estimated my primary specification using only the subset of papers produced through *repeated collaborations*. This subset is useful because the ideas pursued in repeated collaborations are likely to be driven by ideas created in the context of prior research projects. Therefore, even if a first collaboration was originally driven by a researcher explicitly seeking out teammates to execute a specific project, the later collaborations are more likely to be driven by the institutional environment that sustains the research relationship (e.g. Figure 4, Panel E), rather than a selection effect (Figure 4, Panel D). Empirical implementation of these concepts amounts to filtering my sample down to papers where the focal author has prior collaborations with at least some of her co-authors and rerunning the same specification. I present the results of this analysis in Table C3. Panel A focuses on the set of papers where the author has at least one prior collaborator (based on the entire SCOPUS publication record, including papers out of sample). Model (1) shows that the sample restriction only moderately reduces the effect of corporate involvement on FWCI to 10.75%. Model (2) through (4) repeats the same exercise using other key outcome variables from the analysis, similarly showing the robustness of those effects. Further, considering the case where at least half or all of the coauthors are prior collaborators tells a similar story across all outcomes of interest (Panel B & C), though with more noise due to the smaller sample sizes. Overall, all of my results qualitatively hold as I restrict the sample to more stringent prior collaboration criteria, though some of my results lose statistical significance as the sample restrictions get more severe and my estimates get correspondingly noisy.

### 3.3.2 Alternative Causal Mechanisms

A second concern with the primary analysis is that it does not rule out alternative causal mechanisms, such as a causal effect of affiliation on idea generation or on publication itself. Indeed, this concern is not limited to just this paper's empirical strategy; even the ideal econometric scenario that randomizes researchers to firm or university environments would be similarly unable to rule out these alternative causal mechanisms based only on an analysis of an average treatment effect. In this section, I consider these alternative mechanisms as well as the possibility that the effect is driven by just one or two firms and discuss implications for the interpretation of results.

*Idea-Generation 'Resources'* — The theorized mechanism from Section 1 only discusses the effect of affiliation on project choice and project execution. However, corporate affiliation may also affect the quality of ideas that are generated in the first place (Figure 4, Panel E). Higher scientific value of ideas generated in a corporate setting may therefore explain the corporate citation premium. Empirically, my reduced-formed estimates of the main effect capture both mechanisms and are unable to distinguish between higher-quality ideas and higher-quality execution of those ideas.

Nevertheless, while theoretically intriguing[30], for most managerial and policy implica-

[30]The notion that working with firms leads to exposure to technical problems that can inspire scientific

tions of interest, disentangling these effect is irrelevant and can be left to future work. Therefore, the most straightforward conceptual way to handle this alternative explanation is to argue that it is actually the same explanation; in other words, improved idea generation may itself be a 'resource' from firms, though one that may theoretically differ from the type modeled in Section 1. The main advantage of the more parsimonious model developed before is that it highlights specific empirical implications that can be tested in Section 3.2, but that does not preclude other broader ways of modeling resources.

*Selective Publication* — There is ample heuristic evidence that firms block the publication of certain papers after they are completed (e.g. Simonite (2020)); could firms suppress the publication of research after it is completed in a way that explains the corporate citation premium (Figure 4, Panel F)? For this to be the case, firms would have to suppress *lower quality* research. However, it is far more likely that firms suppress publication on the basis of *firm* value rather than scientific value. And, conditional on an idea being developed in a firm, there is no reason to suspect that *firm* value correlates with lower scientific value. Furthermore, selective publication could not explain the results of mechanism tests developed in Section 3.2. I therefore find this explanation unlikely.

*Is the Effect Driven by Google Alone?* — Recall that in Figure B2, I observed that a few firms comprise the majority of top AI conference publications with corporate involvement. Further, I saw in Table 4 that there is significant heterogeneity in my estimated effect as a function of firm scientific resources. Could this indicate that my estimated effect of corporate involvement on citations is driven by just one or two firms?

To explore this possibility, I modify my regression specification to include interactions of my corporate involvement measure with firm-specific dummies, presenting my results in Figure C1[31]. Estimating firm-specific effects is challenging due to the lower sample size per firm leading to significant noise in effect estimates for most firms. However, the figure nevertheless reveals a large variation in the effect of specific firms on scientific quality, allowing us to draw two qualitative conclusions. First, while Google is certainly contributing to the estimated effect (with an estimated effect on Ln(FWCI+1) of approximately 30%), many (over 20) companies contribute to the overall positive effect, including Chinese companies like Sensetime, Tencent, and Baidu. Second, while the effect is not a Google-only effect, it does appear that positive involvement is largely associated with larger technology companies and not companies in other industries (such as Raytheon BBN, Siemens, or Motorola).

Beyond confirming that the main effect is driven by many companies, Figure C1 shows that involvement by some firms leads to a much greater effect on scientific impact than by other firms. However, this does not imply that some firms are not benefiting by hiring these scientists because firms are able to change the research topics that researchers are working on. In other words, citations are the wrong metric by which to measure firm appropriation of value by this strategy. Nevertheless, a greater understanding of this firm-level variation presents an opportunity for future work.

### 3.3.3 Robustness of Estimates to Measurement and Sample Choice

Beyond specific concerns about selection or alternative causal explanations, my results could be driven by any number of the decisions I made in the process of operationalizing

ideas is only lightly discussed in the prior literature, largely in qualitative discussions of the success of researchers in corporate labs in the 20th century (Rosenberg 1982).

[31] Notably, each of these regression coefficients is estimated in comparison to the rest of the (non-listed) firms in the sample

my theory in my empirical setting. To reduce concerns regarding the robustness of results to these decisions, I conducted further tests where I varied my measurement approach or analysis sample. I provide a high-level summary of each class of robustness test and direct interested readers to Appendix C for full details of the results.

*Corporate Involvement Measure* — This analysis depends heavily on the ability to measure corporate involvement at a paper level. In my preferred specification, I use a continue measure (*Percent Private*) which I derive from the affiliation tags associated with each author on the paper. However, my results are robust to alternative operationalizations of this measure. To show this, I run my preferred specification using two classes of alternative measures: binary measures and using the dual-author's affiliation only (Appendix Table C4). In Panels B through E, I show that my core results are robust to measuring corporate involvement as a binary variable (with varying cutoff thresholds) along all outcome measures used in the analysis. Direction and statistical significance hold for all results except one (Q99 using the Binary Corporate Involvement Measure with a Percent Private $> 0.0$ cutoff). In Panel F, I show that my main results hold for a specification run using only variation in the Dual Affiliated Author's own affiliation tag.

*Academic Subject Measure* — One concern with the use of the ASJC Classification system as a way to operationalize alignment with firm commercial interests is that it is assigned at the journal level rather than the paper level. To test the robustness of this measure, I repeat my test but using an alternative paper-level subject classification — the Science Metrix classification system. Science Metrix has developed a machine learning system to classify academic subjects at a paper level, with full details provided in their complementary publication (Rivest et al. 2021). I visualize the results of this test in Appendix Figure C2. The figure shows that corporate involvement is positively conditionally associated with subject measures like 'Information Technology' and 'Strategic Technology' (which have subcategories Energy, Materials, Nanoscience, and other applied engineering fields) and is negative conditionally associated with subject measures like Clinical Medicine, Physics, and Biomedical Research. In other words, using the paper-level SMC classification does not change the qualitative conclusions of my test.

*Timeframe of Analysis* — Is the effect that I am observing driven by the long timeframe of my analysis (with papers as early as 1963)? For example, I may be concerned that the effect of corporate involvement on science was greater in the 20th century when there were large corporate labs like the labs. To ensure that it's not older papers that are driving the positive effect of corporate involvement on citations that I observe, I repeat my analysis but filter my sample to include only papers published after 2000 or 2010. I show these estimates in Appendix Table C5, Panels A and B. The results show that my estimates do not decrease and, in fact, increase by a significant amount. Whereas my original estimate of 12.53% was seen as large, when filtering to papers that were published only after 2010, I actually observed the effect increase to 19.33% (maintaining statistical significance despite the fact that my sample size is cut in half).

*Inclusion of Papers with Multiple Dual Authors* — Finally, I test the robustness of my results to my decision to run analysis at the researcher-paper level, which allowed papers that had two dual-affiliated authors to be included in my sample two times[32]. I rerun my results after filtering my sample to remove one of the two paper authors; in Appendix Table C5, Panel C, I assign the paper to the more productive author (based on publications in my analysis sample) and remove the author-paper corresponding to the less productive

[32]Recall that I exclude papers with more than two dual-affiliated authors on them outright

author. In Panel D, I do the opposite. Neither sample restriction meaningfully changes my results.

# 4 Discussion and Conclusion

## 4.1 Generalization Beyond AI Research

I argue that the resource-constraint tradeoff characterizes a *general* mechanism that drives a causal impact of firm involvement on scientific quality in other fields of science beyond AI. But if this resource-constraint mechanism is so important for other fields, why hasn't it been highlighted in the literature before? The argument that AI research is a 'special case' that does not generalize to other fields of research is unappealing because it doesn't illustrate *why* AI is a special case. While direct evidence of the generality of this mechanism is beyond the scope of this paper (which focuses mainly on AI research), I develop three further explanations for why this mechanism hasn't been reported in the prior literature, bolstering the case for generalization.

The first reason is the lack of causal identification. This paper contributes a novel empirical strategy focused on *dual-affiliated* researchers that enables controlling away the effects of researcher selection into firms based on topic preference or researcher quality. However, dual-affiliated researchers are not nearly as common in other fields of science, and researchers have yet to find other convincing natural experiments that introduce plausibly exogenous variation between university and firm environments.

A second, and perhaps even more basic reason, is the lack of systematic data. To demonstrate this, I provide suggestive evidence that the positive benefit of corporate involvement in science is not just limited to AI (or CS) research. In Figure 5, I plot the uncontrolled five-year citation percent difference between papers with and without corporate involvement across different fields. Specifically, I gather data from the top 10 publication sources for each ASJC group in my data[33], and define corporate involvement as having *Percent Private* $> 0.5$. As expected, given my preceding analysis, Computer Science shows a large positive association between corporate involvement and average citations: papers with corporate involvement average 200% (3x) more citations. Further, as expected based on prior theory, many basic disciplines like Physics, Chemistry, and Earth Science show a negative association between corporate involvement and average citations. However, most surprisingly, there are a number of other fields that show a positive association between corporate involvement and average citations, contrary to prior theory. This includes more commonly studied fields like Pharmaceuticals, Biochemistry, and Medicine. As these differences are completely uncontrolled, I cannot draw conclusions from this figure alone. Nevertheless, I take this as suggestive evidence that firms providing unique resources may be a first-order mechanism operating in other fields of science beyond AI research and strongly encourage future research to understand the variation observed here.

As a third and final reason, the effect of firm resources may vary across scientific fields, and we lack an understanding of the underlying conditions. However, my conceptual framework provides boundary conditions for the mechanism. It highlights that the effect of firm involvement on a given field of science depends on the amount of unique resources that firms can provide for basic research relative to the constraints imposed by the firm. The constraints may be the easiest to reason about — as highlighted in the theoretical model,

[33] See Footnote 23 for details on the data and ASJC groups.

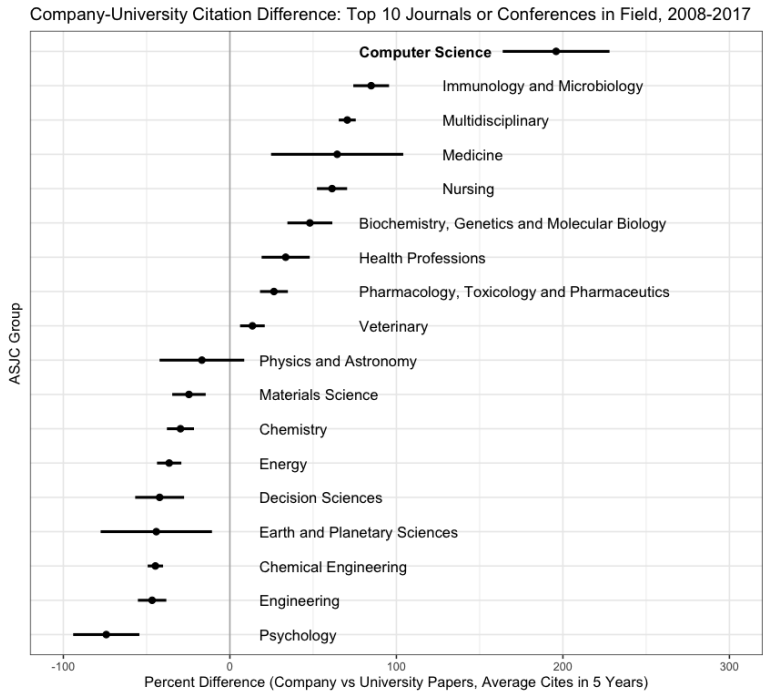


Figure 5. Five-year Citations Percent Difference between Papers with and without Corporate Involvement. The result is shown across every 2-digit ASJC Group Subfield with at least 1% of papers with corporate involvement (leading to the exclusion of many social sciences and arts/humanities fields, where firms are not very active). See Appendix B.3 for details on the construction of this dataset.

one way to think about firm constraints is as (the inverse of) the extent to which firm value and scientific value are correlated. When fields of science investigate problems that more often fall into 'Pasteur's Quadrant', we expect that firm constraints will be less relevant and fewer resources will be required for firms to have a positive impact.

What are these resources? Due to the limitations of the data available, I cannot comment quantitatively on the nature of these resources in this setting. However, this question provides a promising avenue for follow-up work. For now, I briefly highlight external evidence that suggests the likely candidates for these resources: datasets and computing, human capital resources (e.g. talented engineers), and novel technical problems. Prior research articles leveraging publications data and manual coding efforts have largely focused on datasets & computing, the most quantifiable difference in the models produced by industry relative to academia (Thompson et al. 2023). More subtly, AI researchers have noted that large-scale projects often require the creation of complexly engineered systems that require high technical skill and long timeframes, even though the engineering work does not constitute AI research directly. These types of projects may be uninteresting to AI researchers but can be pursued given sufficient access to engineering resources (Togelius and Yannakakis 2023). Finally, as suggested by the model distillation example from the introduction as well as other more informal discussions, working in close proximity to product teams provides a novel perspective into theoretical problems that may be otherwise ignored by those working in universities (Rosenberg 1982; AAAI 2020).

Knowing what scientific resources firms are able to uniquely provide is crucial to understanding how well this mechanism generalizes to other fields of science because the relevance of specific resources to science depends heavily on the topic of research. Therefore, the effect may be smaller (and therefore harder to statistically isolate) in more commonly studied fields like biology, chemistry, and pharmaceuticals[34] due to firms having less relevant resources. For example, the ability to push candidate drugs through clinical trials may not lead to fundamental insights in pharmaceutical research compared to the ability to access large datasets in artificial intelligence research. Nevertheless are plenty of other fields of science beyond AI where, at least anecdotally, access to corporate resources greatly helps research. For example, in the social sciences, the ability to experimentally vary website experiences for millions of people at a time via digital platforms similarly vastly exceeds the ability to experiment at scale in university contexts (e.g. Blake et al. (2021); Rajkumar et al. (2022)).

### 4.2 Contributions to the Literature on University-Firm Relations

I highlight three ways that this paper contributes to the innovation literature on the university-firm relationship. The main contribution of this paper is to be the first large-scale quantitative study of any field of science to demonstrate a direct positive effect of corporate involvement on scientific quality and to describe the underlying mechanism. The case of AI research documented here provides a counter-example to a widely-held belief that corporate involvement has predominantly negative effects on science due to the constraints that firm involvement places on research topics (see Foray and Lissoni (2010) or Perkmann et al. (2013) for a summary), which provides the basis for present thinking on the university-firm relationship in innovation (Arora et al. 2001, 2018). I show empirically that corporate involvement not only adds constraints but also resources. For the case of AI research, these resources provide a greater benefit than the costs of the firm's added constraints, leading to up to a 44% increase in field-weighted citations received and a significant increase in the likelihood of a breakthrough paper. More generally, this mechanism helps answer the puzzle of why firms are so often the source of fundamental breakthroughs in basic science: when firm constraints do not bind but firms have unique scientific resources, they are uniquely positioned to contribute fundamental breakthroughs in science. In identifying this mechanism, this paper joins a small, recent group of papers calling for a more nuanced understanding of the difference between university and firm environments for scientists and the development scientific ideas (Sauermann and Stephan 2013; Bikard 2020; Bikard and Marx 2020; Nagle and Teodoridis 2020; Marx and Hsu 2022), and an even smaller group asking for conditions when firms can provide a helpful, complementary organizational environment to universities in the pursuit of basic knowledge (Azoulay et al. 2009; Bikard et al. 2019; Hartmann and Henkel 2020).

A second contribution to this literature is to provide a novel reason for the significant amount of researcher selection into industry observed in other studies in the AI field (Jurowetzki et al. 2021; Gofman and Jin 2023). Whereas the prior literature frames selection into industry as the outcome of scientists individually evaluating a tradeoff between scientific freedom and higher wages (Stern 2004; Sauermann and Stephan 2013), I show that scientists may choose to work in industry simply because it is a better place to do research. The effect of researcher selection on the relationship between corporate affiliations and ci-

[34] Previous research in the university-firm relations literature and the science-of-science more generally focuses largely on these fields due to their large size and economic importance.

tations is dramatic: whereas the raw difference in citations between university and firm papers is nearly 200% in top conferences, this difference shrinks to between 12% to 44% after controlling for researcher-year fixed effects. This implies that a (relatively) small resource differential between universities and firms can lead to large differences in the quality of researchers at universities versus firms due to the best researchers seeking positions at firms[35].

A third contribution of this paper follows from considering where the resources that drive a positive effect of firm involvement in this context come from. These resources arise as a byproduct of firms' commercial activities in the form of large data sets, engineering talent, and insight into novel technical problems. I showed in this paper that these resources lead to greater scientific impact by the researchers working at these firms. Therefore, the third contribution of this paper is to highlight this unappreciated complementarity between commercial activity and the production of basic scientific knowledge, which suggests that we should *expect* firm prominence in certain fields of science. This type of 'reverse causality', where technology improves basic scientific research, has been largely unappreciated in the literature despite anecdotal evidence arising from corporate laboratories in the 20th century (Rosenberg 1982). The contribution here is to suggest that this complementarity may be a first-order mechanism at work in important fields of scientific inquiry[36].

### 4.3 Managerial and Policy Implications

*Policy Implications* — Should we be concerned about the impact of firm involvement on these AI researchers? Policymakers think so: the highly visible activity of firms in AI research has prompted concern and (ultimately) policy proposals in response to the possibility of negative effects. The general rationale has been articulated by Stanford's Human-centered Artificial Intelligence group as related to inequity of access to resources leading to a socially inefficient allocation of scientific resources (Ho et al. 2022). In July 2021, the US Government released a Request For Information (RFI) for implementation plans for a National AI Research Resource (Register 2021). Companies have responded by recommending funding at over $500MM/year for the creation of a shared national infrastructure for AI researchers (Kaye 2021).

This paper's results have important implications for this policy proposal. Specifically, the current proposed implementation for National AI resources may be inefficient because, by providing blanket general resources for the entire population of AI researchers, such government funding may crowd out corporate involvement in basic AI research. Instead, my results suggest that an effective policy intervention would target the stimulation of research based on the direction of research. Specifically, resources should target projects situated at the nexus of AI research and specific fields of application that are of particular societal interest but not of interest to the firm (such as applications in the earth, physical, or medical sciences). This type of implementation may help to mitigate the downsides of

[35]This type of thinking can be formalized through the use of Roy-style selection models.

[36]Notably, the idea that unique resources enable firms to have superior productivity is unoriginal. The strategy literature (Wernerfelt 1984; Barney 1991; Hartmann and Henkel 2020) has long noted that unique resources are a key source to competitive advantage and has even pointed out its connection to firm dominance in AI research. This study is, however, the first to quantitatively show that firms have resources that can uniquely advance science rather than profitability and therefore represents a novel form of complementarity between commercial activity and scientific research. This connection to the strategy literature is useful because it provides a rich, developed literature from which I can better reason about how firms develop and maintain these unique resources.

corporate involvement (lost projects) while maintaining the upside (private money funding the production of basic science research, a public good).

*Managerial Implications* — Finally, these results have a direct strategic implication for managers of scientifically oriented firms. An outstanding puzzle in innovation strategy is understanding why some firms publish influential papers rather than privately capturing their value. For example, in 2017, Google famously published the Transformer architecture, spurring a revolution in Natural Language Processing. This was no doubt essential to the development of modern AI applications, but did Google "give up the golden goose" in making this public rather than keeping it as a trade secret used to improve their search and other products? The reasons highlighted in the prior literature (e.g. Rotolo et al. (2022)) do not seem sufficient to justify making a major research breakthrough like the Transformer public when the firms like Google can so clearly benefit from keeping it private. Further, there seems to be a divergence of strategies of leading tech firms. Google and Facebook hire researchers and allow them to publish much of their work, but Apple and Amazon, despite hiring researchers, typically do not allow those researchers to publish.

This paper's results begin to suggest a novel economic reason for firm involvement in science: because firm involvement can open up new areas of study that university researchers will subsequently build on. If firms typically have a *negative* impact on follow-on research, then it may make more sense for firms to allow other institutions like universities to coordinate basic research and instead focus on applying those insights to technical problems. However, this paper shows that firms can have a *positive* impact on follow-on research by supporting researchers and allowing them to publish. In doing so, firms may be effectively focusing the follow-on interest of the scientific community on problems with relevance to the firm in a way that accelerates the firm's own innovation agenda — a knowledge *spillback*. Of course, the benefit of this knowledge spillback must be held in tension with the cost of knowledge spillouts to rivals that increases market competition, perhaps explaining why some firms pursue open science while others hire researchers but do not allow them to publish. While these results only suggest this concept, exploring this strategy further provides an exciting and promising avenue for future research on corporate involvement in science.

Regardless of why firms are publishing these important research breakthroughs, we are all fortunate that sometimes firm value does align with scientific value and that firms *can* play a unique role in providing the resources needed to advance scientific research in ways that are useful to our society.

## References


AAAI (2020). AAAI 20 oxford style debate academic AI research in an age of industry labs.

Aghion, P., Dewatripont, M., and Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics*, 39(3):617–635. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-2171.2008.00031.x.

Arora, A., Belenzon, S., and Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, 39(1):3–32. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.2693.

Arora, A., Belenzon, S., Patacconi, A., and Suh, J. (2020). The changing structure of american innovation: Some cautionary remarks for economic growth. *Innovation Policy and the Economy*, 20:39–93. Publisher: The University of Chicago Press.

Arora, A., Fosfuri, A., and Gambardella, A. (2001). *Markets for Technology: The Economics of Innovation and Corporate Strategy*. The MIT Press, first edition edition.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. pages 609–626. Publication Title: The Rate and Direction of Inventive Activity Section: The Rate and Direction of Inventive Activity.

Azoulay, P., Ding, W., and Stuart, T. (2009). The impact of academic patenting on the rate, quality and direction of (public) research output*. *The Journal of Industrial Economics*, 57(4):637–676. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6451.2009.00395.x.

Azoulay, P., Furman, J. L., Krieger, and Murray, F. (2015). Retractions. *Review of Economics and Statistics*, 97(5):1118–1136.

Azoulay, P., Graff Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public r&d investments and private-sector patenting: Evidence from NIH funding rules. *The Review of Economic Studies*, 86(1):117–152.

Azoulay, P., Graff Zivin, J. S., and Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3):527–554. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-2171.2011.00140.x.

Azoulay, P., Stuart, T., and Wang, Y. (2014). Matthew: Effect or fable? *Management Science*, 60(1):92–109. Publisher: INFORMS.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2023). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, page 89.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1):99–120.

Baruffaldi, S., van Beuzekom, B., Dernis, H., Harhoff, D., Rao, N., Rosenfeld, D., and Squicciarini, M. (2020). Identifying and measuring developments in artificial intelligence: Making the impossible possible. Series: OECD Science, Technology and Industry Working Papers Volume: 2020/05.

Bikard, M. (2020). Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8):1528–1543. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3162.

Bikard, M. and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66(8):3425–3443.

Bikard, M., Vakili, K., and Teodoridis, F. (2019). When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. *Organization Science*, 30(2):426–445. Publisher: INFORMS.

Blake, T., Moshary, S., Sweeney, K., and Tadelis, S. (2021). Price salience and product choice. *Marketing Science*, 40(4):619–636. Publisher: INFORMS.

Blumenthal, D., Campbell, E. G., Anderson, M. S., Causino, N., and Louis, K. S. (1997). Withholding research results in academic life science: Evidence from a national survey of faculty. *JAMA*, 277(15):1224–1228.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 535–541. Association for Computing Machinery.

Bush, V. (1945). Science: The endless frontier. *A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development.*

Czarnitzki, D., Grimpe, C., and Toole, A. A. (2015). Delay and secrecy: does industry sponsorship jeopardize disclosure of academic research? *Industrial and Corporate Change*, 24(1):251–279.

Dasgupta, P. and David, P. A. (1994). Toward a new economics of science. *Research Policy*, 23(5):487–521.

David, H. A. and Nagaraja, H. N. (2004). *Order Statistics*. John Wiley & Sons. Google-Books-ID: bdhzFXg6xFkC.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, pages 67–82. PMLR. ISSN: 2640-3498.

Ferreira, K. J., Parthasarathy, S., and Sekar, S. (2022). Learning to rank an assortment of products. *Management Science*, 68(3):1828–1848. Publisher: INFORMS.

Foray, D. and Lissoni, F. (2010). University research and public–private interaction. *Handbook of the Economics of Innovation*, 1:275–314. Publisher: Elsevier.

Ganguli, I. (2017). Saving soviet science: The impact of grants when government r&d funding disappears. *American Economic Journal: Applied Economics*, 9(2):165–201.

Gertner, J. (2013). *The Idea Factory: Bell Labs and the Great Age of American Innovation*. Penguin Books, reprint edition edition.

Ghorbel, M., Pineau, J., Gourdeau, R., Javdani, S., and Srinivasa, S. (2018). A decision-theoretic approach for the collaborative control of a smart wheelchair. *International Journal of Social Robotics*, 10(1):131–145.

Girotra, K., Terwiesch, C., and Ulrich, K. (2010). Idea generation and the quality of the best idea. *Management Science*, page 16.

Gofman, M. and Jin, Z. (2023). Artificial intelligence, education, and entrepreneurship.

Hartmann, P. and Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries*, 6(3):359–381. Publisher: Academy of Management.

Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship. *Econometrica*, 52(4):909–938. Publisher: [Wiley, Econometric Society].

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018a). Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Number: 1.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. (2018b). Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 123–129. Association for Computing Machinery.

Hill, R. and Stein, C. (2020). Scooped! estimating rewards for priority in science. *Working Paper*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.

Ho, D., King, J., Wald, R., and Wan, C. (2022). Building a national AI research resource: A blueprint for the national research cloud. *Stanford HAI*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589. Number: 7873 Publisher: Nature Publishing Group.

Jurowetzki, R., Hain, D., Mateos-Garcia, J., and Stathoulopoulos, K. (2021). The privatization of AI research(-ers): Causes and potential consequences – from university-industry interaction to public research brain-drain?

Kaye, K. (2021). AWS, google and microsoft want in on a national AI research cloud, but google wants first crack at the data. - protocol.

Kernighan, B. (2019). *UNIX: A History and a Memoir*. Kindle Direct Publishing.

Keum, D. D. and See, K. E. (2017). The influence of hierarchy on idea generation and selection in the innovation process. *Organization Science*, 28(4):653–669. Publisher: INFORMS.

Krimsky, S. (2004). *Science in the Private Interest: Has the Lure of Profits Corrupted Biomedical Research?* Rowman & Littlefield. Google-Books-ID: W_qKbgQ5XbwC.

Lacetera, N. and Zirulia, L. (2012). Individual preferences, organization, and competition in a model of r&d incentive provision. *Journal of Economic Behavior & Organization*, (2):550–570.

Machado, J. A. F. and Santos Silva, J. M. C. (2019). Quantiles via moments. *Journal of Econometrics*, 213(1):145–173.

MacroPolo (2020). The global AI talent tracker.

Marx, M. and Hsu, D. H. (2022). Revisiting the entrepreneurial commercialization of academic science: Evidence from "twin" discoveries. *Management Science*, 68(2):1330–1352. Publisher: INFORMS.

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Metz, C. (2021). *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. Dutton.

Mullis, K. (2000). *Dancing Naked in the Mind Field*. Vintage, reprint edition edition.

Murghia (2019). AI academics under pressure to do commercial research. *Financial Times*.

Murray, F. (2010). The oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. *American Journal of Sociology*, 116(2):341–388.

Murray, F., Aghion, P., Dewatripont, M., Kolev, J., and Stern, S. (2016). Of mice and academics: Examining the effect of openness on innovation. *American Economic Journal: Economic Policy*, 8(1):212–252.

Nagle, F. and Teodoridis, F. (2020). Jack of all trades and master of knowledge: The role of diversification in new distant knowledge integration. *Strategic Management Journal*, 41(1):55–85.

Nelson, R. R. (1959). The simple economics of basic scientific research. *Journal of Political Economy*, 67(3):297–306. Publisher: University of Chicago Press.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A., and Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy*, 42(2):423–442.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., and Aral, S. (2022). A causal test of the strength of weak ties. *Science*, 377(6612):1304–1310. Publisher: American Association for the Advancement of Science.

Register, F. (2021). Request for information (RFI) on an implementation plan for a national artificial intelligence research resource.

Rivest, M., Vignola-Gagne, E., and Archambault, E. (2021). Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. *PLOS ONE*, 16(5):e0251493. Publisher: Public Library of Science.

Roach, M. and Sauermann, H. (2010). A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry. *Research Policy*, 39(3):422–434.

Rock, D. (2021). Engineering value: The returns to technological talent and investments in artificial intelligence. *SSRN Electronic Journal*.

Rosenberg, N. (1982). *Inside the Black Box: Technology & Economics, Chapter 7: How Exogenous is Science?* Cambridge University Press.

Rotolo, D., Camerani, R., Grassano, N., and Martin, B. R. (2022). Why do firms publish? a systematic literature review and a conceptual framework. *Research Policy*, 51(10):104606.

Sauermann, H. and Stephan, P. (2013). Conflicting logics? a multidimensional view of industrial and academic science. *Organization Science*, 24(3):889–909. Publisher: INFORMS.

Senoner, J., Netland, T., and Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*, 68(8):5704–5723. Publisher: INFORMS.

Simonite, T. (2020). The dark side of big tech's funding for AI research. *Wired*. Section: tags.

Smith, N. (2021). The dream of bringing back bell labs.

Stephan, P. E. (1996). The economics of science. *Journal of Economic Literature*, 34(3):1199–1235.

Stern, S. (2004). Do scientists pay to be scientists? *Management Science*, 50(6):835–853. Publisher: INFORMS.

Thompson, N., Ahmed, N., and Wahed, M. (2023). The influence of industry in AI research. *Science*.

Togelius, J. and Yannakakis, G. N. (2023). Choose your weapon: Survival strategies for depressed AI academics.

Torvik, V. I., Weeber, M., Swanson, D. R., and Smalheiser, N. R. (2005). A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 56(2):140–158. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20105.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, 5(2):171–180.

# A Theory Details

This appendix section contains details and proofs for the theory described in Section 1.

## A.1 Proofs of Hypotheses

The key probability insight that justifies the modeling approach is the observation that for a firm researcher, $S_p$ is distributed as a *concomitant* of an order statistic, specifically the maximum (David and Nagaraja 2004). This insight is valuable because results on order statistics and their concomitants can be determined analytically for the two-variable case where the variables are linearly related and share the same marginal distribution. Formally, I assume $S_p = \mu_S + \rho\frac{\sigma_S}{\sigma_F}(F_p - \mu_F) + \epsilon$, where $\epsilon$ is an arbitrary mean-zero noise distribution independent of $F_p$ that preserves the marginal distribution[37], and $\mu_S, \mu_F, \sigma_S, \sigma_F$ are standard parameterizations of the mean and variance of the joint distribution. I note that because the marginal distribution of $F_p$ is unconstrained, this distribution is still quite general. Denote the $r$th order statistic of scientific value as $S_{r:n}$, and $F_{r:n}$ for firm value. The concomitant of the $r$th order statistic of firm value is the scientific value associated with that idea, denoted as $S_{[r:n]}$. Using this notation, I am interested in comparing the random variables $S_{p^*_{\text{Univ}}} = S_{1:n}$ and $R \times S_{p^*_{\text{Firm}}} = R \times S_{[1:n]}$.

Recall that I am interested in comparing the distributions of $S_{1:n}$ (University Researcher) and $R \times S_{[1:n]}$ (Firm Researcher), henceforth using the notation of order statistics and concomitants. To shorten proofs, I first write down the first and second moments of these random variables (which are standard results of the concomitants literature) and reference them later (David and Nagaraja 2004):

$$
\begin{aligned}
E[S_{1:n}] &= \sigma_S \left( \frac{E[F_{1:n}] - \mu_F}{\sigma_F} \right) + \mu_S \\
E[S_{[1:n]}] &= \rho\sigma_S \left( \frac{E[F_{1:n}] - \mu_F}{\sigma_F} \right) + \mu_S \\
&= E[S_{1:n}] - (1-\rho)\sigma_S \left( \frac{E[F_{1:n}] - \mu_F}{\sigma_F} \right) < E[S_{1:n}] \\
\text{Var}(S_{1:n}) &= \sigma_S^2 \left( \frac{\text{Var}(F_{1:n})}{\sigma_F^2} \right) \\
\text{Var}(S_{[1:n]}) &= \sigma_S^2 \left( \rho^2 \frac{\text{Var}(F_{1:n})}{\sigma_F^2} + (1-\rho^2) \right) \\
&= \text{Var}(S_{1:n}) + \sigma_S^2(1-\rho^2) \left( 1 - \frac{\text{Var}(F_{1:n})}{\sigma_F^2} \right) > \text{Var}(S_{1:n})
\end{aligned}
$$

I now formally state and prove the following claims:

**Hypothesis 1.** $R > R^*_{Avg} \iff E[RS_{p^*_{Firm}}] > E[S_{p^*_{Univ}}]$. *If firm resources are sufficiently large, then on average, the scientific quality of papers produced by a firm researcher will be greater than that of an (otherwise identical) university researcher.*

Using the notation of concomitants, $R > R^*_{\text{Avg}} \iff E[RS_{[1:n]}] > E[S_{1:n}]$.

[37] To ensure $\mu_S$ and $\sigma_S^2$ keep their typical interpretations, I require $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = (1-\rho)\sigma_S^2$. I require that $S_p$ and $F_p$ have the same marginal distribution for analytic tractability.

*Proof.* Because of the linearity of expectations, I can just isolate $R$ to derive the threshold. Plugging in the corresponding expectation values gives the exact threshold.

$$R \times E[S_{[1:n]}] - E[S_{1:n}] > 0$$
$$\implies R > \frac{E[S_{1:n}]}{E[S_{[1:n]}]} \equiv R^*_{\text{Avg}} > 1$$

□

Note that $R^*_{\text{Avg}} > 1$ because $E[S_{1:n}] > E[S_{[1:n]}]$ . (Intuitively, as $\rho$ decreases, $E[S_{[1:n]}]$ decreases and $R^*_{\text{Avg}}$ increases.)

**Hypothesis 2.** $R > R^*_{QuantileEffect} \iff (Q_{95}[RS_{p^*_{Firm}}]) - Q_{95}[S_{p^*_{Univ}}]) > (Q_{50}[RS_{p^*_{Firm}}] - Q_{50}[S_{p^*_{Univ}}])$. *If firm resources are sufficiently large, then the average effect of firm involvement on scientific quality will be greater at the right tail (e.g. the 95th quantile) of the distribution than at the median.*

I.e., $R > R^*_{\text{QuantileEffect}} \iff (Q_{95}[RS_{[1:n]}]) - Q_{95}[S_{1:n}]) > (Q_{50}[RS_{[1:n]}] - Q_{50}[S_{1:n}])$.

*Proof.* I prove this for the normal marginal distribution[38]. Recall that

$$Q_{95}(X) = E[X] + 2\sqrt{\text{Var}(X)}$$

for normally distributed $X$. Then the LHS of the inequality that I want to prove can be written as

$$(E[RS_{[1:n]}] + 2\sqrt{\text{Var}(RS_{[1:n]})}) - (E[S_{1:n}] + 2\sqrt{\text{Var}(S_{1:n})})$$

Rearranging, I can write this as

$$(E[RS_{[1:n]}] - E[S_{1:n}]) + 2(R\sqrt{\text{Var}(S_{[1:n]})}) - \sqrt{\text{Var}(S_{1:n})})$$

But the first term is exactly the RHS (where I take advantage of the fact that $Q_{50}[X] = E[X]$ for normally distributed $X$), and the second term is positive if and only if

$$R > \sqrt{\text{Var}(S_{1:n})}/\sqrt{\text{Var}(S_{[1:n]})} \equiv R^*_{\text{QuantileEffect}}$$

□

I observe that $R^*_{\text{QuantileEffect}} < 1$ because $\text{Var}(S_{1:n}) < \text{Var}(S_{[1:n]})$. This means that the variation reduction from selection is precisely the quantity that drives the greater effect for greater baseline values. The difference between right-tail outcomes at firms and universities is less than at the mean *even when universities have greater resources than firms.*

I can gain further intuition for this quantile effect by asking a variant of the above hypothesis: At what resource threshold is there an effect at the 95th quantile? Formally, I can show that

$$(E[RS_{[1:n]}] + 2\sqrt{\text{Var}(RS_{[1:n]})}) - (E[S_{1:n}] + 2\sqrt{\text{Var}(S_{1:n})}) > 0 \iff R > R^*_{\text{Q95}}$$

[38] With an unrestricted marginal distribution of scientific value, exact quantile results on the order statistics are not available. Instead, I use results on the 95th quantile of the normal distribution here as an approximation. This logic extends easily to any quantile where the value is a linear function of $R$.

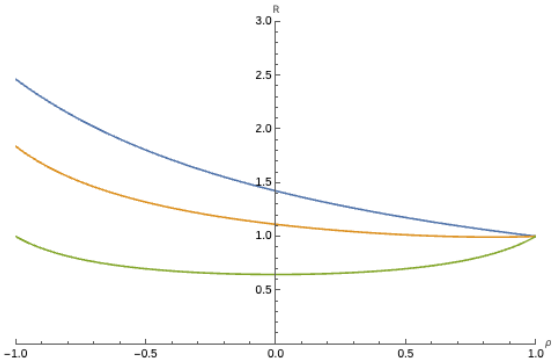


Figure A6. Resource Thresholds for Hypothesis 1 and Hypothesis 2 as a function of $\rho$. Blue depicts $R^*_{\text{Avg}}$, Orange depicts $R^*_{\text{Q95}}$, and Green depicts $R^*_{\text{QuantileEffect}}$.

for some $R^*_{\text{Q95}}$. The proof of this is almost identical to the proof of Hypothesis 1. I find the threshold by isolating $R$ and then plugging in the corresponding moment expressions.

$$R > \frac{E[S_{1:n}] + 2\sqrt{\text{Var}(S_{1:n})}}{E[S_{[1:n]}] + 2\sqrt{\text{Var}(S_{[1:n]})}} \equiv R^*_{\text{Q95}}$$

I note that because $E[S_{[1:n]}] < E[S_{1:n}]$ and $\text{Var}(S_{1:n}) < \text{Var}(S_{[1:n]})$, this implies that[39] $R^*_{\text{Avg}} > R^*_{\text{Q95}} > 1 > R^*_{\text{QuantileEffect}}$. This implies that the condition for the right tail of the firm researcher's scientific quality distribution to exceed the right tail of the university researcher is a weaker condition than for there to be a positive average effect. As before, this effect is also driven by the fact that $\text{Var}(S_{1:n}) < \text{Var}(S_{[1:n]})$, due to university researcher selection reducing variation in the observed scientific quality (but firm researcher selection not reducing the variation as much).

To visualize these conditions, I plotted $R^*_{\text{Avg}}$ and $R^*_{\text{Q95}}$ as a function of $\rho$ assuming a multivariate normal joint distribution with arbitrary parameters, shown in Figure A6. As one might expect, as $\rho$ decreases, $R^*_{\text{Avg}}$ gets much larger. Interestingly, $R^*_{\text{Q95}}$ does not increase much over 1 for most values of $\rho$ — implying that even when firm resources do not dramatically exceed that of universities, the best firm papers can still exceed the quality of the best university papers.

**Hypothesis 3.** $\partial_R(E[RS_{p^*_{Firm}}] - E[S_{p^*_{Univ}}]) > 0$. *The average effect of firm involvement on scientific quality will be greater when a researcher is involved at a firm with more resources.*

Equivalently, $\partial_R(E[RS_{[1:n]}] - E[S_{1:n}]) > 0$.

*Proof.* The derivative equals $E[S_{[1:n]}]$ because of the linearity of expectations, and I assumed that $S_p > 0$. □

**Hypothesis 4.** $E[\theta_{p^*_{Firm}}] > E[\theta_{p^*_{Univ}}]$. *The papers produced by firm researchers will, on average, be more aligned with their firm's commercial interests than the research produced by an (otherwise identical) university researcher.*

[39] Observe that if $A/B > 1$ and $A, B > 0$, then $A/B > (A + C)/(B + C) > (A + C')/(B + C)$, where $C > C' > 0$.

Equivalently, $E[\arctan(F_{1:n}/S_{[1:n]})] > E[\arctan(F_{[1:n]}/S_{1:n})]$.

*Proof.* By definition of order statistics, $F_{1:n} > F_{[1:n]}$ and $S_{1:n} > S_{[1:n]}$. Thus, $F_{1:n}/S_{[1:n]} > F_{[1:n]}/S_{1:n}$ for all possible realizations of project candidates. Further, because arctan is monotonic, $\arctan(F_{1:n}/S_{[1:n]}) > \arctan(F_{[1:n]}/S_{1:n})$. Finally, since this is true in all cases, this must be true in expectation as well. □

### A.2 Discussion of Modeling Assumptions

*Assumptions in modeling firm constraints* — To model firm constraints as the effect of corporate affiliation on project selection, I assume a particularly simple selection function for firm researchers. Although $G$ is fairly general, I do restrict $G$ to a linear relationship between $S_p$ and $F_p$ because it is necessary; given general non-linearity in $G$, I could not derive any hypothesis because the model would be too flexible. Nevertheless, I do not restrict the correlation $\rho$, and I believe that the linear relationship assumed in $G$ is still quite unrestrictive.

*Assumptions in modeling firm resources* — This analysis assumes that resources affect scientific quality multiplicatively. However, it is easily shown that the results hold if I think of resources as entering linearly or as any (positive) affine transformation of the baseline scientific value of an idea. More interestingly, I may alternatively think of resources as project- or field-dependent. However, I leave these possibilities unmodelled for theoretical simplicity.

*Exclusion Restrictions on Corporate Affiliation* — A more subtle (related) assumption is that I exclude firm involvement from impacting other aspects of the research process. For example, it may also be reasonable to assume firm involvement impacts the idea-generation process. I do this for analytic tractability, though I also note that because of the flexible form of $G$, this assumption is relatively innocuous. Nevertheless, I further discuss this possibility in Section 3.3.

Similarly, I assume that corporate affiliation does not affect researchers' ability to publish their work and that they can do so in a relatively unconstrained manner (given a selected project). I justify this assumption with the observation that tens of thousands of papers are published by firm researchers in my sample. However, I acknowledge that my model does not explain *why* firms do this. A robust literature on the topic (Rotolo et al. 2022) highlights various reasons that firms choose to publish, including increased ability to access ideas from scientific communities, increased ability to recruit scientific talent, and improved firm reputation.

Finally, I note that I do not explicitly model the productivity of scientists (that is, the number of papers produced). While this outcome is certainly of scientific and policy importance, I am not able to test such hypotheses in my empirical setting and leave such results for future work.

## B Details of Datasets Used in the Paper

In this appendix section, I describe the process for deriving the various datasets used in this paper from SCOPUS. I use data from top AI conferences to describe the empirical setting in Section 2.1, as well as a way to identify dual-affiliated authors. From the set of dual-affiliated authors, I derive my analysis sample (papers by dual-affiliated researchers) used throughout the paper (described in Section 2.4). Finally, I use a broader sample of

papers from top journals across fields in Section 4.1. I also provide extended details on my method for cleaning and validating the affiliations found in the sample.

## B.1 Top AI Conferences

In order to define artificial intelligence (AI) as a field of research, I take advantage of labels and rankings of computer science conferences provided by the Computing Research and Education Association of Australasia (CORE). I used the CORE rankings and labels because they are created by domain experts and have been validated before in the literature (see Baruffaldi et al. (2020), section 2.2). Using the 2021 rankings, I filtered to conferences that were ranked as 'A*' ("Exceptional") conferences[40], and focused only on conferences labeled as Artificial Intelligence, Machine Learning, or Computer Vision. See Table B1 for details on the conferences included in the sample. While the filtering method used here is quite restrictive in the sense that these are very prominent conferences, it achieves my goal of narrowing down my sample to an elite set of AI researchers. By credibly narrowing my sample to a more limited community of researchers in AI, I increase the validity of my understanding of the institutions they are working within, such as dual-affiliated employment.

Using this list, I set out to find papers corresponding to these conferences in the SCOPUS database. Unfortunately, the database does not consistently connect year-to-year conference proceedings and group them into distinct entities. Instead, I rely on a semi-manual process of string matching the title of each paper's explicit conference string (in SCOPUS, they call this the 'source title') to a set of substrings associated with my conferences of interest, typically the conference abbreviation. In order to exclude false-positive matches, I further reviewed the identified set of papers and wrote exclusions based on other keywords. For example, the International Conference on Automated Planning and Scheduling (ICAPS) is a conference included in my sample, but I *exclude* papers from the ICAPS Workshop, which are intended to be a spot for earlier stage, less developed ideas. Ultimately, I was able to find data on 16/18 of the conferences in my list in SCOPUS. This is consistent with the degree of coverage quoted in Baruffaldi et al. (2020), which examines coverage of top CS conferences across various publications databases.

The data comprises 82,359 papers; the distribution of 'Year Published' and 'Percent Private Affiliation' are shown in Figure B1. The 'Year Published' plot (left) is given as an empirical cumulative density. Papers were published as far back as 1984 (the 22nd Annual ACL meeting). However, the vast majority of the data comes more recently; over 95% of the data comes after 2000, and over 75% of the data comes after 2010. The 'Percent Private Affiliation' plot is given as a bar plot, but with the y-axis given in log scale. Just under 80% of the data has only university affiliation (Percent Private=0), while only 4.35% has pure firm affiliation (Percent Private=1). The rest of the data falls in between, representing mixed affiliation teams.

This process for cleaning the affiliation labels in this data is described at length in Appendix B.4. To validate the affiliation labels and the overall data sample, I filter this sample to the years 2008-2017, aggregate to the affiliation level, and compute the number of 'Paper Credits' associated with each affiliation, where credit is a weighted count of publications by the number of authors on it (so a paper with one Google author and one New York University author would count as a half-credit for both Google and NYU). The results are plotted in Figure B2 and briefly described in Section 2.1.

[40] For descriptions of the CORE conferences rankings, see here

| Abbreviation | Conference Name | CORE Field | In SCOPUS |
|---|---|---|---|
| AAAI | National Conference of the American Association for Artificial Intelligence | Artificial Intelligence | ✓ |
| AAMAS | International Joint Conference on Autonomous Agents and Multiagent Systems | Artificial Intelligence | ✓ |
| ACL | Association of Computational Linguistics | Artificial Intelligence | ✓ |
| ACMMM | ACM Multimedia | Computer Vision | ✓ |
| COLT | Conference on Learning Theory | Machine Learning | × |
| CVPR | IEEE Conference on Computer Vision and Pattern Recognition | Computer Vision | ✓ |
| EC | ACM Conference on Economics and Computation | Artificial Intelligence | ✓ |
| ECCV | European Conference on Computer Vision | Computer Vision | × |
| ICAPS | International Conference on Automated Planning and Scheduling | Artificial Intelligence | ✓ |
| ICCV | IEEE International Conference on Computer Vision | Computer Vision | ✓ |
| ICDM | IEEE International Conference on Data Mining | Machine Learning | ✓ |
| ICLR | International Conference on Learning Representations | Machine Learning | ✓ |
| ICML | International Conference on Machine Learning | Machine Learning | ✓ |
| IJCAI | International Joint Conference on Artificial Intelligence | Artificial Intelligence | ✓ |
| KDD | ACM International Conference on Knowledge Discovery and Data Mining | Machine Learning | ✓ |
| KR | International Conference on the Principles of Knowledge Representation and Reasoning | Artificial Intelligence | ✓ |
| NeurIPS | Advances in Neural Information Processing Systems | Machine Learning | ✓ |
| WSDM | ACM International Conference on Web Search and Data Mining | Machine Learning | ✓ |

Table B1. A* Conferences from the CORE 2021 Rankings in the Categories of Artificial Intelligence (4602), Computer Vision (4603), or Machine Learning (4611).

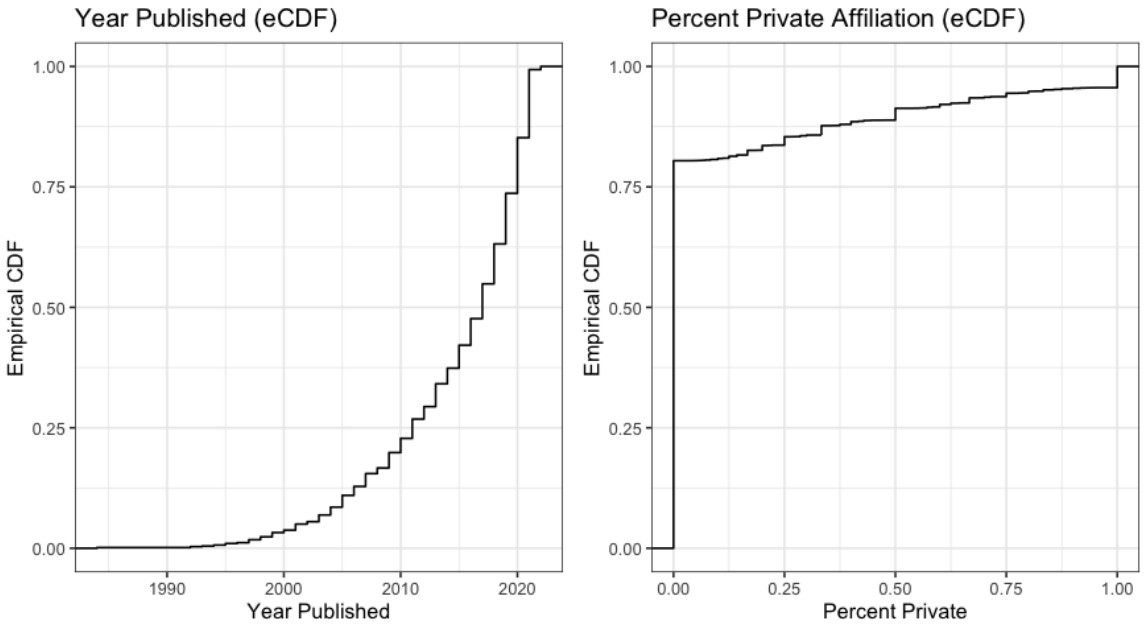


Figure B1. Distribution of Key Variables in Top Conferences Data.

Another use of this data occurs in Figure 1, where I visualize the distribution of 5-year citations for papers in this dataset for papers with vs. without Percent Private $> 0.5$. That figure reveals a fact that motivates the overall paper: that corporate papers tend to receive more citations, and that this difference extends into the right tail of the distribution.

## B.2 Papers by Dual-Affiliated Researchers

I now aim to identify dual-affiliated authors as a subset of the broader set of authors of papers in top AI conferences (derived from the dataset on top AI conferences from

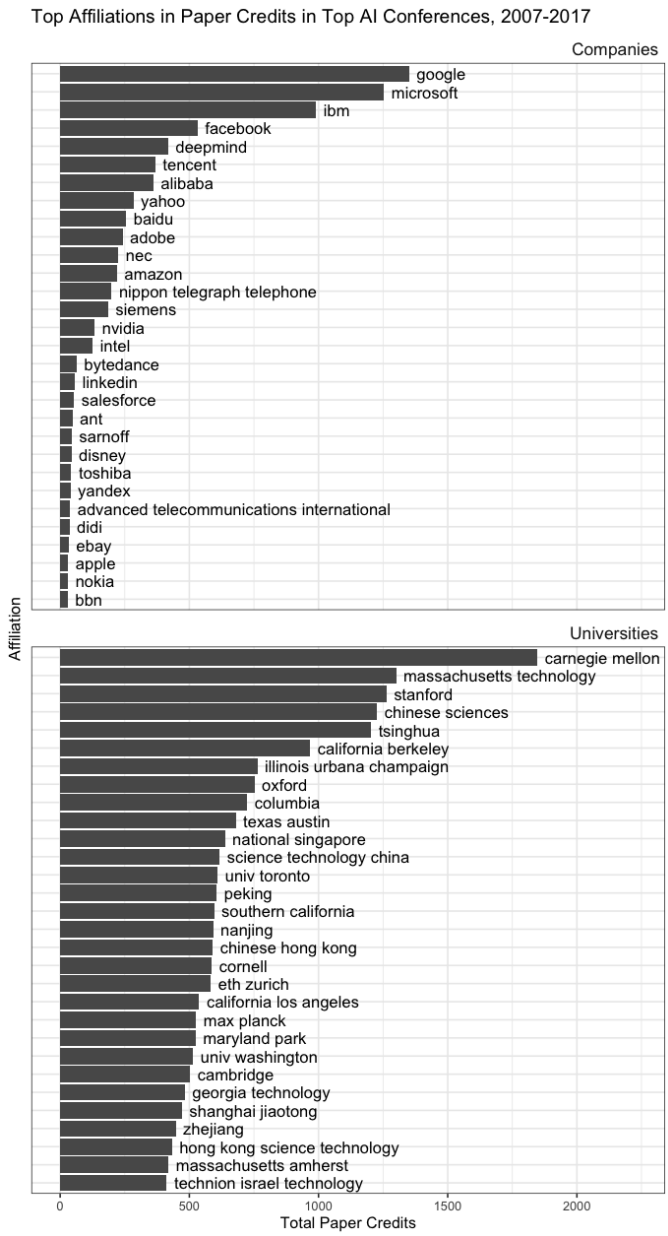


Figure B2. Company and University Involvement in Top AI Conferences. This figure visualizes the extent of Company (Top) and University (Bottom) involvement in top computer science conferences.

Appendix B.1). I expand my dataset to include any publication by any of these top AI conference authors, including papers outside of top AI conferences. Within this expanded set, I look for papers with authors that have both a company and a university affiliation on that same paper. I call this a *dual paper*, the corresponding author a *dual-affiliated author*, and the year of publication a *dual year* (for that author). Using this method, I identify 3965 dual-affiliated researchers. I list some summary statistics on these authors in Table B2 and describe them briefly in Section 2.2.

| Variable Group | Variable | Min | Q25 | Q50 | Mean | Q75 | Max | Std. Dev. |
|---|---|---|---|---|---|---|---|---|
| Publications History | First Year Publishing | 1964 | 1993.00 | 2002.00 | 2000.03 | 2008.00 | 2017.00 | 10.46 |
| | Last Year Publishing | 1986 | 2019.00 | 2021.00 | 2019.55 | 2022.00 | 2023.00 | 4.27 |
| | Total Years Publishing | 1 | 10.00 | 16.00 | 17.82 | 25.00 | 55.00 | 10.14 |
| | Total Publications | 1 | 23.00 | 56.00 | 105.78 | 132.00 | 2183.00 | 154.79 |
| | Average Publications per Year | 1 | 2.20 | 3.40 | 4.79 | 5.59 | 76.30 | 4.93 |
| Any Corporate Involvement | First Year Publishing with any Company Involvement | 1964 | 1998.00 | 2006.00 | 2004.21 | 2012.00 | 2017.00 | 9.44 |
| | Last Year Publishing with any Company Involvement | 1976 | 2015.00 | 2020.00 | 2017.46 | 2021.00 | 2022.00 | 5.66 |
| | Count of Publications (Percent Private > 0) | 1 | 6.00 | 15.00 | 26.92 | 34.00 | 641.00 | 36.82 |
| | Count of Publications (Percent Private > 0.5) | 0 | 2.00 | 7.00 | 14.69 | 17.00 | 389.00 | 24.20 |
| | Average Percent Private | 0 | 0.05 | 0.14 | 0.22 | 0.34 | 0.99 | 0.22 |
| Dual Affiliation | Count of Years with Dual Affiliation | 1 | 1.00 | 1.00 | 1.94 | 2.00 | 24.00 | 1.93 |
| | First Year with Dual Affiliation | 1964 | 2003.00 | 2009.00 | 2007.80 | 2015.00 | 2017.00 | 8.10 |
| | Last Year with Dual Affiliation | 1966 | 2005.00 | 2012.00 | 2009.96 | 2016.00 | 2022.00 | 8.08 |
| | Count of Dual-Affiliated Publications | 1 | 1.00 | 1.00 | 3.30 | 3.00 | 242.00 | 8.16 |
| Citations | Median Citations 5y | 0 | 3.00 | 6.00 | 9.63 | 10.00 | 714.00 | 18.88 |
| | Max Citations 5y | 0 | 50.00 | 107.00 | 328.48 | 252.00 | 56543.00 | 1529.89 |
| Subjects | Percent Papers with ASJC Group = Computer Science | 0 | 0.71 | 0.85 | 0.77 | 0.92 | 1.00 | 0.23 |
| | Percent Papers with ASJC Group = Medicine | 0 | 0.00 | 0.00 | 0.04 | 0.02 | 0.97 | 0.11 |

Table B2. Summary Statistics on Dual-Affiliated Authors. This table presents summary statistics on 3965 dual-affiliated authors identified through the process described in Section 2.2.

Although this process relies heavily on SCOPUS's author identifiers, I argue that it is *conservative* in terms of the identification of dual-authors, in the sense that the papers associated with dual-authors are genuinely by those authors (low 'false positive' matches). Because author disambiguation is a known challenge in scientific databases, several custom solutions have been developed, such as the Author-ity method developed for biomedical publications (Torvik et al. 2005). No such methods exist (to my knowledge) in the AI context. Instead, I rely on SCOPUS's own internal systems for disambiguating authors at the time that they index articles for inclusion in the database. SCOPUS does this by using additional information beyond name, including email, affiliation, subject area, citations, and co-authors, as well as community disambiguation efforts like ORCID. Nevertheless, SCOPUS is conservative in this process, in the sense that two papers by the same author are much more likely to be coded as by different authors ('false negative') than two different authors being coded as the same author ('false positive'). Further, they provide a custom web interface for researchers to notify SCOPUS of errors and request to merge different author profiles if the algorithm is found to have been overly conservative[41].

A conservative matching algorithm means that I will (with high certainty) be comparing papers by the same author. Nevertheless, I do need to make a missing-at-random assumption regarding unlinked papers that end up excluded from the analysis.

Using this set of dual-affiliated authors, I create my analysis sample by identifying any papers by a dual-affiliated author published with a dual-year. That is, if a dual-affiliated author published a dual paper in 2013 but not in 2014, then I only include other papers by that author from 2013. I do this because my empirical strategy is focused on comparing papers from the same year by the same author but with variations in the affiliations of the co-authors on those papers. I filter this sample with two final criteria: A) I remove papers with more than 10 co-authors to increase interpretation that the dual-affiliated author contributed meaningfully to the paper and B) I only include journal articles and conference papers. This step eliminates a wide variety of other cited objects that typically don't represent new validated scientific knowledge like books, pre-prints, reviews, or editorials.

[41] For the claims in this paragraph, see Elsevier's online documentation, including a description of author profiles and a blog post about efforts to merge SCOPUS IDs with ORCID

In total, this leads to a sample of 77,847 papers. 47,411 (60.9%) were published before 2018, meaning they have sufficient time for five-year citations to materialize. I keep the post-2017 papers for the purpose of testing non-citation-based outcomes like academic subjects. The summary statistics of this data are presented in Table 1 and are described in Section 2.4. I further visualize the Year Published and Percent Private Affiliation in Figure B3. Here, we can easily see that while the year-published distribution is comparable to the top conferences, there is much more firm involvement, reflecting the fact that we're focused on authors who we selected because they work with industry.

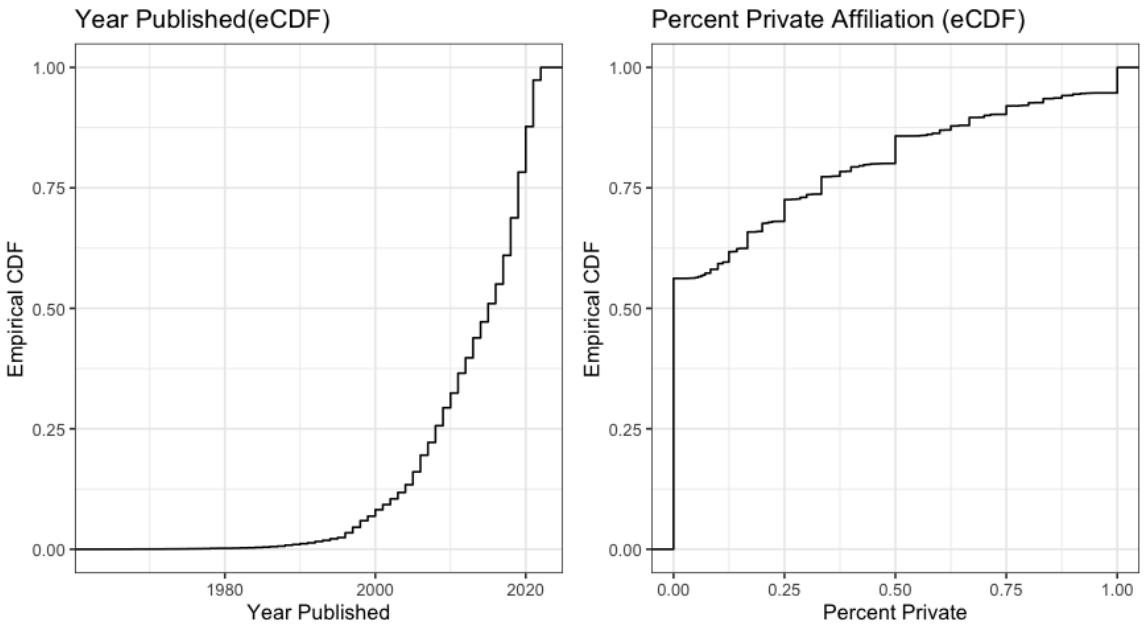


Figure B3. Distribution of Key Variables in Analysis Sample Data.

The most frequent company affiliations[42] found in the sample were comparable to the companies found in Figure B2 but with slightly different frequencies (reflecting that some companies publish more in other venues than top AI conferences). A table listing the most frequent company affiliations is provided in Table B3. Notably, the vast majority of these companies are IT or Telecommunications companies. There are several companies with direct e-commerce businesses (Amazon, Alibaba), but they tend to be conglomerates with many businesses.

## B.3 Top Journals Across Fields

In order to generate a broader set of data to test the association of corporate involvement and citations across fields of science, I created a dataset comprised of publications from 'Top Journals'. To define this set, I took the following steps:

- I created a list of all journals or conference publications in each field, as defined by the 2-digit ASJC coding in SCOPUS (documentation).

[42] Again, see Appendix B.4 for details on how these affiliation labels were cleaned and validated.

| Rank | Weighted Paper Credits | Company | Industry |
|---|---|---|---|
| 1 | 2432 | Google | Information Technology |
| 2 | 2122 | IBM | Information Technology |
| 3 | 2062 | Microsoft | Information Technology |
| 4 | 1153 | Facebook | Information Technology |
| 5 | 1026 | ATR | Telecommunications |
| 6 | 852 | DeepMind | Information Technology |
| 7 | 831 | Tencent | Information Technology |
| 8 | 548 | NVIDIA | Semiconductors |
| 9 | 541 | Nippon Telegraph Telephone | Telecommunications |
| 10 | 456 | Siemens | Conglomerate |
| 11 | 454 | Alibaba | Conglomerate |
| 12 | 432 | Baidu | Information Technology |
| 13 | 414 | Amazon | Conglomerate |
| 14 | 359 | Adobe | Software |
| 15 | 317 | Intel | Information Technology |
| 16 | 266 | Sensetime | Software |
| 17 | 264 | Samsung | Telecommunications |
| 18 | 256 | Disney | Conglomerate |
| 19 | 229 | NEC | Information Technology |
| 20 | 203 | TNO | Defense |
| 21 | 195 | Yahoo | Information Technology |
| 22 | 168 | Hewlett Packard | Information Technology |
| 23 | 145 | AIXTRON | Semiconductors |
| 24 | 142 | Salesforce | Software |
| 25 | 127 | ANT | Conglomerate |

Table B3. Top 25 Most Frequent Firm Affiliations in Dual-Affiliate (Analysis) Sample

- I used a standard journal ranking score to identify the top 10 journals or conference publications of each field. (SCOPUS provides a 'Scimago Journal Ranking' for all journals and conferences, so this was straightforward).
- I then pulled all papers from these publications between the years 2008 and 2017.

The result was a database of 2.24MM papers, including papers from the most recognizable journals in science (Nature, Science, and Cell), mathematics (Annals of Mathematics, Annals of Statistics, Journal of the American Statistical Association), and social science (Quarterly Journal of Economics, American Sociological Review, and Management Science).

Percent Private affiliation was calculated in the same way as described in Section 2.4. Corporate involvement is defined as having at least half of the authorship team having a corporate affiliation on the paper of interest. The result was visualized in Figure 5 and was described in Section 4.1.

## B.4 Cleaning and Validation of Affiliation Entities

A key element of the data-cleaning pipeline used in this paper was the aggregation and validation of the affiliation entities provided by SCOPUS. Fortunately, SCOPUS assigns affiliations to each paper-author according to how they appear on publications and does some aggregation of this information into distinct affiliation 'entities' (in SCOPUS, this is called an 'afid') and labeling of their type (private, university, non-profit, etc.). However,

I needed to be confident in the quality of these affiliation entities and labels because I rely heavily on them for my analysis. This usage includes computing my corporate involvement measure, controlling for university fixed effects, and calculating the resources associated with various firms. In this section, I describe the problem with the default SCOPUS affiliation labels and my solution.

The problem is similar to that of author disambiguation (described in Appendix B.2) — SCOPUS is overly conservative in the merging of distinct affiliation entities at the time of data ingestion. This technical problem is compounded by a conceptual one: unlike for individual authors, it's unclear to what 'level' of organization the affiliations should be aggregated. For example, should aggregation of Harvard be done at the level of the university, Harvard's Computer Science Department, or even labs within that department? Regardless of the answer, SCOPUS is inconsistent in its level of aggregation across organizations, reflecting inconsistency in each author's own listing of affiliations on their publications.

My conceptual solution was to aggregate at the 'brand name' level: This means that Columbia Statistics and Columbia Computer Science are both considered part of Columbia University, but DeepMind and Google are considered separate entities though they are (at least post-2015) financially part of the same organization (Alphabet). I chose this for three reasons:

1. This matches how we may typically think about access to resources; by being an employee or faculty of an institution, you may have an institution login and the ability to apply for additional resources within that organizational structure.

2. This is a relatively high level of aggregation, so I can consistently group to this level even though SCOPUS was inconsistent in its default aggregation. It would have been quite challenging to unbundle SCOPUS's default groups into a lower level of detail.

3. This was also convenient from an implementation perspective, given that the main information that I had to work with was the string of the affiliation written by authors on each paper.

My algorithm worked as follows.

1. I mapped SCOPUS's more granular 'organization type' (associated with each affiliation ID) into a more high-level taxonomy of 'acad' (including law schools, medical schools, colleges, or research institutes), 'priv' (comprised of companies), and 'publ' (including political organizations, funding organizations, military, non-profits).

2. I *cleaned* the affiliation strings given at the paper-author level by:

    - converting to ascii characters only (using `unidecode`), and converting to lowercase.
    - removing generic stopwords (from `nltk.corpus`) as well as a custom list of affiliation stopwords like 'academy', 'institute', 'group', and 'limited'.
    - swapping out common aliases of test substrings based on a database that I developed through iterative visual inspection. For example, I replaced 'deep mind' with 'deepmind' for consistency or replaced 'mit' with 'massachusetts technology'.
    - specially handling universities or colleges named after specific places. Specifically, I added back in the term 'univ' or 'college'.

3. I *collapsed* the affiliation strings by leveraging the fact that short affiliation labels (e.g. 'stanford') tended to be more common than longer related affiliation labels (e.g. 'stanford computer science'). Specifically, with each organization type (e.g. 'acad'), I ordered cleaned affiliation strings by how frequent they were. For each cleaned string, I checked if there was a more frequent cleaned affiliation string that was a substring of that focal string. If so, I swapped it in.

4. Finally, I aggregated to the (affiliation string, organization type) level and counted the number of publications associated with each entity. For a small number of entities, I had to manually change them (for example, Microsoft Research had been labeled as a 'research institute' in SCOPUS, but I changed it to be a private company).

By sorting and visually inspecting the results of this algorithm, it was straightforward to check the outcome for accurate labeling. I developed this algorithm iteratively by inspecting the results of the algorithm and adding corrections, stop words, and aliases until I couldn't find any errors. My validations comprised of expanded versions of Figure B2, which allowed me to verify the correctness of the labels and the meaningfulness of the cleaned affiliation strings created by my algorithm. Importantly, my validations focused heavily on the most represented institutions in the data, so if there were inaccurate coding of affiliation type, then they would only occur for less common affiliations are are therefore highly unlike to drive the results.

# C Analysis Extensions and Robustness

This section contains a series of additional tests useful for supporting the empirical arguments developed in the paper.

## C.1 Splitting Citations by Source

Table C1 leverages the same specification as Table 2, but splitting the citation count based on the type of reference. Specifically, while Panel (A.) shows the full citation count[43], Panel (B.) and (C.) limit these references to papers from Private and Non-Private papers (where Private is defined as having Percent Private $> 0.25$). This table reveals that the percentage effect of Percent Private on Citation Count is significantly higher when only considering references by Private papers. For example, considering the results in Model (2), while the percentage increase is 24.23% for all citations, it increases to 61.09% for citations from Private papers. Nevertheless, citations from Non-private papers still increase by a statistically significant amount: in this case, 11.19%. The results are consistent when considering Linear, Logged, or Poisson specifications. Overall, Table C1 shows that the citation premium is driven not only by increased attention from researchers at private firms but also by increased attention from researchers in universities. (Importantly, the citation effect is not driven solely by a citation increase from other researchers at the focal researchers' own firm.)

[43]This citation count is done over all-time as opposed to in the immediate five years after publication, an artifact of how the reference data was aggregated in SCOPUS.

| | (1) | (2) | (3) |
|---|---|---|---|
| Specification: | $Y$ (Linear) | Ln($1+Y$) | Poisson |
| *Panel A. Y = Citations (All-time)* | | | |
| Percent Private | 10.9484* | 0.2423*** | 0.3168** |
| | (6.3315) | (0.0407) | (0.1377) |
| Mean(DV) | 41.6332 | 2.3258 | 41.6332 |
| *Panel B. Y = Citations (All-time) by Private Papers* | | | |
| Percent Private | 4.1060*** | 0.6109*** | 1.106*** |
| | (0.5824) | (0.0293) | (0.1284) |
| Mean(DV) | 3.4366 | 0.6895 | 3.4366 |
| *Panel C. Y = Citations (All-time) by Non-Private Papers* | | | |
| Percent Private | 6.8430 | 0.1119** | 0.2396* |
| | (5.8600) | (0.0403) | (0.1414) |
| Mean(DV) | 38.1965 | 2.2429 | 38.1965 |
| Author-Year FE | Y | Y | Y |
| University FE | Y | Y | Y |
| Team Controls | Y | Y | Y |
| Observations | 47411 | 47411 | 43823 |
| *Notes* | | $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$ | |

Table C1. Main Effects with Citations Split by Source.

## C.2 Exploring Researcher Selection

To explore researcher selection, it's useful to explore specifications where I vary the use of my controls and fixed effects in order to understand the magnitude of researcher-selection forces in my sample. I present such results in Table C2, where each row constitutes an outcome from my prior analysis, and each column progressively adds additional controls to my specification. In Panel A, I review the primary outcome of my analysis: FWCI (logged), previously used in Table 2. Here, I see the dramatic effect of researcher selection between Model (1) and (2): by adding (interacted) author fixed effects, the association between corporate involvement and FWCI drops from 17.86% to 10.87% (a 39% drop). Panel B presents the same models but using the Q99 quantile dummy from Table 4 as an outcome. For this outcome, I see that selection does not have as large a role in the determination of the effect of corporate involvement: whereas the uncontrolled specification has an association of 1.19%, the fully controlled specification (including Academic subject and source) has an association of 1.00% (a drop of 15.9%). Together, Models (1) and (2) show that researcher selection is a first-order effect driving uncontrolled FWCI between papers with corporate involvement and university papers in my sample but not the Q99 effect. Further, after controlling for this selection, my main result is quite robust to controlling for the specific firm, the academic subject, and the publication source.

## C.3 Assortative Matching (Selection)

Consider a dual-affiliated researcher choosing whether to use a firm ($F = 1$) or university ($F = 0$) to develop a candidate scientific idea. Let ideas be identically distributed random

| | (1) Year Only | (2) Author-Year | (3) +Univ | (4) +Team | (5) +Firm | (6) +Subj | (7) +Source |
|---|---|---|---|---|---|---|---|
| *Panel A. DV = Ln(FWCI 5y+1)* | | | | | | | |
| Percent Private | 0.1786*** | 0.1087*** | 0.1571*** | 0.1253*** | 0.1097*** | 0.1091*** | 0.0567 |
| | (0.0367) | (0.0192) | (0.0226) | (0.0224) | (0.0324) | (0.0325) | (0.0365) |
| *Panel B. DV= Q99 (FWCI 5y)* | | | | | | | |
| Percent Private | 0.0119** | 0.0100*** | 0.0103*** | 0.0077** | 0.0137** | 0.0136** | 0.0147** |
| | (0.0056) | (0.0029) | (0.0036) | (0.0036) | (0.0054) | (0.0055) | (0.0069) |
| Year FE | Y | N | N | N | N | N | N |
| Author-Year FE | N | Y | Y | Y | Y | Y | Y |
| University FE | N | N | Y | Y | Y | Y | Y |
| Team Controls | N | N | N | Y | Y | Y | Y |
| Firm FE | N | N | N | N | Y | Y | Y |
| Subject FE | N | N | N | N | N | Y | Y |
| Publication FE | N | N | N | N | N | N | Y |

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table C2. Effect of Various Controls.

vectors $(S, R, C)$, where $S$ is the scientific value of the idea, $R$ is the amount that the idea benefits from firm resources, and $C \leq 1$ is the extent of firm constraints that operate on an idea. I assume either $(R + S)C$ (additive resource) or $RCS$ (multiplicative resources) is the (counterfactual) scientific value of an idea developed in the context of a firm (and therefore $C$ is the fraction of scientific value remaining after the idea is modified in order to be publishable in the context of the firm.) In this notation, the "Average Treatment Effect" of firm resources is denoted by $E[RCS] - E[S]$, and the "Average Treatment Effect on Treated" is $E[RCS \mid F = 1] - E[S \mid F = 1]$. By contrast, my empirical strategy estimates (for the case of multiplicative resources):

$$\begin{aligned} E[RCS \mid F = 1] - E[S \mid F = 0] = & \quad \textit{Estimated} \\ E[RCS \mid F = 1] - E[S \mid F = 1] & \quad ATT \\ + E[S \mid F = 1] - E[S \mid F = 0] & \quad \textit{Selection} \end{aligned}$$

In this decomposition, *Estimated* is my empirical estimand, while $ATT$ is the casual estimand of theoretical interest.

This decomposition makes clear that the sign of *Selection* depends on the correlation of the selection function $F$ and scientific value $S$. What could this selection function depend on? Three reasonable candidates include firm value (parameterized here by greater $C$), or scientific value added in the additive resource ($RC$) or multiplicative resource ($RCS$) case. In the case of $C$ and $RC$, we have no ex-ante reason to expect a correlation between these variables and $S$[44]. In the multiplicative resource case ($RCS$), we would expect these to be positively correlated.

My overall argument is that *Estimated* could be positive only if $ATT$ were positive. To argue this, I consider three cases:

[44]For all we know, ideas that benefit more from resources may be *worse* scientific ideas in the university context, e.g. if they were simply incremental conceptual advances that could not be shown to be worth consideration apart from considerable additional resources.

1. *Selection* $< 0$, $ATT > 0$. This occurs when $C$ or $RC$ are the key variables in the selection function, and they are negatively correlated with $S$. In this case, my empirical estimate is an underestimation of the resource effect.

2. *Selection* $> 0$, $ATT > 0$. This happens if the selection function is based on scientific value added (e.g. $RCS$), and resource-benefit $RC$ is sufficiently correlated with $S$. In this case, the only reason a dual-affiliated researcher works on better ideas in a firm (*Selection* $> 0$) is because better ideas benefit more from firm resources ($ATT > 0$).

3. *Selection* $> 0$, $ATT < 0$. For this to be true, the selection function would have to be positively correlated with $S$ but negatively correlated with $RC$. In other words, dual-affiliated would have to prefer working on *better* ideas in the context of a firm even though the firm made those ideas *worse* off. This could theoretically occur if firms paid scientists for the quality of ideas (e.g. amount of citations) that they brought to the firm. However, this does not match how such contracts are written in practice (indeed, citations take too long to materialize to serve as good metrics for contracts) and contradicts a considerable literature emphasizing that scientists seek scientific credit (e.g. Stern (2004); Foray and Lissoni (2010)). I argue that there is, therefore, no reasonable selection function that could lead to this case.

In summary, I have developed a simple selection model to more carefully consider whether idea-level selection by dual-affiliated researchers can explain away the corporate citation premium. I argued that *no*, given *Estimated* has been empirically shown to be positive, $ATT$ must also be positive. In the case where the selection effect is negative, *Estimated* underestimates the true magnitude of $ATT$). In the case where the selection effect is positive, I argue that the only reasonable way this occurs is if better scientific ideas benefit more from firm resources. But in that case, we are essentially assuming $ATT > 0$ in order to justify the positive selection!

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| DV: | Ln(1+ FWCI 5 Yr) | Q99 (FWCI 5y) | CompSci | Medicine |
| *Panel A. At Least One Prior Collaborator* | | | | |
| Percent Private | 0.1075*** | 0.0093** | 0.0637*** | −0.0255*** |
| | (0.0277) | (0.0043) | (0.0098) | (0.0050) |
| Observations | 39495 | 39495 | 65854 | 65854 |
| *Panel B. At Least Half Prior Collaborators* | | | | |
| Percent Private | 0.0798** | 0.0042 | 0.0519*** | −0.0174*** |
| | (0.0326) | (0.0049) | (0.0127) | (0.0058) |
| Observations | 30787 | 30787 | 49475 | 49475 |
| *Panel C. Only Prior Collaborators* | | | | |
| Percent Private | 0.0596 | 0.0070 | 0.0322** | −0.0078 |
| | (0.0393) | (0.0062) | (0.0157) | (0.0061) |
| Observations | 24683 | 24683 | 37910 | 37910 |
| Mean(DV) | 0.9137 | 0.0126 | 0.7538 | 0.0484 |

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table C3. Sample Restricted to Papers with Prior Collaborations Only.

## C.4 Alternative Causal Mechanisms

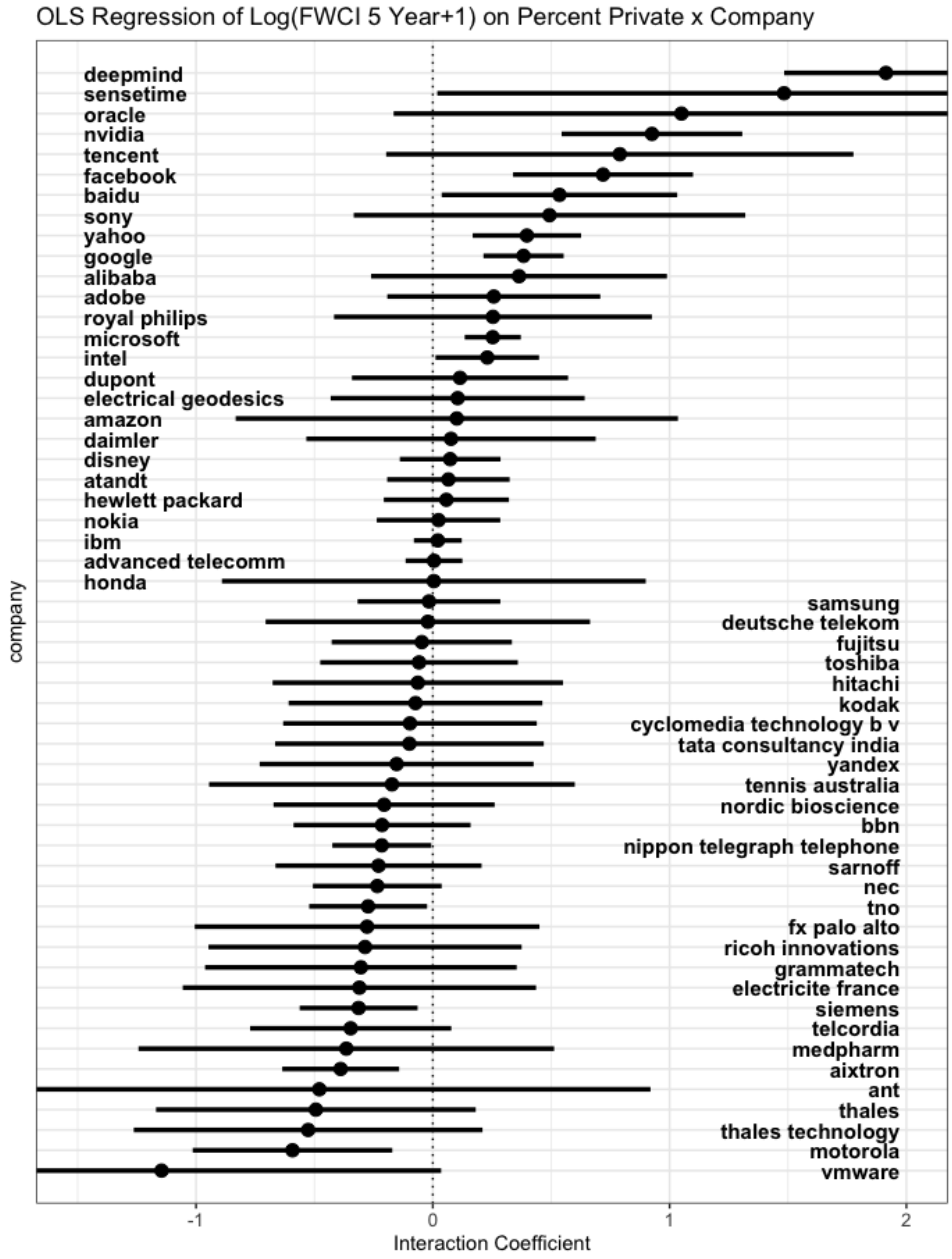


Figure C1. Estimates of Effect of Specific Firm Involvement on FWCI.

## C.5 Robustness of Estimates to Measurement and Sample Choice

| DV: | (1) Ln(1+ FWCI 5 Yr) | (2) Q99 (FWCI 5y) | (3) CompSci | (4) Medicine |
|---|---|---|---|---|
| *Panel A. Percent Private (Original Measure)* | | | | |
| Percent Private | 0.1253*** | 0.0077** | 0.0679*** | −0.0256*** |
| | (0.0224) | (0.0036) | (0.0081) | (0.0042) |
| *Panel B. Binary Measure (0.5 Cutoff)* | | | | |
| Percent Private > 0.5 | 0.0673*** | 0.0053*** | 0.0369*** | −0.0119*** |
| | (0.0133) | (0.0021) | (0.0049) | (0.0024) |
| *Panel C. Binary Measure (0.33 Cutoff)* | | | | |
| Percent Private > 0.33 | 0.0529*** | 0.0037** | 0.0377*** | −0.0110*** |
| | (0.0121) | (0.0018) | (0.0045) | (0.0022) |
| *Panel D. Binary Measure (0.25 Cutoff)* | | | | |
| Percent Private > 0.25 | 0.0482*** | 0.0031** | 0.0369*** | −0.0109*** |
| | (0.0112) | (0.0016) | (0.0042) | (0.0021) |
| *Panel E. Binary Measure (0.0 Cutoff)* | | | | |
| Percent Private > 0.0 | 0.0555*** | 0.0018 | 0.0304*** | −0.0154*** |
| | (0.0107) | (0.0016) | (0.0041) | (0.0021) |
| *Panel F. Dual-Author's Affiliation Only* | | | | |
| Percent Private (Dual Author) | 0.1052*** | 0.0040 | 0.0376*** | −0.0175*** |
| | (0.0181) | (0.0028) | (0.0071) | (0.0035) |
| Mean(DV) | 0.8897 | 0.0121 | 0.7442 | 0.0482 |
| Observations | 47411 | 47411 | 77847 | 77847 |

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table C4. Robustness Tests for Corporate Involvement Measure.

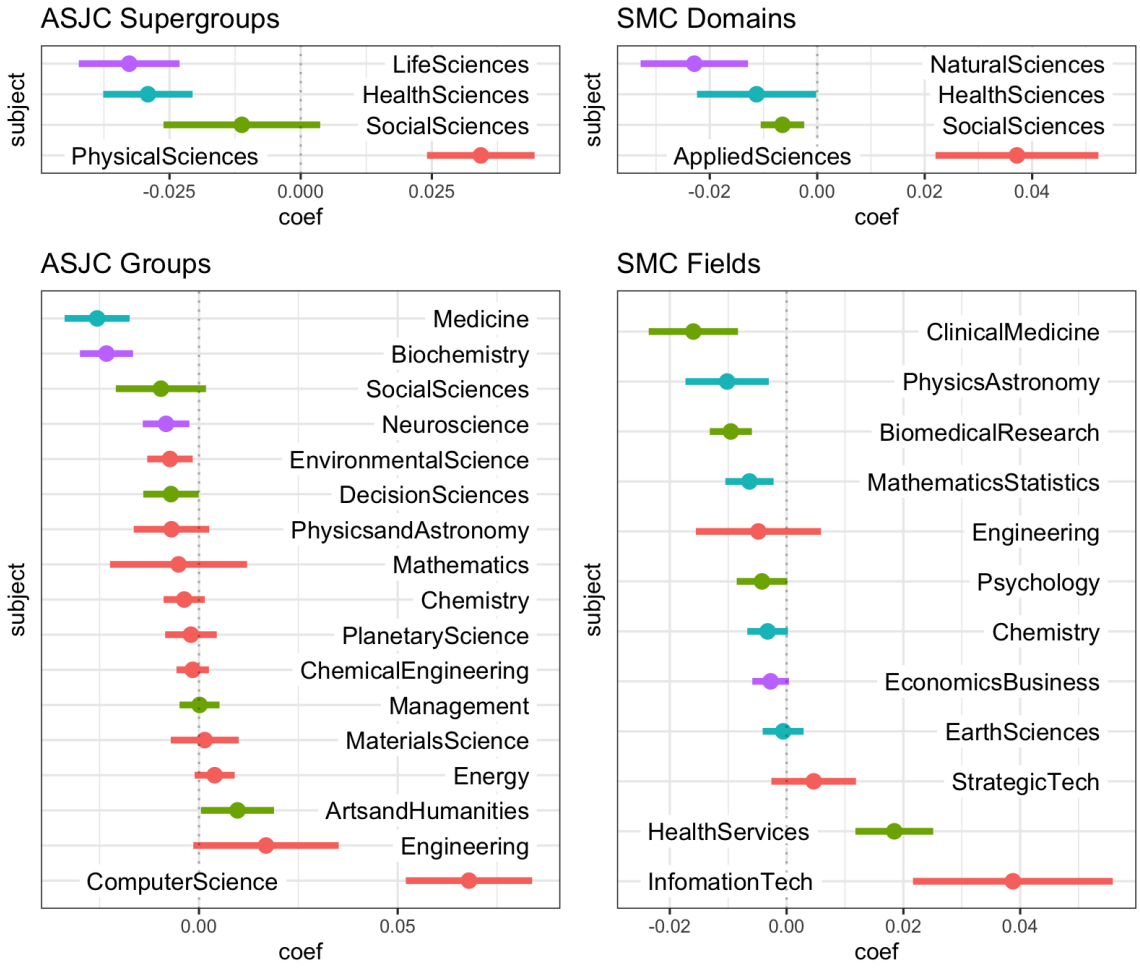


Figure C2. Robustness Tests for Academic Subject Measure.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| DV: | Ln(1+ FWCI 5 Yr) | Q99 (FWCI 5y) | CompSci | Medicine |
| *Panel A. After 2000* | | | | |
| Percent Private | 0.1436*** | 0.0129*** | 0.0296** | −0.0214*** |
| | (0.0257) | (0.0042) | (0.0125) | (0.0061) |
| Observations | 42145 | 42145 | 42145 | 42145 |
| *Panel B. After 2010* | | | | |
| Percent Private | 0.1933*** | 0.0215*** | 0.0412*** | −0.0161** |
| | (0.0347) | (0.0062) | (0.0159) | (0.0078) |
| Observations | 24663 | 24663 | 24663 | 24663 |
| *Panel C. Remove Less Productive Dual Author* | | | | |
| Percent Private | 0.1147*** | 0.0109** | 0.0650*** | −0.0264*** |
| | (0.0266) | (0.0043) | (0.0095) | (0.0048) |
| Observations | 38894 | 38894 | 62036 | 62036 |
| *Panel D. Remove More Productive Dual Author* | | | | |
| Percent Private | 0.1087*** | 0.0081** | 0.0543*** | −0.0260*** |
| | (0.0265) | (0.0041) | (0.0103) | (0.0052) |
| Observations | 38894 | 38894 | 62036 | 62036 |
| Mean(DV) | 0.8854 | 0.0124 | 0.6779 | 0.0550 |

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table C5. Robustness Tests for Alternative Sample Criteria.