Supplemental Appendix:
Cassatts in the Attic: Is there a gender gap in the commercialization of science?

Marlène Koffi and Matt Marx

Table of Contents

This Internet Appendix contains supplementary discussions and analyses, which we organize as follows:

A Predict commercial potential

A.1 The use of text data

Articles abstracts are highly appropriate for the incorporation of text data into the predictive algorithm. In fact, they offer a succinct overview of the main outcomes, methodology, conclusions, and significant contributions of the study, along with its distinguishing characteristics that may indicate its potential quality. As a result, abstracts are less prone to digressions and exhibit a more focused and structured format. This feature proves valuable in minimizing noise within the prediction task. Thus, on top of the computational constraints, this explains the extensive utilization of abstracts when analyzing academic research papers.

We use Bidirectional Encoder Representations from Transformers to process the texts and compute their vector representation (BERT, Devlin et al. 2018). BERT utilizes a bidirectional language model, allowing it to consider both each word's left and right context. To achieve high accuracy, BERT undergoes pretraining on a massive corpus consisting of billions of words (BooksCorpus and English Wikipedia). By leveraging the contextual associations from those sources, BERT can generate 768-dimensional vector representations for words within a text block. These representations consider the surrounding words to capture the contextual meaning of each word. BERT is designed to accomplish two essential tasks. First, it learns to predict masked words within a sentence, with approximately 15% of the words being masked. The model then predicts the masked words based on the context provided by the surrounding words. This masked language modeling task aims to minimize the cross-entropy loss between the predicted probabilities and the actual masked tokens. Second, BERT is trained to understand connections between sentences through a task called next-sentence prediction. Pairs of sentences are used, and the model is trained to classify whether the second sentence follows the first sentence (labeled as"IsNex") or if it is a randomly chosen sentence (labeled as "NotNext").

The main objective of the BERT model is to minimize the combined loss function, which consists of the cross-entropy loss from the masked token task and the binary loss from the next sentence prediction task. To further enhance sentence-level understanding, an extension of BERT called Sentence-BERT (SBERT, Reimers and Gurevych 2019) is utilized. SBERT focuses on generating fixed-length vector representations (embeddings) specifically for sentences or short texts. Unlike BERT, which primarily focuses on word-level tasks, SBERT aims to capture entire sentences' meaning and semantic similarity. SBERT has demonstrated improved speed compared to BERT while maintaining the accuracy achieved by BERT (Reimers and Gurevych 2019). Table G.1 shows the value of the tuning parameters.

A.2 Training and model accuracy

We train and fine-tune our models for 65 distinct groups, employing stochastic gradient descent learning models and tree-based methods. Stochastic gradient descent learning models are a set of algorithms using stochastic gradient descent to improve the speed of the algorithm. They are particularly useful for huge data. The 65 groups represent Web of Science (WoS) field aggregations designed to enable the algorithms to consider field-specific features. After identifying the optimal model for each category, we observe that the average f1-score across all models is around 80% both for the model with and without data, while the average area under the curve (AUC) for both models is above 90%.

Furthermore, we conducted additional evaluations considering twin pairs, where one twin is commercialized and the other is not. For approximately 75% of the uncommercialized twins, our predictive algorithms indicate a probability of commercialization over 50%, i.e., the uncommercialized twin could have as well gotten commercialized. Once again, this ensures the accuracy of our metrics.

B Robustness checks for PPP definition

Our primary indicator for whether a scientific discovery is whether the paper can be "paired" with a patent as described in section 2.2. This definition may undercount the true rate of commercialization if, for example, the paired patent in the PPP does not belong to a commercial firm but to a university,

and the university subsequently licenses the patented discovery to a firm. The license would generally be unobservable to us, leading commercialized discoveries to be classified as uncommercialization.

If, for example, female PIs were more likely to pursue this path, our estimates may reflect that choice. On the one hand, this might seem a reasonable possibility given prior findings that women scientists are less likely to engage in commercial activities (Tartari and Salter 2015). On the other, the fact that we fail to find a gap in the self-commercialization process (see section 4.1) argues against the notion that women would tend to be "hands off" in the commercialization process.

Again, licensing is not fully visible to us, but as reported in Section 2.2.4, we attempt to include in our dependent variable possible licensing activities. In this expanded definition, we do not require that the paper be cited by a patent belonging to a firm, which can include a PPP assigned to a university. This expanded set of commercialization instances—which are added to our set of PPPs—is assembled in three ways. We begin with the set of patents paired to papers where the assignee is *not* a firm, which would not have been counted as a commercialization PPP in our main analysis. We do not count these non-firm PPPs as commercialization as it is unclear whether commercial activity took place. The university, government agency, or other assignee may have obtained patent protection but not taken the discovery further in the commercialization process.

We first check whether the paired non-firm patent appears in OrangeBook (Durvasula et al. 2023), a linkage between small-molecule drugs approved by the FDA and the patents that protect them. This will primarily confirm commercialization instances in biopharmaceuticals. If the patent is in OrangeBook, we take this as evidence that the paired paper was commercialized. (Note: this does not entail that every paper cited by the OrangeBook patent is commercialized, only the paired paper.)

Second, we undertake a similar exercise using the (Rassenfosse and Higham 2020) database of Virtual Patent Markings (VPM)s. VPMs appear primarily for physical consumer goods. Third, we do the same for RoyaltySource, an inventory of licensed patents. RoyaltySource 2021 is primarily constructed using 10-K filings and other public sources, so it is likely limited to publicly traded firms. Therefore, this expanded commercial measure, although more inclusive, is not claimed to be exhaustive.

C Twins human audit: protocol

We use the following grid to evaluate the twin papers;

- **3 - Identical Twins:** Both papers address an identical research question or topic and, subsequently, arrive at congruent conclusions. This suggests that they are functionally analogous and exhibit a mutual substituability.
- **2 - Partial Twins:** The papers discuss the same overarching research question or topic and present conclusions that, while being largely aligned, possess nuanced differences. The two articles largely overlapped.
- **1 - Related yet Distinct:** While both papers pertain to a similar broad theme, they delineate distinct facets or pose varied sub-questions within that theme. Their substitutability nature is not direct, as they provide disparate pieces of information under the shared thematic umbrella. They look more complementary.
- **0 - Divergence or Irrelevance:** The papers either draw antithetical conclusions from a shared research question or topic, or they address wholly unrelated subjects. Furthermore, any paper identified as a literature review or one that offers a generalized overview without specific conclusions shall be categorized as this score.

D Demand side: commercialization and gender homophily

We present evidence of gender homophily in the commercialization process. In an ideal experiment, we would randomly seed commercialization partners with heterogenous gender composition and assess

the likelihood of commercializing scientific articles of otherwise identical quality but heterogeneous composition of the scientific teams. The analytic approach used so far (i.e., Equation 2) cannot accomplish this because the setup is at the level of the academic paper and thus cannot compare the gender composition of the paper with that of the patent.

Thus, we switch the level of analysis from the paper to a patent-paper dyad that *potentially* forms a patent-paper pair. As before, we account for the quality and nature of the paper with "twin" discoveries. To approximate the random seeding of patents that could form a patent-paper pair, we adopt a case-control setup. We reduce our set of twin papers to those where one or the other indeed formed a pair with some patent. A dyad is formed both for the patent and the paper with which it *is* paired as well as for the patent and the twin of the (actually paired) paper, with which the paper was *not* paired but should be about as likely to have been paired. This unrealized patent-paper pairing forms a counterfactual for our case-control analysis.

For the patent, we calculate the percentage of male vs. female inventors on the patent using USPTO (U.S. Patent and Trademark Office 2023) inventor-level classification (therefore, this step is limited to USPTO-issued patents only). To avoid biasing this measure in favor of the paper that was actually cited, in calculating the gender composition of the patent, we omit the inventor(s) who were matched to authors on the paper in the (realized) patent-paper pair. We then estimate the following equation, which deviates from Equation 2 by including a) the gender composition of patent j b) fixed effects for patent j.

$$\begin{aligned} COMM_{ijt} = \alpha_0 + \alpha_1 PaperGender_i + \alpha_2 PatentGender_j + \\ \alpha_3 PaperGender_i X PatentGender_j + \\ \beta X_{it} + TwinDiscoveryFE + \epsilon_{ijt} \end{aligned} \tag{1}$$

We then test the presence of gender homophily in team formation for the commercialization process. To do this, we shift focus from individual papers to potential patent-paper dyads. We continue to use "twin" papers to control for quality and adopt a case-control setup, narrowing our focus to twin papers where at least one formed a pair with a patent. We create dyads for both the actual patent-paper pairs and their corresponding (unpaired) twins, the latter serving as a counterfactual for comparison as it should be about as likely to have been paired. This method allows us to approximate the random pairing of patents and papers in our analysis. For patents, we compute the gender ratio of inventors using USPTO inventor-level classification. To ensure an unbiased measure of a patent's gender composition, we exclude inventors from this calculation if they are matched to authors in the realized patent-paper pairs. In Table G.18, Column (1) confirms our primary finding using a female last author dummy. Column (2) explores the interaction between female last authorship and the percentage of male inventors on the citing patent, revealing a negative, significant coefficient. This suggests that articles with female authors are less likely to be commercialized by patents, with a higher percentage of male inventors. Similar trends are observed in Columns (3) and (4), accounting for the percentage of non-last female authors.

E Demand side: accessibility of articles

We explore the influence of increased visibility of scientific discoveries through a natural experiment facilitated by open access mandates for federally funded research. Post-2008, there was a staggered introduction of a public access policy, mandating that publicly funded academic research be made freely accessible. Our findings indicate that the implementation of open access mandates has, paradoxically, widened the gender gap in the commercialization of scientific projects led by women. This outcome implies that enhancing the accessibility of research articles disproportionately benefits the commercialization prospects of male authors over those of female authors, highlighting a significant discrepancy in how improved article visibility impacts the commercialization potential across genders.

We discuss the impact of awareness about one's scientific discovery as a potential demand-side factor. Collaboration between a researcher and a firm can occur in different ways: published papers and reports, public conferences and meetings, informal information exchange, and consulting, geographic hubs (Cohen, Nelson, and Walsh 2002, Markman, Siegel, and Wright 2008, Bikard and Marx 2019).

While it is challenging to quantify the relative share of each method that can lead to commercialization, it is widely acknowledged that access to scientific publications promotes scientific collaboration (Gowers and Nielsen 2009, Friesike et al. 2015, McKiernan et al. 2016). Thus, it seems plausible that knowledge about one research could lead to a collaboration with a firm.

Therefore, the gender gap in commercialization could also be salient in an environment where access to information on scientific articles is not perfectly distributed and gender-specific.[1] In particular, if scientific articles with women are less visible than male-authored scientific articles, this could prevent companies from accessing the former's publications and, therefore, reduce their probability of commercializing relative to the latter. If such a hypothesis turns out to be accurate, then a shock that would increase awareness and access to scientific research should contribute to reducing the gender gap in commercialization.

To test this, we use a natural experiment provided by the open-access mandates for federally-funded research. In many scientific fields, most articles and working papers are not freely available (Bjork, Roos, and Lauri 2009, Khabsa and Giles 2014, Ware and Mabe 2015). At the same time, one of the most common rationales behind the evolution of scientific discovery is to expand the frontier of knowledge by building upon previously available research. In fact, many authors have shown that limited awareness (limited access or openness constraints) about scientific production can limit the use of science (Furman and Stern 2011,Staudt 2020, Bryan and Ozcan 2021). This channel could be more important in the commercialization of academic research as firms may need to explicitly collaborate with a researcher from an academic institution.

In 2008, the National Institutes of Health (NIH) leveraged an initiative to make freely available the academic research they funded such that any article accepted for publication after April 7, 2008, must be archived in the open-access PubMed Central (PMC) database within 12 months of publication.[2] In 2013, the White House Office of Science Technology Policy mandated agencies with an R&D budget of $100M in order to develop plans to make the results of the federally funded research freely available. This gave rise to a staggered implementation of the "Public Access policy" (PAP), with, for example, the Department of Energy (DOE) implementing this policy in 2014 and the National Science Foundation (NSF) in 2016.

Our empirical model takes advantage of the gradual implementation of the PAP by constructing an event study where the event date is the starting year of the PAP for one agency. Therefore, an article in the database is considered to be "treated" in a given year if a federal agency financed this paper, and during that year, this agency started to implement the PAP. In this setup, we are particularly interested in the triple difference that captures the effect on the commercialization of federally-funded publications written by women after the implementation of the PAP relative to those written by men. Therefore, assessing the effect of the PAP on narrowing the gender gap in commercialization. We further add the control variables similar to our baseline and include the journal, year, and field fixed effects. Our identifying assumption is that there are no shocks correlated with the introduction of the PAP that differentially affect scientific teams with men/women commercialization likelihood. To address concerns regarding heterogeneous treatment effects, we use a robust staggered difference and difference approach by Sun and Abraham 2021. Other procedures to solve this issue have been proposed by Goodman-Bacon 2021, Callaway and Sant'Anna 2021. (Baker, Larcker, and Wang 2022 shows an interesting equivalence between those different procedures.)

Figure H.1 shows the result of this estimation. We use one year before the introduction of the PAP as the reference year. Figure H.1 presents the difference-in-difference estimate separately for papers with a woman as the last author (i.e., lab manager) and papers with a man as the last author. There is no statistically significant pre-trend. In fact, although the pre-trend for each gender group is not non-significant, impeding the interpretation of the simple difference, we clearly see that both genders are moving in an almost perfect one-to-one mapping before PAP. There seems to be a sharp jump in the commercialization of science following the advent of Open Access mandates, but we do not see a

1. Indeed, dissemination of academic research via social media, for example, has been shown to increase the visibility and the likelihood of citation (Eysenbach 2011 and Klar et al. 2020).

2. Most of the paper in the literature of open access on academic citation finds a non-negative effect. In particular, Bryan and Ozcan 2021 show that after 2008, NIH-funded researches were 12 to 27% more likely to get cited, while Staudt 2020 finds a positive but more moderate effect.

material convergence of the gender gap. Rather, the gap widens starting in year 1, diverging further in years 2-4.

We conclude that, contrary to prior expectations that increased information might help close the gender gap, the introduction of open access mandates has, in fact, exacerbated the gender gap in commercialization for scientific projects led by women.

F Becker's outcome test for bias on the part of potential commercialization cooperative partner firms

We attempt to build an instrument for identifying marginal PPPs. The intuition behind the instrument is that PPPs with assignees who rarely commercialize scientific discoveries are probably stronger. In contrast, if the PPP's assignee tends to commercialize more frequently, the discovery may be marginal.

Our measure of how frequently an assignee commercializes scientific discoveries is computed as follows. Although we could count the number of commercializations for each assignee each year, we want to add a denominator to this count to represent the fraction of science "known to the assignee" that is commercialized. Of course, we cannot detect whether any employee of the assignee has read a focal paper. We proxy for awareness by whether the focal assignee cited a paper, which reflects that someone working at the assignee not only knew about the paper but considered it relevant enough to the firm's innovation to cite it in a patent.

We start with all the papers and collect the patents that *cited* those papers, as well as the application year and assignee. The data are then collapsed to the paper-year-assignee level, representing all assignees that *cited* a paper in a given year—whether or not that citation was part of a PPP. This is used to build a count of the number of PPP papers *cited* by every assignee each year and then calculate the percentage of those paper-to-patent citations that were actually part of a PPP.

The final step is to create an observation for each paper-assignee-year (where the paper was in a PPP) that contains two variables: 1) whether the focal paper was commercialized by that assignee 2) the average commercialization percentage for that assignee in that year. Note that for the commercialization percentage, we subtract from the numerator the focal paper if commercialized by that assignee and from the denominator all papers by the PI.

As shown in columns (1) and (2), for either male or female PIs, PPPs that are cited by assignees with a higher likelihood of commercializing papers by the *other* PIs they cite are considerably more likely to be commercialized. For outcome variables, one might consider forward citations a measure of patent value. However, Hochberg et al. 2023 show that patent citations are gender-biased. Therefore, we instead employ a dependent variable of the financial value of a patent as calculated by Kogan et al. 2017 based on stock market reactions to its issuance. Following Huang, Mayer, and Miller 2022, columns (3) and (4) of Table G.19 estimate the second stage, with an absolute but not statistically significant difference in means.

However, this analysis is tentative because it contains numerous reservations. For example, selecting a sample of only publicly traded companies could threaten the exogeneity condition. Additionally, missing values for KPSS cannot be replaced with zero (or another random value) as in Huang, Mayer, and Miller 2022 and Benson, Li, and Shue 2019 because a missing KPSS value does not mean that the patent is not valuable.

G Additional tables

Table G.1: Tuning parameters

Field	Tuning without Language model	Tuning with Language model
Acoustics	(0.01, 'hinge', 'l2')	(0.0007, 'hinge', 'l1')
Agricultural Engineering	(0.007, 'hinge', 'l1')	(0.001, 'log_loss', 'l1')
Allergy	(0.03, 'hinge', 'l2')	(0.0005, 'log_loss', 'l1')
Andrology	(0.01, 'hinge', 'l1')	(0.01, 'log_loss', 'l1')
Biochemical Research Methods	(0.007, 'hinge', 'l1')	(0.0005, 'hinge', 'l1')
Biochemistry, Molecular Biology	(0.005, 'hinge', 'l2')	(0.03, 'hinge', 'l2')
Biotechnology, Applied Microbiology	(0.005, 'hinge', 'l1')	(0.001, 'log_loss', 'l1')
Cardiac, Cardiovascular Systems	(0.007, 'hinge', 'l2')	(0.0003, 'hinge', 'l1')
Cell, Tissue Engineering	(0.03, 'hinge', 'l1')	(0.09, 'hinge', 'l1')
Cell Biology	(0.001, 'hinge', 'l1')	(0.0007, 'log_loss', 'l1')
Chemistry, Analytical	(0.005, 'hinge', 'l2')	(0.0003, 'log_loss', 'l1')
Chemistry, Applied	(0.0005, 'log_loss', 'l1')	(0.0003, 'hinge', 'l1')
Chemistry, Inorganic, Nuclear	(0.001, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Chemistry, Medicinal	(0.07, 'hinge', 'l2')	(0.0005, 'log_loss', 'l1')
Chemistry, Multidisciplinary	(0.01, 'hinge', 'l2')	(0.0007, 'log_loss', 'l1')
Chemistry, Organic	(0.01, 'hinge', 'l2')	(0.0003, 'log_loss', 'l1')
Chemistry, Physical	(0.005, 'hinge', 'l2')	(0.0001, 'hinge', 'l1')
Computer Science, Artificial Intelligence	(0.03, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Computer Science, Cybernetics	(0.0005, 'log_loss', 'l1')	(0.005, 'log_loss', 'l1')
Computer Science, Hardware, Architecture	(0.007, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Computer Science, Information Systems	(0.0007, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Computer Science, Software Engineering	(0.007, 'hinge', 'l1')	(0.0001, 'log_loss', 'l1')
Computer Science, Theory, Methods	(0.005, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Developmental Biology	(0.001, 'hinge', 'l1')	(0.0005, 'log_loss', 'l1')
Electrochemistry	(0.05, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Endocrinology, Metabolism	(0.007, 'hinge', 'l2')	(0.0001, 'hinge', 'l1')
Energy, Fuels	(0.0005, 'hinge', 'l1')	(0.0005, 'log_loss', 'l1')
Engineering, Biomedical	(0.07, 'hinge', 'l1')	(0.0007, 'log_loss', 'l1')
Engineering, Chemical	(0.005, 'hinge', 'l1')	(0.0007, 'log_loss', 'l1')
Engineering, Electrical, Electronic	(0.001, 'hinge', 'l1')	(0.03, 'hinge', 'l2')
Engineering, Manufacturing	(0.0005, 'hinge', 'l1')	(0.0001, 'hinge', 'l1')
Genetics, Heredity	(0.01, 'hinge', 'l2')	(0.0007, 'log_loss', 'l1')
Hematology	(0.007, 'hinge', 'l2')	(0.0007, 'log_loss', 'l1')
Imaging Science, Photographic Technology	(0.005, 'log_loss', 'l1')	(0.05, 'hinge', 'l1')
Immunology	(0.005, 'hinge', 'l1')	(0.07, 'hinge', 'l1')
Instruments, Instrumentation	(0.005, 'hinge', 'l1')	(0.0003, 'log_loss', 'l1')
Limnology	(0.0001, 'hinge', 'l2')	(0.7, 'hinge', 'l2')
Materials Science, Ceramics	(0.0007, 'hinge', 'l1')	(0.001, 'hinge', 'l1')
Materials Science, Characterization, Testing	(0.001, 'log_loss', 'l1')	(0.05, 'hinge', 'l1')
Materials Science, Coatings, Films	(0.005, 'log_loss', 'l1')	(0.007, 'hinge', 'l2')
Materials Science, Multidisciplinary	(0.03, 'hinge', 'l2')	(0.0003, 'hinge', 'l1')
Materials Science, Paper, Wood	(0.005, 'log_loss', 'l1')	(0.3, 'log_loss', 'l2')
Medicine, Research, Experimental	(0.005, 'hinge', 'l1')	(0.5, 'log_loss', 'l2')
Microbiology	(0.03, 'hinge', 'l2')	(0.0005, 'log_loss', 'l1')
Multidisciplinary Sciences	(0.05, 'hinge', 'l1')	(0.001, 'hinge', 'l1')
Nanoscience, Nanotechnology	(0.0007, 'hinge', 'l1')	(0.03, 'hinge', 'l1')
Neurosciences	(0.01, 'hinge', 'l2')	(0.0001, 'hinge', 'l1')
Oncology	(0.01, 'hinge', 'l2')	(0.0001, 'hinge', 'l1')
Ophthalmology	(0.0007, 'hinge', 'l1')	(0.0005, 'log_loss', 'l1')
Optics	(0.005, 'hinge', 'l1')	(0.0003, 'log_loss', 'l1')
Parasitology	(0.0005, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
Peripheral Vascular Disease	(0.0001, 'hinge', 'l1')	(0.0007, 'hinge', 'l1')
Pharmacology, Pharmacy	(0.05, 'hinge', 'l1')	(0.0003, 'log_loss', 'l1')
Physics, Applied	(0.07, 'hinge', 'l1')	(0.0003, 'log_loss', 'l1')
Physics, Mathematical	(0.001, 'hinge', 'l1')	(0.007, 'hinge', 'l1')
Plant Sciences	(0.005, 'hinge', 'l2')	(0.0003, 'hinge', 'l1')
Polymer Science	(0.01, 'hinge', 'l1')	(0.001, 'hinge', 'l1')
Radiology, Nuclear Medicine, Medical Imaging	(0.005, 'hinge', 'l1')	(0.0005, 'log_loss', 'l1')
Reproductive Biology	(0.0003, 'log_loss', 'l1')	(0.005, 'log_loss', 'l1')
Rheumatology	(0.0007, 'log_loss', 'l1')	(0.0001, 'hinge', 'l1')
Robotics	(0.0007, 'log_loss', 'l1')	(0.001, 'log_loss', 'l1')
Spectroscopy	(0.007, 'hinge', 'l1')	(0.0003, 'log_loss', 'l1')
Toxicology	(0.005, 'log_loss', 'l1')	(0.005, 'log_loss', 'l1')
Virology	(0.01, 'hinge', 'l1')	(0.09, 'hinge', 'l1')
All the others	(0.0001, 'hinge', 'l1')	(0.0001, 'log_loss', 'l1')

Notes: Table shows the tuning parameters for the machine learning models for the different field categories used for the training and the tuning of the model. The models were built to optimize the f1-score.

Table G.2: Cross-sectional commercialization vs. citation: predicting commercializability with language model

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author	-0.0306	-0.0227	-0.0133	-0.0098
	(0.0010)	(0.0011)	(0.0008)	(0.0008)
Pct. female authors (not last)		-0.0490		-0.0214
		(0.0014)		(0.0011)
Prob. comm. (w/LM)			0.8662	0.8656
			(0.0013)	(0.0013)
Observations	1011036	1011036	1011036	1011036
Mean of DV	0.2358	0.2358	0.2358	0.2358
Field-year fixed effects	y	y	y	y
Journal fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization on the cross-sectional data of MAG academic research papers, controlling for the commercial potential. The commercial potential was obtained by predictive algorithms using language model described in Appendix A. The prediction is based on a structured database where all the available information on the paper is used and text-based data. As described in Section 3.2, this is performed on a 1.5% subsample of the MAG data. For the gender variable, the reference category is male last author. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. In all the specifications, the standard errors are robustly estimated.

Table G.3: Balance check at the twin level (Part 2)

	Ln average citations per institution			
	(1)	(2)	(3)	(4)
Female last author	-0.0237	-0.0302	-0.0158	-0.0158
	(0.0591)	(0.0805)	(0.0892)	(0.0839)
Observations	27436	27436	6276	6276
Mean of DV	3.2992	3.2992	3.3625	3.3625
Twin-paper fixed effects	n	y	n	y

	Ln average citations per author			
	(1)	(2)	(3)	(4)
Female last author	-0.2766	-0.2638	-0.2760	-0.2760
	(0.0306)	(0.0349)	(0.0450)	(0.0355)
Observations	27436	27436	6276	6276
Mean of DV	5.8632	5.8632	5.8907	5.8907
Twin-paper fixed effects	n	y	n	y

	% female authors (not last)			
	(1)	(2)	(3)	(4)
Female last author	0.1918	0.1637	0.1584	0.1584
	(0.0093)	(0.0128)	(0.0135)	(0.0133)
Observations	27436	27436	6276	6276
Mean of DV	0.5138	0.5138	0.4706	0.4706
Twin-paper fixed effects	n	y	n	y

	Ln authors			
	(1)	(2)	(3)	(4)
Female last author	0.0043	-0.0216	-0.0247	-0.0247
	(0.0090)	(0.0102)	(0.0136)	(0.0106)
Observations	27436	27436	6276	6276
Mean of DV	1.8235	1.8235	1.8397	1.8397
Twin-paper fixed effects	n	y	n	y

Notes: This table presents a balance check for the twin sample, segmented by gender. For each variable, we estimate the coefficient of the linear model where the independent variable of interest is the gender of the last author. In each table, columns (1) and (2) present estimates based on the full twin sample, while columns (3) and (4) focus on a subsample of the twin sample comprising pairs where one principal investigator (PI) is female and the other is male. In all the specifications, the standard errors are robustly estimated.

Table G.4: Commercialization gap by increasing importance (Twins sample)

Panel A: Commercialization gap by increasing academic citation (Twins)

	Commercialization			
	Below-median citations		*Above-median citations*	
	(1)	(2)	(3)	(4)
Female last author	0.000738	0.000898	-0.0233	-0.0230
	(0.00361)	(0.00366)	(0.00997)	(0.0101)
Pct. female authors (not last)		-0.00109		-0.00209
		(0.00167)		(0.00434)
Observations	11612	11612	11279	11279
Mean of DV	0.0121	0.0121	0.0901	0.0901
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Panel B: Commercialization gap by increasing patent citation (twins)

	Commercialization			
	Below-median citations		*Above-median citations*	
	(1)	(2)	(3)	(4)
Female last author	0.00164	0.00172	-0.0377	-0.0369
	(0.00102)	(0.00105)	(0.0143)	(0.0144)
Pct. female authors (not last)		-0.000551		-0.00674
		(0.000411)		(0.00715)
Observations	14060	14060	8137	8137
Mean of DV	0.00135	0.00135	0.135	0.135
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Notes: This table analyzes the evolution of the gender gap in commercialization as the "quality" of the scientific discovery increases in the twin sample. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. For the gender variable, the reference category is male last author. Each column of panels A and B estimates a subset of the sample based on citation-count above or below the median. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. In all the specifications, the standard errors are robustly estimated.

Table G.5: Alternative dependent variables

Panel A: All papers

	No social sciences	Remove firm affiliation	PI on paired patent	Remove Transitive PPP
	(1)	(2)	(3)	(4)
Female last author	-0.0006	-0.0004	-0.0006	-0.0005
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Pct. female authors (not last)	-0.0020	-0.0014	-0.0006	-0.0019
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Observations	61064733	67382643	69614186	69614186
Mean of DV	0.0041	0.0029	0.0021	0.0034
Field-year fixed effects	y	y	y	y
Journal fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Panel B: Twins sample

	No social sciences	Remove firm affiliation	PI on paired patent	Remove Transitive PPP
	(1)	(2)	(3)	(4)
Female last author	-0.0121	-0.0110	-0.0084	-0.0098
	(0.0044)	(0.0044)	(0.0036)	(0.0042)
Pct. female authors (not last)	-0.0027	-0.0026	0.0001	-0.0022
	(0.0018)	(0.0018)	(0.0016)	(0.0018)
Observations	27020	25471	27398	27398
Mean of DV	0.0464	0.0410	0.0284	0.0422
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization on the cross-sectional data of MAG academic research papers and in the twins sample. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. Additionally, the twin regression presented in Panel B incorporates controls for the twin paper that was published first. Column (1) excludes social sciences. In Column (2), we remove the case where one of the authors on the paper has listed a firm as an affiliation. In Column (3), we consider the case where the principal investigator (last author) is also on the patent. In Column (4), we remove "transitive" PPPs defined in Section 2.2.2 from the dependent variable. For the gender variable, the reference category is male last author. The number of observations is lower in columns (1) and (2) due to the omission of social science papers and firm-affiliated papers, respectively. In all the specifications, the standard errors are robustly estimated.

Table G.6: University patent-paper pairs and patent citations to papers

Panel A: All papers

			Ln citations from patents		
	Only UPPPs	Paper cited by patent	Front-page citations	In-text citations	Front-page or in-text citations
	(1)	(2)	(3)	(4)	(5)
Female last author	-0.0002	-0.0026	-0.0044	-0.0024	-0.0051
	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Pct. female authors (not last)	-0.0007	-0.0076	-0.0124	-0.0050	-0.0139
	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Observations	69614186	69614186	69614186	69614186	69614186
Mean of DV	0.0018	0.0550	0.0669	0.0329	0.0821
Field-year fixed effects	y	y	y	y	y
Journal fixed effects	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y

Panel B: Twins sample

			Ln citations from patents		
	Only UPPPs	Paper cited by patent	Front-page citations	In-text citations	Front-page or in-text citations
	(1)	(2)	(3)	(4)	(5)
Female last author	-0.0002	-0.0020	-0.0108	-0.0169	-0.0167
	(0.0033)	(0.0074)	(0.0130)	(0.0104)	(0.0141)
Pct. female authors (not last)	-0.0022	-0.0088	-0.0171	-0.0098	-0.0183
	(0.0022)	(0.0051)	(0.0087)	(0.0068)	(0.0094)
Observations	27398	27398	27398	27398	27398
Mean of DV	0.0228	0.3924	0.6700	0.5623	0.8635
Twin-paper fixed effects	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y

Notes: This table shows the estimation results of the gender dynamic with alternative dependent variables. Panel A uses all the papers in the MAG database, and Panel B focuses on the twins sample. Column (1) of both tables sets the dependent variable to all university patent-paper pairs (UPPPs). Column (2) considers papers cited by any patent as a dependent variable, regardless of how often they are cited. In column (3), the dependent variable is the number of front-page patent citations (likely to be legally binding). In column (4), the dependent variable is in-text patent citations (less likely to be legally binding and more likely to be added by the scientists). In column (5), the dependent variable is the sum of front-page citations and in-text patent citations. Citations to front-page articles are from Marx and Fuegi 2020, and in-text citations are from Marx and Fuegi 2022. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. Additionally, the twin regression presented in Panel B incorporates controls for the twin paper that was published first. For the gender variable, the reference category is male last author. In all the specifications, the standard errors are robustly estimated.

Table G.7: Manual verification of twins papers

	Commercialized					
	Full verification subsample		Partial or Identical Twins		Identical Twins	
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author	-0.1982	-0.1899	-0.2279	-0.2238	-0.2098	-0.1660
	(0.0949)	(0.0982)	(0.0943)	(0.0971)	(0.1152)	(0.1178)
Pct. female authors (not last)		-0.1133		-0.0604		-0.4946
		(0.2377)		(0.2419)		(0.2810)
Observations	200	200	184	184	108	108
Mean of DV	0.4700	0.4700	0.4674	0.4674	0.4444	0.4444
Twin-paper fixed effects	y	y	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in section 3.2.2 for the random sample of 100 twin pairs (200 articles), each pair consisting of one article with a female principal investigator (last author) and another with a male principal investigator, for which we conducted the human audit. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. In all the specifications, the standard errors are robustly estimated. Columns (1-2) estimate Equation 2 on all 200 twins in the random sample that we attempted to verify, regardless of the outcome. Columsn (3-4) limit the sample to those twins we confirmed to be substitutes, i.e., have scores of 2 or 3 in the rubric from Appendix C. Columsn (5-6) limit the sample to those twins we confirmed to be identical twins, i.e., have scores of 3 in the rubric from Appendix C.

Table G.8: Twins based on biological sequence and structure

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author	-0.0279	-0.0292	-0.0237	-0.0221
	(0.0115)	(0.0116)	(0.0116)	(0.0116)
Paper in twin published first			0.0355	0.0357
			(0.0056)	(0.0056)
Pct. female authors (not last)				-0.0243
				(0.0146)
Observations	6268	6250	6250	6250
Mean of DV	0.0479	0.0478	0.0478	0.0478
Bio-twin-paper fixed effects	y	y	y	y
Controls	n	y	y	y
Country/state fixed effects	n	y	y	y

Notes: This table shows the estimation results of the gender dynamic in commercialization for the subset of "twin" articles based on identical biological sequence and structure as defined in Section 3.3.2. The estimations include fixed effects for the bio-twin scientific discovery. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. When "Controls" is set to "y", the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor and a variable indicating whether a paper is in a hub. For the gender variable, the reference category is male last author. In all the specifications, the standard errors are robustly estimated.

Table G.9: Conditional logit and by-hand gender classification

	Commercialized			
	Conditional logit		Every author hand-coded	
	(1)	(2)	(3)	(4)
main				
Female last author	-0.414	-0.409	-0.152	-0.149
	(0.152)	(0.152)	(0.0565)	(0.0564)
Pct. female authors (not last)		-0.201		-0.117
		(0.191)		(0.0790)
Observations	1927	1927	1950	1950
Mean of DV	0.500	0.500	0.404	0.404
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	n	n	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2 for the conditional logistic regression (columns (1)-(2)) and a subset of hand-collected data (columns (3)-(4)). Every author was hand-coded for the subset of papers in columns (3-4) where one or the other twin in the simultaneous discovery was commercialized. Twin discoveries where neither paper was commercialized are excluded. Conditional logit necessarily excludes these. The estimations include fixed effects for the twin scientific discovery. For the gender variable, the reference category is male last author. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. In all the specifications, the standard errors are robustly estimated.

Table G.10: Robustness checks for gender classification: Alternative percentage of known authors

Panel A: All papers

	Percentage of known authors is greater than				
	0%	33%	50%	75%	100%
	(1)	(2)	(3)	(4)	(5)
Female last author	-0.0005	-0.0005	-0.0005	-0.0006	-0.0005
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Pct. female authors (not last)	-0.0020	-0.0020	-0.0021	-0.0024	-0.0019
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Observations	69614186	69608750	64908881	50761076	42799903
Mean of DV	0.0037	0.0036	0.0036	0.0034	0.0024
Field-year fixed effects	y	y	y	y	y
Journal fixed effects	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y

Panel B: Twins

	Percentage of known authors is greater than				
	0%	33%	50%	75%	100%
	(1)	(2)	(3)	(4)	(5)
Female last author	-0.0117	-0.0127	-0.0144	-0.0208	-0.0475
	(0.0044)	(0.0051)	(0.0056)	(0.0084)	(0.0180)
Pct. female authors (not last)	-0.0029	-0.0052	-0.0069	-0.0209	-0.0374
	(0.0018)	(0.0031)	(0.0039)	(0.0082)	(0.0194)
Observations	27398	22672	20338	12935	6159
Mean of DV	0.0459	0.0546	0.0592	0.0813	0.1372
Twin-paper fixed effects	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y

Note: This table displays the estimation results of the gender gap in commercialization within the twin sample specified in Section 3.2.2, as well as in the cross-sectional sample, using different thresholds for the proportion of authors' team whose gender identification probability surpasses 90%. For the gender variable, the reference category is male last author. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. Additionally, the twin regression presented in Panel B incorporates controls for the twin paper that was published first. Standard errors are robust.

Table G.11: Robustness checks for gender classification: Alternative threshold for auto-generated gender

Panel A: cross-section

	Auto-gender threshold			
	50%	75%	90%	95%
	(1)	(2)	(3)	(4)
Female last author	-0.0006	-0.0006	-0.0005	-0.0007
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Pct. female authors (not last)	-0.0009	-0.0009	-0.0020	-0.0009
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Observations	69614186	69614186	69614186	69614186
Mean of DV	0.0037	0.0037	0.0037	0.0037
Field-year fixed effects	y	y	y	y
Journal fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Panel B: twins

	Auto-gender threshold			
	50%	75%	90%	95%
	(1)	(2)	(3)	(4)
Female last author	-0.0092	-0.0102	-0.0132	-0.0122
	(0.0044)	(0.0045)	(0.0048)	(0.0051)
Pct. female authors (not last)	-0.0046	-0.0035	-0.0034	-0.0026
	(0.0039)	(0.0037)	(0.0030)	(0.0025)
Observations	27398	27398	27398	27398
Mean of DV	0.0459	0.0459	0.0459	0.0459
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Note: This table displays the estimation results of the gender gap in commercialization within the twin sample specified in Section 3.2.2, as well as in the cross-sectional sample, using different thresholds for the probability of gender assignment. For example, in column (1), the probability of being assigned a (fe)male gender should be higher than 0.5. For the gender variable, the reference category is male last author. When the probability cutoff is not met, the assigned gender is undetermined. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. Additionally, the twin regression presented in Panel B incorporates controls for the twin paper that was published first. Standard errors are robust.

Table G.12: Robustness: alternative measures of gender composition

	Commercialized				
	(1)	(2)	(3)	(4)	(5)
At least one female	-0.0094				
	(0.0038)				
Pct. female authors (incl. last)		-0.0220			
		(0.0061)			
First author female			-0.0092		
			(0.0047)		
First or last author female				-0.0121	
				(0.0040)	
Female first / Male last author					-0.0076
					(0.0055)
Male first / Female last author					-0.0158
					(0.0073)
Female first / Female last author					-0.0211
					(0.0075)
Observations	27398	24850	27398	27398	27398
Mean of DV	0.0459	0.0503	0.0459	0.0459	0.0459
Twin-paper fixed effects	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2 for alternative definitions of the gender structure of the authors on a given article. The estimations include fixed effects for the twin scientific discovery. Column (1) measures the gender structure using a binary variable equal to 1 if at least one female author is on the team. Column (2) measures the gender structure using the percentage of female authors on the team. Column (3) measures the gender structure using a binary variable equal to 1 if the first author is female. Column (4) measures the gender structure using a binary variable equal to 1 if the first or the last author is female. Column (5) measures the gender structure using a four-dummy model with labels: male first author- male last author (reference category), female first author- male last author, male first author- female last author, and female first author- female last author. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. The standard errors are robust.

Table G.13: Cross-sectional commercialization: additional controls

	Commercialized					
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author	-0.0005	-0.0007	-0.0004	-0.0006	-0.0004	-0.0003
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Pct. female authors (not last)	-0.0020	-0.0022	-0.0012	-0.0014	-0.0014	-0.0013
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Ln num commercializations at institution(s)		0.0000		0.0000	-0.0000	-0.0000
		(0.0000)		(0.0000)	(0.0000)	(0.0000)
Any author previously commercialized			0.0128	0.0133	0.0111	0.0107
			(0.0000)	(0.0000)	(0.0001)	(0.0001)
Observations	69614186	46381723	69614186	46381723	28957061	29063649
Mean of DV	0.0037	0.0050	0.0037	0.0050	0.0051	0.0051
Field-year fixed effects	y	y	y	y	y	y
Journal fixed effects	y	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y	y
PI institution fixed effects	n	n	n	n	y	n
Most frequent institution fixed effects	n	n	n	n	n	y

Notes: This table shows the estimates of the cross-sectional gender gap in commercialization with prior commercialization and fixed effects for the institution on the paper. For the gender variable, the reference category is male last author. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. In all the specifications, the standard errors are robustly estimated.

Table G.14: Twins commercialization: additional controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female last author	-0.0117	-0.0122	-0.0111	-0.0115	-0.0148	-0.0486	-0.0536
	(0.0044)	(0.0045)	(0.0044)	(0.0045)	(0.0060)	(0.0260)	(0.0258)
Pct. female authors (not last)	-0.0029	-0.0030	-0.0026	-0.0027	-0.0041	-0.0080	-0.0144
	(0.0018)	(0.0019)	(0.0018)	(0.0020)	(0.0027)	(0.0149)	(0.0164)
Ln num commercializations at institution(s)		0.0000		0.0000	0.0000	0.0001	0.0001
		(0.0000)		(0.0000)	(0.0000)	(0.0001)	(0.0001)
Any author previously commercialized			0.0179	0.0185	0.0192	0.0010	-0.0029
			(0.0035)	(0.0037)	(0.0049)	(0.0201)	(0.0207)
Observations	27398	25580	27398	25580	23090	2152	1992
Mean of DV	0.0459	0.0472	0.0459	0.0472	0.0508	0.0618	0.0622
Twin-paper fixed effects	y	y	y	y	y	y	y
Country/state fixed effects	y	y	y	y	y	y	y
Journal fixed effects	n	n	n	n	y	y	y
Most frequent affiliation fixed effects	n	n	n	n	n	y	n
Primary affiliation fixed effects	n	n	n	n	n	n	y

Notes: This table shows the estimation results of the gender gap in commercialization for the "twin" sample defined in section 3.2.2. For the gender variable, the reference category is male last author. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable that controls for the twin paper that was published first. In all the specifications, the standard errors are robustly estimated.

Table G.15: Commercialization and gender representation in scientific fields

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author	-0.0122	-0.0205	-0.0118	-0.0202
	(0.0043)	(0.0135)	(0.0044)	(0.0136)
Pct. female authors in field that year	0.0396	0.0343	0.0398	0.0343
	(0.0305)	(0.0335)	(0.0305)	(0.0335)
Female last author × Pct. female authors in field that year		0.0308		0.0312
		(0.0453)		(0.0452)
Pct. female authors (not last)			-0.0029	-0.0029
			(0.0018)	(0.0018)
Observations	27354	27354	27354	27354
Mean of DV	0.0460	0.0460	0.0460	0.0460
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2, focusing on the effect of female representation in scientific fields. The estimations include fixed effects for the twin scientific discovery. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. "Pct female in field-year" is defined as the share of authors publishing in the same field in that same year. Fields are determined by probabilistically crosswalking Microsoft Academic Graph keywords to 251 Web of Science categories. The number of observations is somewhat lower than our main twins analysis given that some papers are missing Web of Science categories.

Table G.16: Commercialization and Networks

	Ln prior coauthors of last author at firms		Commercialized	
	(1)	(2)	(3)	(4)
Female last author	-0.1971	-0.1912	-0.0106	-0.0112
	(0.0207)	(0.0206)	(0.0044)	(0.0048)
Pct. female authors (not last)	0.0170	0.0238	-0.0030	-0.0029
	(0.0134)	(0.0133)	(0.0018)	(0.0018)
Ln prior coauthors of last author at firms			0.0056	0.0058
			(0.0022)	(0.0024)
Female last author × Ln prior coauthors of last author at firms				0.0013
				(0.0051)
Observations	27382	25455	27382	27382
Mean of DV	0.7001	0.6548	0.0459	0.0459
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y
Focal paper has an author at a firm	y	n	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2, focusing on the industry network of the authors. The estimations include fixed effects for the twin scientific discovery. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. "Ln prior coauthors of last author at firms" is the logarithm of the count of coauthors of the last author *not on the focal paper* with industrial affiliations. In columns (1)-(2), the left-hand side variable is the "Ln prior coauthors of last author at firms". In columns (3)-(4), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. Column (2) omits articles that have an industry author. The number of articles is smaller than our main twins' analysis because some institutional affiliations were missing or could not be reliably classified.

Table G.17: Commercialization, gender, and attention

	Paper has boastful words		Commercialized	
	(1)	(2)	(3)	(4)
Female last author	-0.0004	0.0030	-0.0117	-0.0116
	(0.0001)	(0.0066)	(0.0044)	(0.0045)
Pct. female authors (not last)	-0.0013	0.0014	-0.0029	-0.0029
	(0.0001)	(0.0047)	(0.0018)	(0.0018)
Paper has boastful word(s)			0.0024	0.0024
			(0.0061)	(0.0072)
Paper has boastful word(s) × Female last author				-0.0014
				(0.0171)
Observations	69614186	27398	27398	27398
Mean of DV	0.0580	0.0834	0.0459	0.0459
Twin-paper fixed effects	n	y	y	y
Field-year fixed effects	y	n	n	n
Journal fixed effects	y	n	n	n
Country/state fixed effects	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2 (except column (1), which is using all the publications in MAG), focusing on self-promotion. Self-promotion is measured by the use of "boastful words". "Paper has boastful words" indicates that the title or abstract uses one or more words such as "breakthrough" which are defined by Lerchenmueller, Sorenson, and Jena 2019 as boasting, with the exception that "novel" is not treated as a boasting word when it appears in a bigram with "coronavirus." In columns (1)-(2), the left-hand side variable is "Paper has boastful words". In columns (3)-(4), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.2. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, and a variable indicating whether a paper is in a hub. Additionally, the twin regression presented in columns (2-4) incorporates controls for the twin paper that was published first.

Table G.18: Commercialization and gender homophily (Counterfactual paper-patent twin dyads)

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author	-0.0932	0.6397	-0.0967	0.6283
	(0.0531)	(0.1591)	(0.0531)	(0.1588)
Female last author × Pct male inventors		-0.9924		-0.9811
		(0.1925)		(0.1916)
Pct. female authors (not last)			-0.0948	-0.0826
			(0.0633)	(0.0624)
Observations	2382	2382	2382	2382
Mean of DV	0.4996	0.4996	0.4996	0.4996
Twin-paper fixed effects	y	y	y	y
Country/state fixed effects	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twin sample defined in Section 3.2.2, focusing on homophily in team composition. In particular, the table estimates Equation 1, which instead of paper-level analysis, performs patent-paper level analysis of *possible* PPPs. The sample is limited to twin discoveries where one or the other twin is commercialized. The commercializing patent in the realized PPP is then artificially paired with the uncommercialized article in the twin to create a counterfactual PPP, given the intuition that the uncommercialized article in the twin discovery might well have been paired with the patent that commercialized the other article in the twin. The percentage of male inventors on the focal patent is calculated using USPTO's inventor-gender file, but excludes inventors who are also on the paper in order to avoid biasing toward realized PPPs. All the models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, the journal impact factor, a variable indicating whether a paper is in a hub, and a variable indicating the twin paper that was published first. In all the specifications, the standard errors are robustly estimated.

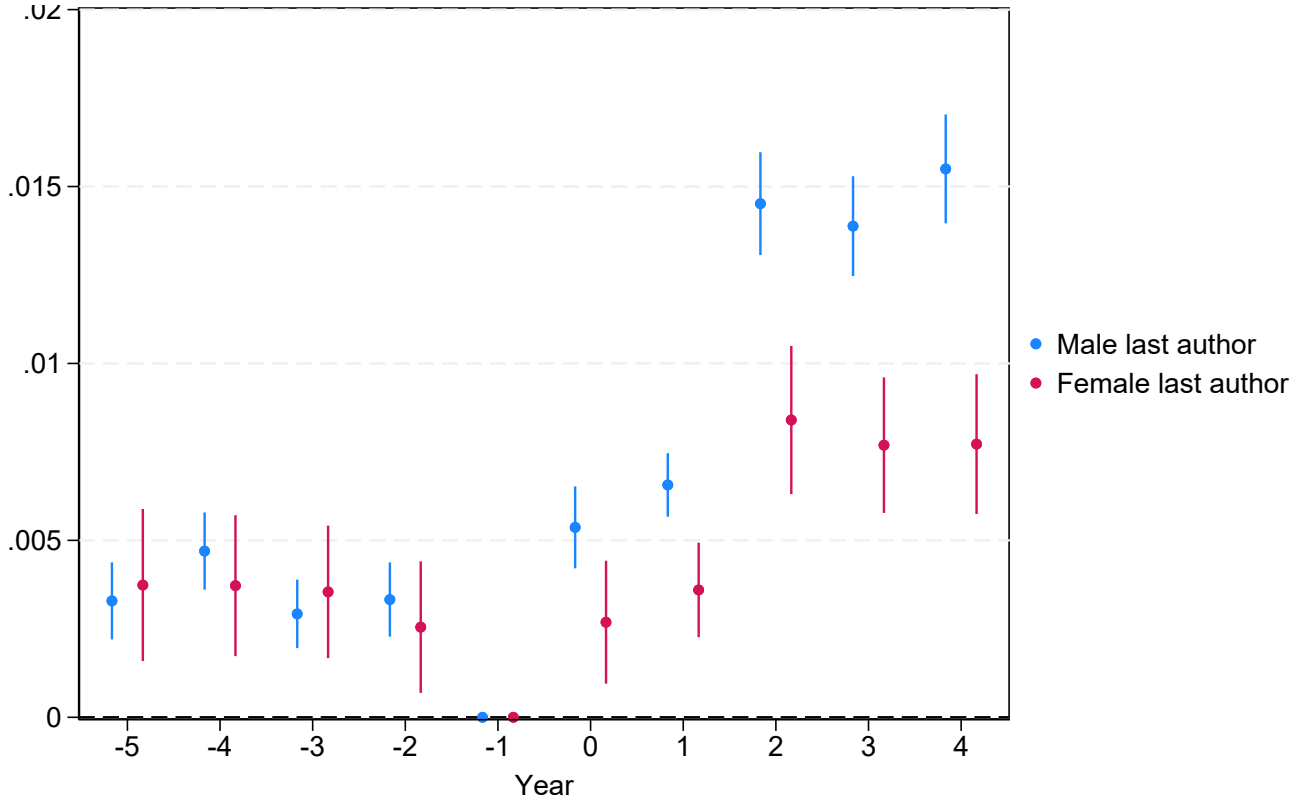
Table G.19: Becker's outcome test: IV estimates

	Financial value of patent			
PI gender	male	female	male	female
	(1)	(2)	(3)	(4)
leave-out assignee-year % comm.	0.6886	0.6876		
	(0.0063)	(0.0190)		
Commercialized			1.7567	1.8648
			(0.0279)	(0.0817)
Observations	609734	73663	609734	73663
Mean of DV	0.0737	0.0628	0.2340	0.2065
Year fixed effects	y	y	y	y
Assignee fixed effects	y	y	y	y

Notes: This table shows the estimation results of Becker's outcome test using an instrumental variable approach: the leave-out assignee-year share of commercialization. Columns (1) and (2) represent the first stage of this two-stage-least square estimation and show that the instrument predicts the likelihood of commercialization. Columns (3) and (4) represent the second stage of this two-stage least squares estimation for each gender category. The dependent variable in columns (3) and (4) is financial value of a patent as calculated by Kogan et al. 2017 based on stock market reactions to its issuance. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. In all the specifications, the standard errors are robustly estimated.

H Additional Figures

Figure H.1: Impact of Open Access mandates on commercialization



Notes: Figure H.1 shows the estimation results of the staggered triple difference to assess the effect of Open Access on the gender gap in commercialization. Panel A shows the staggered difference-in-difference separately for male-authored (last author male) and female-authored papers (last author female). The unit of observation is the academic article. The dependent variable is the commercialization measured by the patent-paper-pair whose assignee is a firm. All estimates include controls for the number of authors, the authors' average prominence and institutions, the fields, and years dummies. The coefficient for event time -1 is omitted to normalize the gender commercialization gap to zero in the year prior to the policy.

References for Appendix

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Benson, Alan, Danielle Li, and Kelly Shue. 2019. "Promotions and the peter principle." *The Quarterly Journal of Economics* 134 (4): 2085–2134.

Bikard, Michaël, and Matt Marx. 2019. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science.*

Bjork, Bo-Christer, Annikki Roos, and Mari Lauri. 2009. "Scientific journal publishing: yearly volume and open access availability." *Information Research: An International Electronic Journal* 14 (1).

Bryan, Kevin A, and Yasin Ozcan. 2021. "The impact of open access mandates on invention." *Review of Economics and Statistics* 103 (5): 954–967.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science* 48 (1): 1–23.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Durvasula, Maya, C Scott Hemphill, Lisa Larrimore Ouellette, Bhaven Sampat, and Heidi L Williams. 2023. "Data for: The NBER Orange Book dataset: A user's guide." *Research Policy* 52 (7): 104791.

Eysenbach, Gunther. 2011. "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact." *Journal of medical Internet research* 13 (4): e2012.

Friesike, Sascha, Bastian Widenmayer, Oliver Gassmann, and Thomas Schildhauer. 2015. "Opening science: towards an agenda of open science in academia and industry." *The journal of technology transfer* 40:581–601.

Furman, Jeffrey L, and Scott Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative research." *American Economic Review* 101 (5): 1933–1963.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225 (2): 254–277.

Gowers, Timothy, and Michael Nielsen. 2009. "Massively collaborative mathematics." *Nature* 461 (7266): 879–881.

Hochberg, Yael, Ali Kakhbod, Peiyao Li, and Kunal Sachdeva. 2023. *Inventor Gender and Patent Undercitation: Evidence from Causal Text Estimation.* Technical report.

Huang, Ruidi, Erik J Mayer, and Darius P Miller. 2022. "Gender bias in promotions: Evidence from financial institutions." *SMU Cox School of Business Research Paper,* nos. 21-18.

Khabsa, Madian, and C Lee Giles. 2014. "The number of scholarly documents on the public web." *PloS one* 9 (5): e93949.

Klar, Samara, Yanna Krupnikov, John Barry Ryan, Kathleen Searles, and Yotam Shmargad. 2020. "Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work." *PloS one* 15 (4): e0229446.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. 2017. *Data for: Technological Innovation, Resource Allocation, and Growth.* https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data. Accessed: 2023.

Lerchenmueller, Marc J, Olav Sorenson, and Anupam B Jena. 2019. "Gender differences in how scientists present the importance of their research: observational study." *bmj* 367.

Markman, Gideon D, Donald S Siegel, and Mike Wright. 2008. "Research and technology commercialization." *Journal of Management Studies* 45 (8): 1401–1423.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal.*

———. 2022. "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations." *Journal of Economics & Management Strategy* 31 (2): 369–392.

McKiernan, Erin C, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg, et al. 2016. "How open science helps researchers succeed." *elife* 5:e16800.

Rassenfosse, Gaétan de, and Kyle Higham. 2020. "Wanted: a standard for virtual patent marking." *Journal of Intellectual Property Law & Practice* 15 (7): 544–553.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084.*

RoyaltySource. 2021. *RoyaltySource.* https://www.royaltysource.com/.

Staudt, Joseph. 2020. "Mandating access: assessing the NIH's public access policy." *Economic policy* 35 (102): 269–304.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

Tartari, Valentina, and Ammon Salter. 2015. "The engagement gap:: Exploring gender differences in University–Industry collaboration activities." *Research Policy* 44 (6): 1176–1191.

U.S. Patent and Trademark Office. 2023. *PatentsView.* https://www.patentsview.org. Accessed: 2023.

Ware, Mark, and Michael Mabe. 2015. "The STM report: An overview of scientific and scholarly journal publishing."