Startups attenuate demographic disparities in biomedical, cultural, and commercial innovation

Authors: Šimon Trlifaj[1]*, Yian Yin[2]*

Affiliations:

[1] Central European University; Vienna, Austria

[2] Cornell University; Ithaca, United States of America

*Corresponding authors. Emails: trlifaj_simon@phd.ceu.edu, yian.yin@cornell.edu

Abstract: New inventions often focus on the privileged and neglect the underrepresented. Large incumbent firms are typically the focus of public, regulatory and business initiatives to attenuate gender, race, and ethnic disparities in innovation. Here, we present large-n evidence that startups, in contrast to incumbents, play a disproportionate role in innovating for underrepresented groups across new products in biomedical (patents), cultural (movies), and commercial (retail) domains. This pattern is consistently stronger in categories and time periods of greater underrepresentation. Leveraging a regulatory shift in biomedical innovation, our results show that as that as female-oriented innovation mainstreamed, the role of startups in innovating for female weakened. This dynamic is not due to startups employing more underrepresented inventors and creators (social push) or attending to lower- or higher-income consumers (market pull). We propose a simple profit-seeking explanation for this pattern (competitive fill), where startups attend to the needs of underrepresented groups as they seek underserved consumer niches.

Main Text:

New inventions often focus on the privileged and neglect the underrepresented (*1*, *2*). Patients of race or ethnic minority may receive suboptimal supplemental oxygen as pulse oximeters perform worse on darker skin, even decades after their invention (*3*). Women continue being more vulnerable in car crashes as seat belts, airbags and other car safety features had been developed for male body types, despite recent progress (*4*). Black people may feel underrepresented in mainstream culture as, for instance, the first superhero blockbuster with a Black lead arrived only decades after the genre had taken hold (*5*). Nowhere is this tension between efficiency and inequity more visible than in private sector—which accounts for most innovation globally (*6*). If we are to achieve equitable progress, this naturally raises a question: which private sector firms can we expect to address the needs of underrepresented groups?

Public attention typically focuses on large incumbent companies to demand equity. A common view—grounded in business, policy, and moral theories—holds that incumbents possess both the practical means and the moral obligation to serve underrepresented groups (*7*, *8*). Regulators, business leaders, and consumer advocates have codified this logic in initiatives that press private firms toward greater inclusivity such as corporate social responsibility, fair-access obligations, pubic pressure campaigns, non-financial disclosure requirements, or equal opportunity policies (*9*, *10*). These typically focus on established companies, as does public attention to the recent discontinuations of these policies (*10*). This yields a clear baseline expectation: that serving minority communities scales with firm size and equitable progress emerges from established market participants rather from new entrants.

Countering this common view, an equity-oriented reading of a major business theory gives reasons to believe startups might be an unexpected pioneer of minority-focused progress. The innovator's dilemma theory proposes that to succeed on the market, new entrants need to find a consumer group underserved by the mainstream market (*11*). To the extent that societally underrepresented groups constitute such a market niche, this would suggest that startups might drive innovation towards such groups. This then yields a clear alternative hypothesis: that startups would disproportionately serve minority communities and drive equitable progress.

It is well documented that young firms play a disproportionate role in economic growth and in generating novel, high-impact innovations (*12–14*). Recent empirical research also revealed that factors such as identity and life experiences of the inventor (*15–19*) and income of the consumer influence which needs innovation focuses towards (*1*, *20–22*). At the intersection of firm age and direction of innovation, however, it remains empirically unanswered if startups play a role in directing innovation towards underrepresented groups.

Here, we set out to investigate this question empirically, building on three large-scale datasets spanning biomedical inventions (157 thousand patents, 1976–2013), movie productions (473 million ratings and 42 thousand movies, 1951–2023), and retail purchases (451 million purchases on 2 million new products, 2007–2020). Contrary to the baseline hypothesis, our analysis presents

systematic evidence that startups—in contrast to incumbents—play a distinctive yet previously unrecognized role in serving underrepresented groups. Across all three domains, we find startups are systematically more likely than incumbents to introduce new products attending to underrepresented groups. This pattern is consistently stronger in categories and time periods of greater underrepresentation across race, ethnicity, and gender. Together, our analysis presents systematic evidence that as opposed an obligation carried by incumbents in spite of their business interests, progress oriented towards underrepresented groups is disproportionately delivered by startups, and may well be in their business interests.

Results

We start by analyzing the role of startups in the context of a policy-induced underrepresentation in biomedical innovation (Study 1), using 1,530 U.S. biomedical patents attributable to male or female needs (*15*), filed between 1976 and 1993. During this period, research into drugs that addressed female medical needs in the U.S. was constrained by restrictions on female clinical trials participation. This led to the underrepresentation of female patients in drugs innovations (*15*, *23*–*25*), representing a notable case of policy-induced underrepresentation. The restrictions affected all firms equally, establishing the baseline hypothesis that incumbents and startups would be similarly constrained in innovating for female. Alternatively, if startups tend innovate for underrepresented groups, we should see them disproportionately innovating towards female needs during this period of heightened underrepresentation.

Figures 1**A, D** show that during this period, startups were disproportionately more likely to file drug patents addressing female medical needs. Prior to the 1993 reform, 38 % of patents were female focused (fig. A1). Applying the appropriate transformation $1 - e^{coefficient}$ to log-odds regression estimates (fig. A2), startups were 65 % (P .025) more likely to introduce female-oriented rather than male-oriented patents, compared to incumbents.

Many factors may explain this association. Notably, recent studies have uncovered a widespread association between female inventors and biomedical patents oriented towards female health needs (*15*, *16*). To this end, we include controls for inventor gender and patent category and year interactions throughout our results. Prior to 1993, the startup association is at least as large as the association between female inventors and female focus of biomedical patents found in prior research (*15*), and it is robust to its inclusion, suggesting this pre-1993 startup association is not due to higher presence of female inventors in startups.

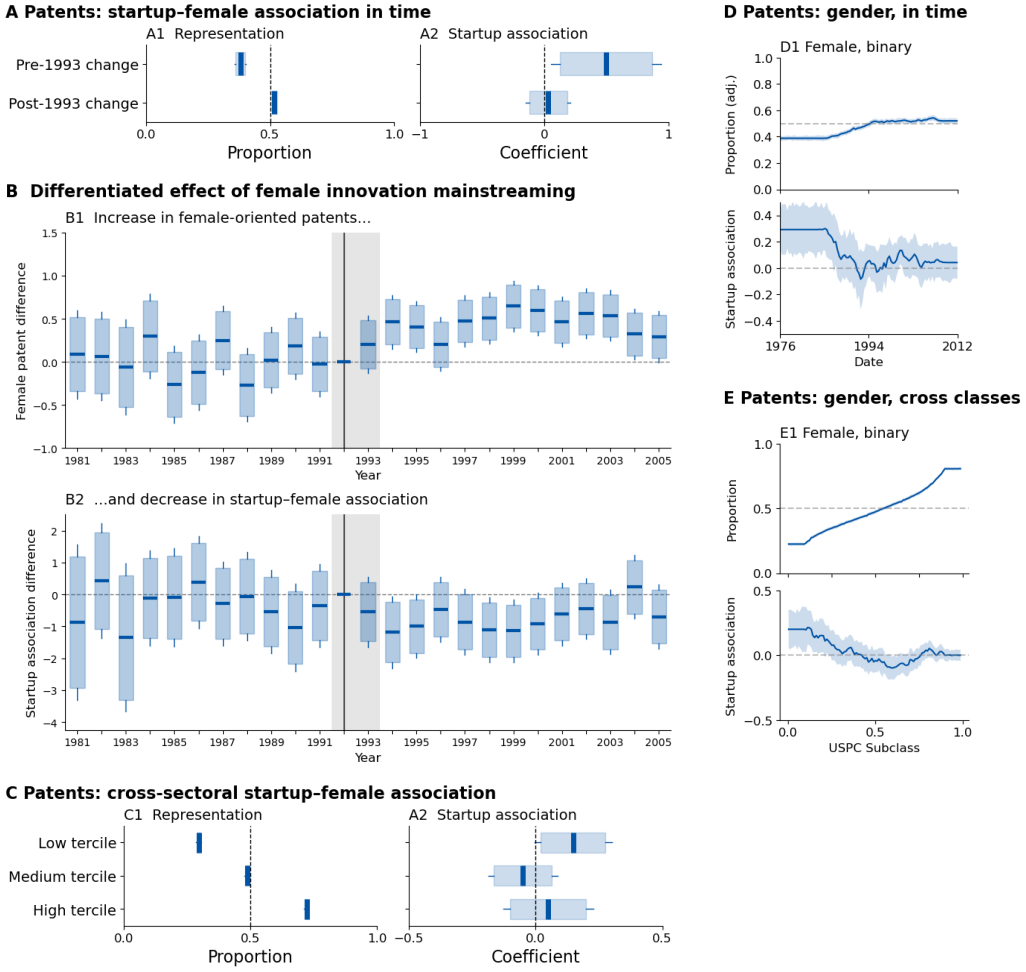


Fig 1: In biomedical inventions, startups tend to introduce female-oriented inventions during periods and in categories of greater underrepresentation. (A, C) Proportion of female-oriented patents (A1, C1) and startup–female association in log odds ratio regression coefficients (A2, C2) by time and patent class tercile of female representation, tables S1 and S5 respectively. (B) Increase in female patent orientation in drugs patents comparted to other biomedical patents after FDA policy change (B1) and relative decrease of startup–female in log odds ratio (B2), triple difference regression. (D, E) Proportion of female-oriented patents (D1, E1) and startup–female association in log odds ratio (D2, E2) as moving averages across time and patent classes by female proportion, bootstrapped CIs. In all panels, light blue marks 90 percent CI.

Changing female underrepresentation in biomedical innovation

More importantly, the discovered association is absent after the policy restriction is lifted. In 1993, the Federal Drugs Administration (FDA) weakened the clinical trials participation restrictions,

which mainstreamed female-oriented biomedical patenting (*23–25*). Descriptively, we find an increase in female-oriented drugs patents, and a decrease in startup female association (fig. 1A, D) after 1993. In Study 2, we leverage this policy change and examine if the orientation of startup innovation changes as the underrepresentation of a group alleviates.

To examine the FDA policy change and startup–female association, we leverage the fact that this change only affected drugs in a triple-difference model. We analyze 112 thousand biomedical patents filed between 1981 and 2005 in drugs and other non-drug biomedical patents. The model absorbs pre-1993 levels and post-1992 changes in female-orientation in drug, other non-drug, startup, and incumbents patents, and isolates the differentiated change in startup drugs female orientation following the policy change (see SI S2 for full specification).

Figure 1**B** shows that as the policy shift increased overall female-oriented drugs patenting, compared to other non-drugs biomedical patents (B1), it also decreased relative startup-orientation towards female needs (B2). By focusing on the difference between drugs and other patents, we can isolate the change among patents affected by the FDA policy from a broader policy change in biomedical patenting during the period (*23–25*). Results from the model show that following the lifting of the restrictions, the relative probability of startups innovating towards female decreased by 33 % (P .028, SI S2). The post-reform decline is similar in magnitude to the pre-1993 startup–female association, consistent with startups and incumbents converging to comparable rates of female-oriented invention as the reform mainstreamed female-oriented innovation.

Fig. 1A–C suggest that it is the level of underrepresentation on the market—rather than gender per se—that drives the differing rate at which startups and incumbents innovate toward female. This raises another interesting question: does the pattern of heightened startup focus in areas of lower representation extend to variation between categories? To this end, we leverage the fact that the underrepresentation of female-related patents is non-uniform across categories. In study 3 we consider 1,643 patent subclasses in our data, dividing them into terciles of low (mean 30 %, e.g. subclass 606/198: *Surgery, Internal pressure applicator, Expanding dilator*), medium (49 %) and high (72 %, e.g. subclass 606/45: *Surgery, Applicators, Cutting*) female representation, excluding those with fewer than ten patents. We then analyze startup association with female patenting in each tercile.

Fig. 1**D** shows that with increasing female representation in patent classes (D1), the female-startup association decreases (D2). For patents in lowest third of classes with lowest female representation, startups are estimated to be 16 % ($P < 0.06$) more likely to file a patent oriented towards female medical needs. There is no association in medium- and high-representation classes (P 0.48 and 0.58, respectively). This confirms the consistent pattern that startups focus on female biomedical innovation in a context of higher underrepresentation.

Underrepresentation in mass culture

Together, Studies 1–3 document a strong link between startups and innovations for underserved communities in biomedical patents. We next investigate if this pattern generalizes to other domains. To this end, we consider cultural innovations in movie production—an industry broadly reflective popular culture of the past several decades, and also associated with persistent underrepresentation of female and race and ethnic minorities (*26*). We draw on data about 42 thousand U.S. feature movies sources from TMDB, a comprehensive free database of movies. Motivated by theories of identity representation and survey evidence on gendered viewing preferences (*26*, *27*), we consider lead-actor gender as a proxy for films targeting female audiences, and investigate if new production companies feature female main characters at higher rates.

Study 4 again leverages cross-sectional variation by ranking 348 movie genre combinations by female representation, excluding combinations with fewer than ten movies. The broad pattern presented in fig. 2**A**, appears remarkably consistent with our biomedical analysis (Study 3). The higher the proportion of female lead characters, the stronger the startup association (A1). For movies in genre combinations with lowest lead female representation (18 %, e.g. *drama, war and tv*), startups are estimated to be 24 % ($P < .01$) more likely to feature female lead actor (fig. 2**B**). There is no significant association for movies in medium-representation genres (P 0.28), and a negative association (P .02) in high representation genre combinations with highest lead female proportion (55 %, e.g. *drama, romance, music*).

A similar pattern is clear across time (fig. 2**B**): during the second half of the 20th century, female lead actors were less common, and startups were disproportionately more likely to feature them. As female lead actors became more common since the 1990s, this association waned. What emerges is a consistent pattern of increased startup orientation towards female audiences in genres and time periods where mainstream movie production underserves them.

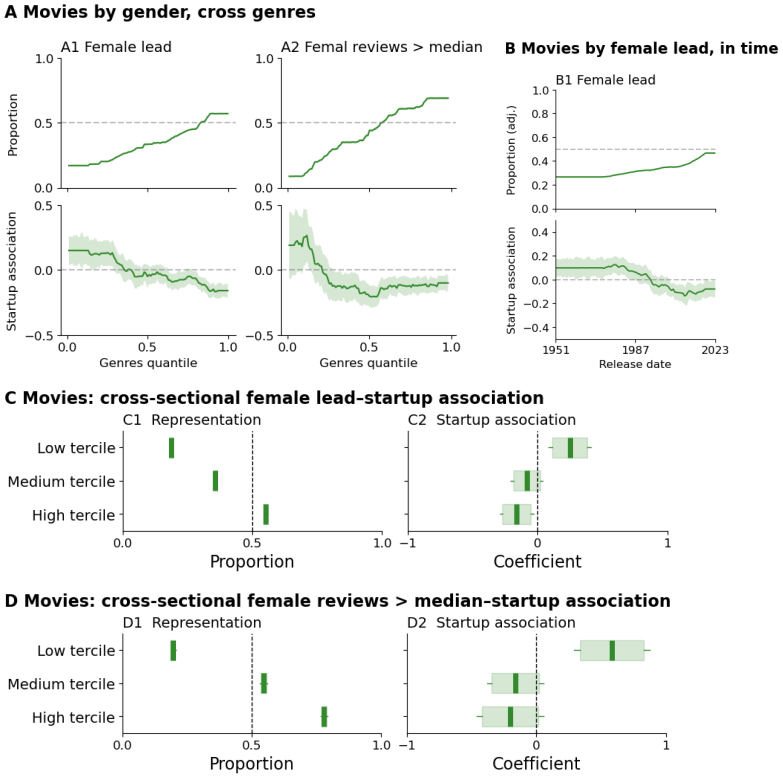


Fig. 2: In movies, startups feature more female lead actors and attract more female viewers at time periods and in genres of greater underrepresentation. (A, B) Proportion of female lead actor movies and startup–female association in log odds ratio (A1, B1), proportion of above median female reviews and startup–female association as log odds ratio (A2), moving averages across movie genres by female representation and time, bootstrapped CIs. (C, D) Proportion of female lead actors and above-median female reviewers (C1, D1) and startup–female association with startups in log odds ratio regression coefficients (C2, D2) by movie genres tercile of female representation, tables S6 and S8, respectively. In all panels, light blue marks 90 percent CI.

To validate that these casting choices of movie studios correspond to audience preferences, study 5 considers an alternative approach to measure gendered consumption of movies. We connect our movies dataset to gender-specific counts of 472 million rankings from IMDB, the largest database of movie ratings. In this dataset, movies with male lead actors receive an average of 16 % female ratings, compared to 25 % for movies with female lead actor, a difference of 56 % ($P < .01$). We then analyze if movies popular among female audiences, measured as above median (.18) proportion of female voters, are more likely to be produced by startups. Fig. 2A shows a pattern consistent with the female lead results: in genres with fewer movies popular among female audiences, startups are disproportionately more likely to introduce such movies (A2).

Figure 2**B** presents results of a formal test of this association. For movies in genres with lowest proportion of movies popular among female audiences, startups are estimated to be 80 % (P < .01) more likely to produce a movie popular among female audiences (B2). There is no significant startup association for movies in medium-representation genres (P .2), and negative association (P .08) in genre combinations most popular among female audiences. In S3.3, we further validate these results with a model that counts reviews by gender.

In a common pattern across the analysis of gender thus far, startups are not always disproportionately introducing products targeting female consumers overall (Studies 1–5). They do so in time periods and product categories of strong underrepresentation. This corresponds to the fact that female represent approximately half of the population, which is hardly a niche consumer group—but may be at time periods and product categories of underrepresentation. We next extend our analysis to groups which do constitute a small proportion of population overall.

Underrepresentation of race and ethnic minorities

We next analyze startup orientation towards race and ethnic minority groups. Our hypothesis shifts in this dimension of underrepresentation, as in the U.S., non-white and Hispanic population currently represents about 40 % of the population, and less in earlier periods from which some of our data originates. If startups tend to innovate for underrepresented groups, we should therefore see them disproportionately innovating towards race and ethnic minorities across categories and time.

First, Study 6 investigates if startups are more likely to feature race or ethnic minority lead character (proxied by actor's race or ethnic minority). Fig. 3**A** shows that startup production companies are disproportionately more likely to feature minority lead actor in their movies across genres. The pattern is stronger in genres of lower representation. In genres with lower minority lead character representation, startups are 54 % (P < .01) more likely to feature a minority lead actor movie than incumbents (fig. 3**C**). The association is weaker but still significantly positive in mid (22 %, P = .02) and high (31 %, P < .01) representation genres. Over time, too, startups are disproportionately more likely to feature minority lead characters in periods of higher underrepresentation (fig. 3**B**).

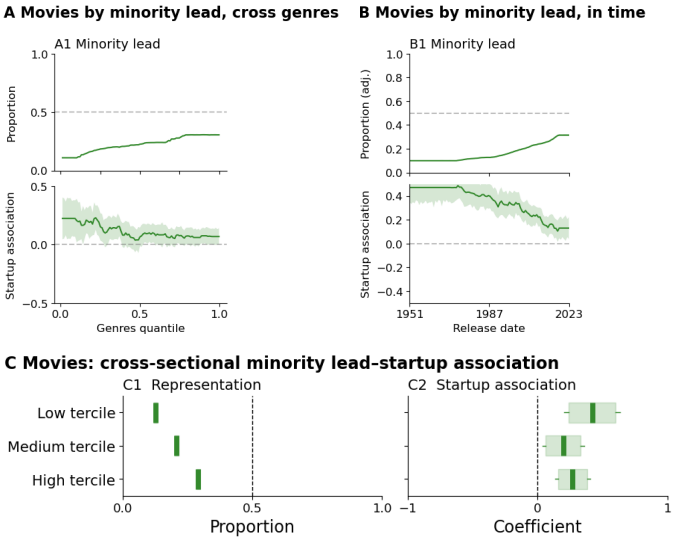


Fig. 3: In movies, startups feature more minority lead actors, especially at time periods and in genres of greater underrepresentation. (A, B) Proportion of race or ethnic minority lead actor movies and startup–minority association in log odds across genres ranked by minority representation (A1) and time (B1), bootstrapped CIs. (B) Proportion of minority lead actors (B1) and startup–minority association in log odds ratio regression coefficients (B2) by movie genres tercile of minority representation, table S7. In all panels, light blue marks 90 percent CI.

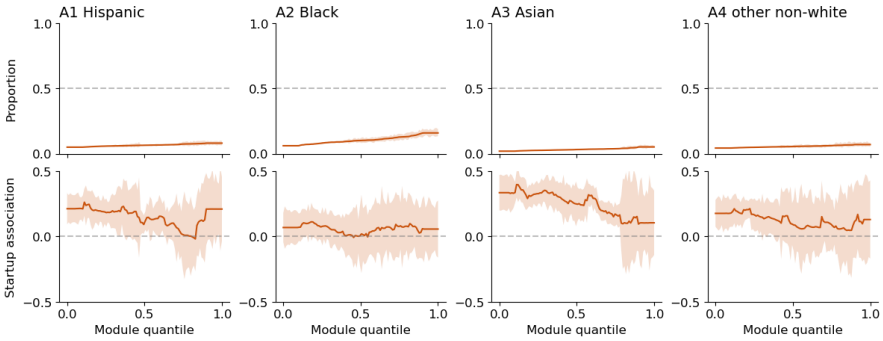
Next, we test these patterns of startup innovation on a dataset which links consumer and firm characteristics in retail purchases. Study 7 leverages barcode-level purchase data from the Nielsen Consumer Panel, a sample of 60 thousand U.S. households with self-reported race and ethnic identity on the household level. Fig. 4**A** shows that startups are more likely to introduce product that attract disproportionately more purchases by consumers of race (Black, Asian, and other non-white, compared to white) end ethnic (Hispanic, compared to non-Hispanic) minority status.

This positive association is persistent across product categories and stronger when representation is weaker. Results from regressions in fig. 4**B–E** show that for all these race and ethnic identities, the strongest association with startup products occurs in the third of categories with lowest representation of each of those groups, while controlling for product-level characteristics. In those product categories, startups are estimated to attract 20 % ($P < .01$) more purchases by Hispanic households, 9 % ($P < .01$) by Black households, 43 % ($P < .01$) by Asian households, and 20 % ($P < .01$) by other non-white households. In mid- and high- representation product categories, the association ranges from 28 % ($P < .01$) by Asian households to positive but insignificant 1 % ($P = .13$) by Black households in mid-representation categories.

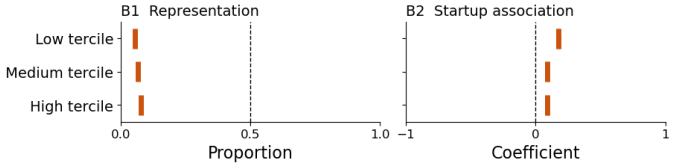
Throughout these minority groups, startups are consistently more likely than incumbents to introduce new products that attract race and ethnic minority consumers, consistent with the hypothesis that they do attend to underrepresented groups. We also report startup association with

the count of female household members in SI 3.3., which are consistent with our earlier findings on gender, although complicated by data aggregation.

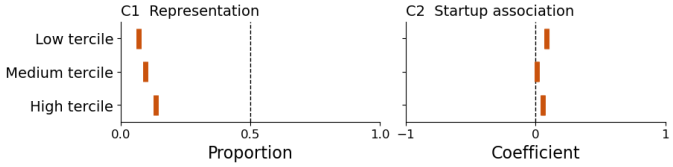
A Retail by minority households, cross categories



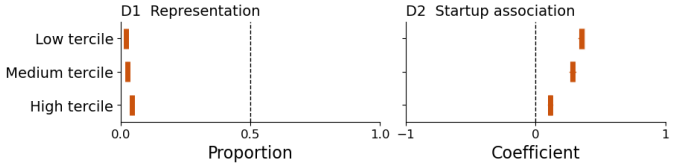
B Retail: cross-sectional Hispanic / Latino association



C Retail: cross-sectional Black association



D Retail: cross-sectional Asian association



E Retail: cross-sectional other non-white association

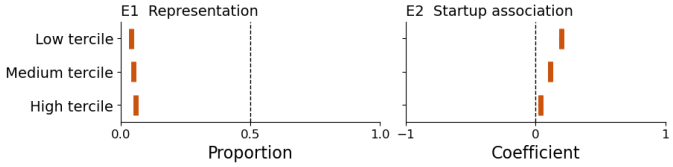


Fig. 4: In commerce, startups introduce product that attract disproportionately more purchases from race and ethnic minority households. (A, B) Proportion of minority purchases and startup–minority association in log odds across product categories ranked by minority representation (A1–A4), bootstrapped CIs. (B–E) Proportion of minority household purchases (A1–E1) and startup–minority association with startups product purchases (A2–E2) by product category tercile of minority representation, tables S10–S13. In all panels, light blue marks 90 percent CI.

Interactions with income and identity

We further examine interactions with income and creator identity. Race and ethnic minority households typically possess lower income (*28*), raising the question that whether the startup association can be explained by startups focusing on introducing low-quality products to low-income households. Our extended results (table S16) find the opposite income pattern: startups are more likely to introduce products popular among high-income households, and less likely among low-income households. The startup association with race and ethnic minorities increases or remains comparable among mid- and high-income households within each group, showing that income does not drive the results. Neither do we find startups introducing lower quality, less-innovative products. Startups tend to introduce products of higher quality (SI S6), including for underrepresented consumers, in patents (quality measured as 10-year citations) and retail (nutri-score).

Finally, we investigate the interaction between startup firms and identity of the inventor and director (S4). In biomedical inventions, we expand on models presented in (*15*) using biomedical data covering 1976–2013. The results show both the association between startup innovation and female orientation being robust to the gender composition of the inventor team, and vice-versa. Strikingly, these two associations reinforce each other—the co-occurance of startups and female inventor teams has the highest chance of inventing for female—while remaining independently significant and positive. In movies, we include an interaction for move director gender. The results show female directors both featuring significantly more female lead characters and attracting more female viewers. The startup–female association in low tercile of genres remains unchanged for both female lead (table S16) and female reviews (table S17), again suggesting the association between startups and female orientation of movies is independent of movie director identity.

Discussion

By analyzing datasets spanning 157 thousand biomedical patents, 42 thousand movies, and 2 million consumer products, we uncover a simple yet remarkably consistent empirical regularity: startups are comparatively more likely to innovate for the underrepresented gender, race and ethnic groups. This startup association is not static: it is strongest precisely when and where marginalization is deepest, and weakens as a group representation increases. This pattern suggests that startups play a distinctive corrective role where market biases have historically or structurally underserved demographic groups.

Together, these results demonstrate that 'who innovation is for' is shaped not only by *social push*: the identities and experiences of inventors and creators inside firms (*15–19*) and *market pull*: consumer income and market growth rate (*1*, *20–22*), but also by *competitive fill*: the comparative position of the firms and consumers on the market. Prior literature has shown that female and minority innovators are more likely to produce for people like themselves. Our results show that startups exhibit a comparably strong tilt even after accounting for both inventor identity and consumer income. The startup–minority link is therefore not simply a compositional artifact of who chooses to work at startup firms, but a feature of how those firms position themselves relative to incumbents. At least in one domain we studied, biomedical patents, these patterns reinforce one another, indicating that innovator identify and organizational position can function not only distinct but also complementary levers shaping the demographic reach of innovations.

Much of existing theory and practice around equitable innovation rests on two linked assumptions: that large incumbent firms are the primary vehicles through which markets will become more inclusive, and that equitable progress is mainly driven by normative commitments inside these organizations—through CSR programs, pledges, or internal targets (*9*, *10*, *29*, *30*). Our results challenge both, suggesting that prevailing, incumbent-centric and norm-centric approaches consider only one of the mechanisms which drive innovation equity. Practical efforts to change who innovation serves should also act on the conditions of entry and survival for new firms.

Analyses of market structure and innovation in economics have largely focused on efficiency and the rate of innovation—from Schumpeterian arguments that market power can finance risky R&D to Arrow-style arguments that competition sharpens incentives to innovate (*14*, *31*, *32*). Our results add a third dimension to this debate: whom that innovation serves. Startups are the firms most likely to introduce innovations for underrepresented groups, yet they are also the most exposed to liquidity contains, regulatory compliance costs, and killer acquisitions (*33–35*). Through this lense, the vulnerability of startups has direct consequences for diversity and inclusion in innovation. Policies and market structures that raise the cost of entry, privilege scale, or accelerate consolidation may risk weakening a key channel through which new products have historically

served women, minorities, and other under-served groups. Fostering entrepreneurial environments that protect and nurture startups—by lowering barriers to entry, expanding access to capital, or slowing premature absorption—is likely to affect both the rate and direction of innovation, opening the possibility of progress that is efficient and equitable.

Materials and Methods

We compile three large-scale datasets to analyze a link between startups and minority groups, the details of which are laid out in Supplementary Information S1. In Patents, we start with all U.S. utility patents from PatentsView in the NBER category drugs and medical (*36*). We use male and female orientation of patents based on the Medical Text Indexer collected by (*15*), firm incorporation date based on matching to OpenCorporates by (*37*), and differentiate between drugs and other biomedical patents based on the NBER classification (*36*). As a proxy for quality, we calculate 10 years of forward citations.

In Movies, we start with The Movie Database (TMDB) database, which we query for information about producers, cast, genres, and release date. We restrict the sample to U.S. non-documentary non-animated feature films. To identify startups, we match production companies by name to OpenCorporates and consider the older production company in the case of co-productions (robustness in S). We use lead actor gender information from TMDB and estimate race and ethnic minority status and gender when unavailable using a large language model (*38*, *39*, SI 1.2). We further link to an aggregate dataset of reviews from the International Movie Database (match rate 34 percent).

In Retail, we start with the Nielsen Consumer Panel dataset, which continuously collects barcode-level purchase information from a representative sample of approximately 60 thousand U.S. households. We rely on data from 2008 to 2020 for household characteristic, including race and ethnicity, and from 2004 to identify product introduction. To identify startups, we link barcodes to GS1 data, and further match using company name and address to OpenCorporates, a dataset of firm incorporation information. We use OpenStreetMap API to geolocate addresses in the matching process. On a subset of food products, we estimate quality using nutri-score provided by OpenFoodFacts (*40*), which we match on barcode (match rate 24 percent). Throughout these datasets, we consider products introduced within three years of company establishment as startup.

These datasets are limited in several ways. We observe gender as a female – male binary, as data on other gender identities are unavailable (biomedical, retail, movie reviews) and because of the very small number of movies with non-binary lead actors (S1.2). We only observe race and ethnicity in Nielsen Consumer Panel at the household level, which flattens within-household variation, and which we interpret as the identity of the main reporting household member. We also do not observe corporate structures in firm incorporation data and therefore cannot distinguish

independent startups from those backed by an incumbent, a practice especially prevalent in movies. We note that such practice would still be broadly consistent with the innovator's dilemma theory, which encourages incumbents to back startups to develop niche market opportunities.

5

References and Notes

1. X. Jaravel, The Unequal Gains from Product Innovations: Evidence from the U.S. Retail Sector. *The Quarterly Journal of Economics* **134**, 715–783 (2019).

2. K. A. Wailoo, V. J. Dzau, K. R. Yamamoto, Embed equity throughout innovation. *Science* **381**, 1029–1029 (2023).

3. E. R. Gottlieb, J. Ziegler, K. Morley, B. Rush, L. A. Celi, Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit. *JAMA Internal Medicine* **182**, 849–858 (2022).

4. M. Z. Abrams, C. R. Bass, Female vs. male relative fatality risk in fatal motor vehicle crashes in the US, 1975-2020. *PLoS One* **19**, e0297211 (2024).

5. E. Claverie, "Blade and Black Panther: Strategic Blackness and the Rights of Kings" in *Copyright Vigilantes: Intellectual Property and the Hollywood Superhero*, E. Claverie, Ed. (University Press of Mississippi, 2024; https://doi.org/10.14325/mississippi/9781496851338.003.0006), p. 0.

6. National Science Board, National Science Foundation, Research and Development: U.S. Trends and International Comparisons, Science and Engineering Indicators 2024 (2024); https://ncses.nsf.gov/pubs/nsb20246/.

7. E. Garriga, D. Melé, Corporate Social Responsibility Theories: Mapping the Territory. *Journal of Business Ethics* **53**, 51–71 (2004).

8. G. Aras-Beger, F. D. Taşkın, "Corporate Social Responsibility (CSR) in Multinational Companies (MNCs), Small-to-Medium Enterprises (SMEs), and Small Businesses" in *The Palgrave Handbook of Corporate Social Responsibility*, D. Crowther, S. Seifi, Eds. (Springer International Publishing, Cham, 2021; https://doi.org/10.1007/978-3-030-42465-7_69), pp. 791–815.

9. N. Nguyen, C.-V. Priporas, M. McPherson, S. Manyiwa, CSR-related consumer scepticism: A review of the literature and future research directions. *Journal of Business Research* **169**, 114294 (2023).

10. C. Valor, P. Antonetti, G. Zasuwa, Corporate social irresponsibility and consumer punishment: A systematic review and research agenda. *Journal of Business Research* **144**, 1218–1255 (2022).

11. C. M. Christenson, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail* (Harvard Business School Press, Boston, Massachusetts, USA, 1997).

12. S. J. H. Graham, C. Grim, T. Islam, A. C. Marco, J. Miranda, Business dynamics of innovating firms: Linking U.S. patents with administrative data on workers and firms. *Economics Manag Strategy* **27**, 372–402 (2018).

13. U. Akcigit, W. R. Kerr, Growth through heterogeneous innovations. *Journal of Political Economy* **126**, 1374–1443 (2018).

14. P. Aghion, U. Akcigit, P. Howitt, The Schumpeterian Growth Paradigm. *Annu. Rev. Econ.* **7**, 557–575 (2015).

15. R. Koning, S. Samila, J.-P. Ferguson, Who do we invent for? Patents by women focus more on women's health, but few women get to invent. *Science* **372**, 1345–1348 (2021).

16. J. Feng, X. Jaravel, E. Einio, Social Push and the Direction of Innovation. *Proceedings* **2021**, 10143 (2021).

17. G. Dossi, Race and Science. [Preprint] (2024).

18. A. Bell, R. Chetty, X. Jaravel, N. Petkova, J. Van Reenen, Who Becomes an Inventor in America? The Importance of Exposure to Innovation*. *The Quarterly Journal of Economics* **134**, 647–713 (2018).

19. M. W. Nielsen, J. P. Andersen, L. Schiebinger, J. W. Schneider, One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nat Hum Behav* **1**, 791–796 (2017).

20. A. Eizenberg, Upstream Innovation and Product Variety in the U.S. Home PC Market. *The Review of Economic Studies* **81**, 1003–1045 (2014).

21. D. Comin, A. Danieli, M. Mestieri, Income-driven Labor Market Polarization. NBER [Preprint] (2020). https://doi.org/10.3386/w27455.

22. C. Broda, D. E. Weinstein, Product Creation and Destruction: Evidence and Price Implications. *American Economic Review* **100**, 691–723 (2010).

23. V. Michelman, L. Msall, Sex, Drugs, and R&D: Missing Innovation from Regulating Female Enrollment in Clinical Trials. [Preprint] (2024).

24. C. M. Woods, B. Applebaum, Y. Green, D. L. Kallgren, E. Kappeler, Women's health: 30 years of progress in the U.S. Department of Health and Human Services. *Public Health Rep* **130**, 123–127 (2015).

25. R. B. Merkatz, S. W. Junod, Historical background of changes in FDA policy on the study and evaluation of drugs in women. *Academic Medicine* **69**, 703–7 (1994).

26. A.-C. Ramón, M. Tran, D. Hunt, "Hollywood Diversity Report 2023" (UCLA, 2023).

27. B. Gaut, Empathy and Identification in Cinema. *Midwest Studies In Philosophy* **34**, 136–157 (2010).

28. M. Kollar, Z. Scherer, "Income in the United States: 2024" (U.S. Census Bureau, 2025).

29. E. Garriga, D. Mele, Corporate Social Responsibility Theories: Mapping the Territory.

30. G. Aras-Beger, F. D. Taşkın, Corporate Social Responsibility (CSR) in Multinational Companies (MNCs), Small-to- Medium Enterprises (SMEs), and Small Businesses.

31. K. J. Arrow, "Economic Welfare and the Allocation of Resources for Invention" in *The Rate and Direction of Inventive Activity: Economic and Social Factors* (Princeton University Press, 1962).

32. K. A. Bryan, H. L. Williams, "Innovation: market failures and public policies" in *Handbook of Industrial Organization*, K. Ho, A. Hortaçsu, A. Lizzeri, Eds. (Elsevier, 2021; https://www.sciencedirect.com/science/article/pii/S1573448X21000133)vol. 5 of *Handbook of Industrial Organization*, pp. 281–388.

33. W. R. Kerr, R. Nanda, Democratizing entry: Banking deregulations, financing constraints, and entrepreneurship. *Journal of Financial Economics* **94**, 124–149 (2009).

34. C. Cunningham, F. Ederer, S. Ma, Killer acquisitions. *Journal of political economy* **129**, 649–702 (2021).

35. L. Klapper, L. Laeven, R. Rajan, Entry regulation as a barrier to entrepreneurship. *Journal of Financial Economics* **82**, 591–629 (2006).

36. B. H. Hall, A. B. Jaffe, M. Trajtenberg, "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools" (Working Paper 8498, National Bureau of Economic Research, 2001); https://doi.org/10.3386/w8498.

37. M. Ewens, M. Marx, Firm Age and Invention: An Open-Access Dataset.

38. K. K. F. Law, J. Tan, Diversity Tokenism. *J of Accounting Research*, 1475–679X.70019 (2025).

39. K. AlNuaimi, G. Marti, M. Ravaut, A. AlKetbi, A. Henschel, R. Jaradat, Enriching Datasets with Demographics through Large Language Models: What's in a Name? arXiv arXiv:2409.11491 [Preprint] (2024). https://doi.org/10.48550/arXiv.2409.11491.

40. S. Hercberg, M. Touvier, J. Salas-Salvado, The Nutri-Score nutrition label: A public health tool based on rigorous scientific evidence aiming to improve the nutritional status of the population. *International Journal for Vitamin and Nutrition Research* **92**, 147–157 (2022).

Acknowledgements: The work on this paper received financial support from Central European University, Cornell University, and the Fulbright Program. It benefited from comments from Cristina Corduneanu-Huci, Michael Dorsch, Mihaly Fazekas, Yian Yin and Shane M. Greenstein and from participants of the 2025 International Conference on Science of Science and Innovation. The authors declare no conflict of interest.

Supplementary materials

S1 Data construction

Our analysis relies on accurately combining firms incorporation data with domain-specific consumer characteristics. For firm information, we use OpenCorporates, a comprehensive source of administrative records on firms in the U.S. We consider products introduced within three years of company establishment as startups. For product characteristics, we use domain-specific large-scale datasets in retail, movies, and biomedical patents.

S1.1 Retail

In Retail, we start with the Nielsen Consumer Panel dataset, which continuously collects barcode-level purchase information from a representative sample of approximately 60 thousand U.S. households.[1]. The dataset is comprehensive for each household, but not necessarily for each product: we observe fewer than ten purchases for 51 percent of products. To establish product introduction date, we consider all purchases from 2004, the earliest date available. To reliably compare product introduction and firm incorporation, and to take advantage of a larger panel size, we only consider purchases of products introduced after 2007.

We derive consumer characteristics on the household level, on which we observe purchases. Race and ethnicity are self-reported by the household on the household level. For households of members with multiple racial and ethnic identities, we assume that the identity of the primary respondent for the household is being reported. For gender, which is reported on individual level, we count the number of male and female members.

To identify startups, we link product barcodes to GS1 data, which yields company name and address. We clear company name of punctuation, leading and trailing 'the', common legal suffixes, and retrieve ZIP code from the address, with which we match to OpenCorporates. We find ZIP codes more reliable than other location identifiers, as they are broadly available in both GS1 and OpenCorporates, avoid variations and abbreviations in geographical names, and appropriately provide higher granularity than city names for more dense areas.

We begin with exact match on name and ZIP code, which covers 44 percent of companies. On the remaining set, we perform fuzzy match based on a match score comprising of ZIP code distance, Jaro-Winkler name similarity, match of first letter in name, and match of U.S. State. We exclude matches to companies established after the company introduced its first product in the Nielsen dataset. We use OpenStreetMap API to geolocate ZIP codes for the distance calculation.

[1]More information at `https://www.chicagobooth.edu/research/kilts/research-data/nielseniq`, accessed October 31, 2025.

We measure quality using nutri-score provided by OpenFoodFacts,[2] which we match on barcode (24 percent match rate on food products). Nutri-score evaluates the nutritional quality of food based on their nutrient composition. The score is calculated by subtracting positive points—accrued from favorable components (fiber, protein, fruits, vegetables, nuts, etc.) from negative points (energy density, saturated fatty acids, simple sugars, and sodium) [2]. OpenFoodFacts collects composition of food products from open sources and user input, and calculates the nutri-score based on this information.

S1.2 Movies

In Movies, we start with The Movie Database (TMDB) dataset, which we query for information about producers, cast, genres, and release date. We restrict to 42 thousand U.S. non-documentary non-animated feature films. To identify startups, we match production companies to OpenCorporates by name, and consider the older production company in the case of co-productions.

We can only use produced company name, as address is unavailable in TMDB. At the same time, TMDB covers most movies produced in the United States, providing good coverage on its own. In our main results (panels A in tables S6–S9), we subset to exact name match to OpenCorporates on companies established prior to the earliest movie release, yielding match rate of 49 percent. We also report results without matching to OpenCorporates, where firms establishment is defined as the introduction of the first movie (panels B).

We identify the gender and race and ethnic minority identity of lead actors. Ideally, we would observe self-identified gender and race and ethnic identity. TMDB includes some information about lead actor gender crowd-sourced from user input. We only consider male and female leads, as the number of non-binary actors is insufficient for statistical analysis: there are only 32 movies non-binary lead character movies in our sample. The database does not include information about race or ethnic identity.

As the second best option, we use use a GPT4 mini 2024-07-18 to classify lead actors by race and ethnic minority identity and by gender when gender information is unavailable (29 percent of movies)—an approach that achieves high accuracy on public figures [3, 4]. We use this method on assumption that self-identified information about lead actors is often in public domain, and that common perceptions of gender, race and ethnic identity matter for audience choosing a movie [5, 6]. We cross-validate the results with a human reviewer with access to the internet including bibliographic information on the actors. In gender classification, they reviewed 25 actors labeled

[2] `https://world.openfoodfacts.org/data`, accessed October 31, 2025.

as female and male each, and concurred with all the labels.

In race and ethnicity, the Human reviewer cross-validated 25 actors labeled as white non-Hispanic or as identifying as race or ethnic minority. They concurred with all white non-Hispanic labels, but not with six actors labeled as race or ethnic minorities. Treating the assessment of the human coder as true values, this yields estimated precision of $19/26 = 0.76$ and, given that the estimated proportion of minority actors is 0.23, estimated accuracy $0.23 \times 0.76 + 0.77 \times 1 = 0.94$. This is comparable to other projects following a similar approach [3, 4].

We further link to the International Movie Database (IMDB) aggregate review by gender dataset, which we match on IMDB ID.[3] The dataset is restricted to movies with at least 100 male and female reviews, 34 percent match rate.

S1.3 Biomedical

In the biomedical domain, we start with patents from PatentsView. We establish male and female orientation based on data from the U.S. National Library of Medicine's Medical Text Indexer (MTI) data from [7]. Unfortunately, publicly available MTI has been discontinued, which prevents us from expanding the dataset. For incorporation date, we rely on match to OpenCorporates provided by [8], which makes the results more reproducible than doing our own matching.

To control for inventor gender, we add information from the PatentsView *Disambiguated Inventor* table, which relies on WIPO's *Worldwide Gender-Name Dictionary 2.0* to approximate gender of inventors. We aggregate information on the patent level based on female share and following [7], we code no female inventors as No_female, between 0 and 0.5 as $Minority_female$, between 0.5 (inclusive) and 1 as $Majority_female$, and 1 as All_female. We count unavailable gender as non-female. As a proxy for quality, we calculate 10 years of forward citations.

To distinguish between drugs and non-drug patents, we use NBER patent classes based on USPC main class, and code $drugs = 1$ for mainclass 424 and 514. [4]

As a proxy of quality, we calculate 10 year citations based on PatentsView data, and code natural log + 1 as $Citations(log)$. [TODO: review] For class fixed effects, we code CPC patent classes as $Class$. CPC classes have the advantage of being more granular than NBER classes, and less granular than NBER subcategories.

[3] Available at `https://github.com/sahildit/IMDB-Movies-Extensive-Dataset-Analysis`, accessed November 18, 2025

[4] Inconveniently, PatentsView lacks USPC classification for the years 2002–2004 per Release Notes, available at `https://patentsview.org/release-notes`. We supplement PatentsView with GoogleBigQuery to to circumvent this limitation.

S2 FDA policy shock in biomedical patents

Figure 3C in the main text estimates a link between under-representation in biomedical patents and orientation of startups toward female medical needs. This section elaborates on the empirical setup and robustness of the results.

S2.1 Female-oriented biomedical research and FDA policy change

Until late-1980s, female health had been systematically under-researched in the U.S. This came to focus in the landmark 1985 Report of the Public Health Service Task Force on Women's Health Issues, written by the first-ever U.S. Public Health Service Task Force on Women's Health Issues.[9] The report recommended increased research on women's health, and started a series of initiatives aimed ad towards goal. The 1986–87 NIH Guide for Grants and Contracts encouraged female participation in studies, and in 1991, the National Institutes of Health (NIH) launched set of clinical trials as part of the Women's Health Initiative. In 1993, the U.S. Congress wrote inclusion policy into the NIH Revitalization Act.[9–11]

A major obstacle to research on female health had been the exclusion of women of childbearing potential in early phases of clinical drug trials. This likely caused female medical needs to be under-researched compared to male medical needs. The Food and Drug Administration (FDA) revised its guidance in 1993, allowing female participation and requiring gender-specific analysis of safety and efficacy of new drugs. In estimates presented by [11] and confirmed in our analysis, this lead to a significant increase in female-oriented research post-1992. While the other policy changes affected biomedical research broadly, the FDA policy change only affected drugs.

This presents an attractive empirical setup to test the hypothesis that startups tend to invent more for underrepresented groups. Prior to the changes in the 1990s, female were more underrepresented in biomedical innovation. The reforms mainstreamed female-focus biomedical innovation, and the 1993 FDA reform in particular made it significantly easier to introduce female-oriented drugs. The question is if this mainstreaming of female-oriented drugs research decreased the relative female focus of startup patents compared to incumbent patents.

S2.2 Female under-representation and startups

The FDA policy change made it easier for all firms to include female in clinical trials. Directly, any increase in female-oriented biomedical research resulting from this policy change should affect both startups and incumbents similarly. This establishes our baseline hypothesis: incumbents and

startups would be similarly affected by the 1993 FDA policy change.

If it is the case that startups innovate more for underrepresented groups, the odds of startups inventing for female should decrease after the policy change. As an alternative hypothesis, we would therefore expect startups to have comparatively lower odds of female-oriented patenting in the post-1992 period.

We test this hypothesis using patent data in a difference-in-difference framework between startups (treatment) and incumbents (control). We further leverage the fact that the FDA reform affected only drugs, which we can leverage in a triple-difference setup [11].

S2.3 Empirical strategy

To test if startups were less likely to invent female drugs post-1992, we use a difference-in-difference framework:

$$\begin{aligned}\text{Logit}(Pr(Female_i = 1)) =& \beta_0 + \beta_1 \cdot Startup_i + \beta_2 \cdot Post1992_i \\ &+ \beta_3 \cdot Startup_i \cdot Post1992_i \\ &+ \gamma \cdot \hat{Team}_i + \delta \cdot \hat{Year}_i + \gamma \cdot \hat{Subcategory}_i + \varepsilon_i\end{aligned} \tag{1}$$

Where $Female_i = 1$ if a patent i is female-oriented, $Post1992_i = 1$ for patents filed in 1993 and later, $Startup_i = 1$ if the assignee was a startup at the time of filing. $\hat{Team}_i$ is a vector of inventor gender composition [7], $\hat{Subcategory}_i$ is a vector of NBER subcategory fixed effects, $\hat{Year}_i$ is a vector of filing year fixed effects, excluding the years 1989 and 1993 to avoid collinearity with the intercept and $Post1992_i$.

We estimate this specification on drugs patents with application filing date between 1991 and 2005 (12 years prior and after the policy change). We exclude patents filed between October 1992 (when a major conference after which the FDA policy change has been decided) and July 22, 1993 (when FDA announced the policy shift) [9, 10]. We further present results from a six year period prior and after the policy change.

We are primarily interested in the coefficient β_3, which estimates the decrease in log odds ratio of female-oriented patenting for startups after the 1992 reform; separated from the odds of female-oriented patenting among startups, post 1992, team gender composition, and year and class-fixed effects. Secondarily, we are also interested in the coefficient β_1, which represents the pre-1993 association between startups and female patents, and β_2, which represents the post-1992 increase in female-oriented patenting.

The identification rests on the assumption of parallel trends: startups and incumbents must

evolve in parallel absent the policy change. As it is impossible to evaluate this assumption directly, we provide standard indirect checks. In addition to the year fixed effects included in all specifications, in table S3, we interact our variables of interest with a linear time pre-trend, $Pre - trend$. While there is a statistically significant pre-trend in some specifications, this is positive, suggesting that any deviation in the main coefficient of interest due to a violation in pre-trend would be positive, counter to our hypothesis.

A theoretical reason for there to be non-parallel time-trends is that the FDA policy change occurred during a period of broader policy shift. Incumbent companies might have been incentivized to take-up on female-oriented research as this was mainstreamed by other funding and guidance changes such as those by the NIH. This dynamic would be in line with our hypothesis, but it would also conflate the changes due to the FDA policy change.

To isolate the effect of the FDA policy change, we compare the effects of the policy change on drugs differentiated other non-drug biomedical patents. This triple-difference approach adds a drug interactions to the regression specification and includes both drug and other medical patents:

$$\begin{aligned}\text{Logit}(Pr(Female_i = 1)) =& \beta_0 + \beta_1 \cdot Startup_i + \beta_2 \cdot Post1992_i + \beta_3 \cdot Drug_i \\ &+\beta_4 \cdot Startup_i \cdot Post1992_i + \beta_5 \cdot Startup_i \cdot Drug_i + \beta_6 \cdot Post1992_i \cdot Drug_i \\ &+\beta_7 \cdot Startup_i \cdot Post1992_i \cdot Drug_i \\ &+\gamma \cdot \hat{Team}_i + \delta \cdot \hat{Year}_i + \gamma \cdot \hat{Subcategory}_i + \varepsilon_i \end{aligned} \tag{2}$$

In addition to the variables outlined in the prior specification, we add interactions with $Drug_i = 1$ if patent i has main USPC class in NBER drugs patents. We are primarily interested in the coefficient β_7, which estimates the increase in log odds ratio of female-oriented drugs patents for startups after the 1992 reform. Secondarily, we are also interested in the coefficient β_5, which represents the pre-1993 association between startups and female patents, and β_6, which represents the post-1992 increase in female-oriented patenting.

In this setup, we can relax the assumption that startups and incumbents were exposed to similar trends in the female-orientation of their drugs patenting. This comes at the expense of assuming that drug and non-drug medical patents were exposed to similar trends, other than the post-1993 FDA policy change. We evaluate this assumption visually in figure 3C, and by adding linear time trend in table S3. Similarly to the time difference-in-difference results, there is a statistically significant time-startup-drug interaction, but it is positive, suggesting that any violation of parallel trends would bias our main results upward, counter to the hypothesis. Finally,

this model also assumes no spillovers between drug and medical research. As [11] note in a similar design, the existence of such spill-overs would bias the estimator downwards.

We further estimate a dynamic treatment effect specification, where we estimate differentiated changes for each year instead of the pooled post-1993 dummy:

$$\begin{aligned}\text{Logit}(Pr(Female_i = 1)) =& \beta_0 + \beta_1 \cdot Startup_i + \beta_2 \cdot Post1992_i + \beta_3 \cdot Drug_i \\ &+ \beta_4 \cdot Startup_i \cdot Post1992_i + \beta_5 \cdot Startup_i \cdot Drug_i + \beta_6 \cdot Post1992_i \cdot Drug_i \\ &+ \sum_{t=1981, t\neq 1992}^{2005} \Big[\eta_t \cdot \{Year_i = t\} + \lambda_t \cdot (Startup_i \cdot \{Year_i = t\}) \\ &+ \mu_t \cdot (Drug_i \cdot \{Year_i = t\}) + \delta_t \left(Startup_i \cdot Drug_i \cdot \{Year_i = t\}\right) \Big] \\ &+ \gamma \cdot \hat{Team}_i + \zeta \cdot \hat{Subcategory}_i + \varepsilon_i \end{aligned} \tag{3}$$

where $t_0 = 1992$ is omitted as the reference year. We are primarily interested in the vector of coefficients δ_t, which we report in figure 3B1, and in μ_t, which we report in figure 3B2.

S2.4 Results

As a baseline descriptive result, reported in Figure 1A of the main text, we split the sample of exclusively male and female patents by the year 1993. The results, reported in table , shows that the female association was positive before the policy change and neutral after the change.

Across the specifications, we consistently find that as female-oriented biomedical patenting increased following the FDA policy change, startups became significantly less likely to produce female-oriented drug patents than incumbents. The main triple difference estimates in table S2 shows an increase in odds of female-oriented patenting overall, but a similarly-sized relative decrease in startup patenting of $exp(-0.404) = 0.67$ or 34 percent among startups. This effect is consistent with the difference-in-differences results on drug patents (columns 1–2).

Descriptively, the coefficient on *Startup* (columns 1–2 on startup patents) shows that prior the policy change, startups were more likely than incumbents to file female-oriented patents. The relative decrease in the post-period all but erases this positive association in its size, showing that the positive association between startups and female-oriented patenting is concentrated to a period of greater under-representation.

The results are broadly robust to sub-setting to 6 years time period, and to the inclusion of startup-specific linear time trends. Table S4 shows that subletting to a six year period retains the

coefficient sizes, even if with a lower significance. Results with linear time pre-trend in table S3 show no negative linear pre-trend, and the main coefficient of interest is similar to or higher than the results without pre-trend. Some specifications show a positive pre-trend, the presence of which would bias our main results downwards. Figure 3C in the main text visualizes these temporal dynamics.

This shows strong evidence that pre-1993, startups were more likely to produce female-oriented patents than incumbents, and that the FDA reform closed this gap, mostly by mainstreaming female-oriented innovation among incumbents.

S3 Terciles regressions

Figure 1 of the main text shows the relationship between underrepresented groups and startups by visualizing coefficients from a series of regressions. Here, we specify these models and report full results in tables S5 (biomedical), S6, S7, S8, and S9 (movies), S10, S11, S12, and S13 (retail),

S3.1 Biomedical

In biomedical patents, we estimate a logistic regression:

$$\text{Logit}(female_i = 1)) = \alpha + \beta_1 \cdot Startup_i + \gamma \cdot \bar{Subcategory} \times Year_i + \varepsilon_i \tag{4}$$

where $Female_i$ indicates if patent i is female-oriented compared to the baseline of male, $Startup_i$ indicates if it has been introduced by a startup company, $\bar{subcategory} \times year_i$ is a vector of main patent NBER subcategory and year interactions [12]. β_1 is our primary coefficient of interest, estimating how much more likely are startups to introduce female-oriented patents, compared to the baseline of male-oriented patents.

Table S5 shows estimates of this specification on the full sample of patents (column 1), and on subsample split into approximately equal terciles by the proportion of female patents in each USPC patent subclass (columns 2–4), excluding patents in subclasses with fewer than ten patents.

S3.2 Movies

In movies, we estimate a logistic regression:

$$\text{Logit}(Pr(Minority_i = 1)) = \alpha + \beta_1 \cdot Startup_i + \gamma \cdot Year \bar{\times} Genre_i + \varepsilon_i \tag{5}$$

for each movie i, where $Minority_i$ is a binary variable equal to 1 if the main character is female (table S6) or race or ethnic minority (table S7), or if it attracted more female reviews than half of movies (table S8). $year \times genre_i$ is a vector of release year and first genre interactions.

Tables S6–S8 show estimates of this specification on the full sample of movies (column 1), and on subsamples split into approximately equal terciles by the proportion of minority movies in each relevant genre combination (columns 2–4), excluding movies in genres combinations with fewer than ten movies.

We further estimate the a negative binomial regression for the counts of ratings by gender:

$$\begin{aligned} Log(NRatings_{i,j}) = \alpha + \beta_1 \cdot Female_j + \\ \beta_2 \cdot Female_j \cdot Startup_i + \beta_3 \cdot Startup_i + \gamma \cdot \bar{Movie}_i + \varepsilon_i \end{aligned} \tag{6}$$

where we estimate the count of ratings for each movie separately for male and female reviewers, and include a vector of movie fixed effects, $\bar{movie}_i$. Table S9 shows estimates of this specification on the full sample of movies (column 1), and on subsamples split into approximately equal terciles by the proportion of female reviewers in each relevant genre combination (columns 2–4), excluding movies in genres combinations with fewer than ten movies. Throughout these results, we report startups restricted to exact name match with OpenCorporates, and startups defined by first release date in TMDB.

S3.3 Retail

For retail, we estimate a negative binomial regression:

$$\begin{aligned} Log(NPurchases_{i,j}) = \alpha + \beta_1 \cdot Minority_j + \\ \beta_2 \cdot Minority_j \cdot Startup_i + \beta_3 \cdot Startup_i + \gamma \cdot \bar{Product}_i + \varepsilon_{i,j} \end{aligned} \tag{7}$$

where $NPurchases_{i,j}$ is the count of purchases of product i, counted separately for majority ($j = 0$) and minority ($j = 1$) households, $Startup_i$ indicates if the product has been introduced by a startup company, $\bar{product}_i$ is a vector of product fixed effects. β_2 is our primary coefficient of interest, estimating by how much more purchases do startup product attract in a minority group compared to the majority group and incumbents.

Tables S10–S13 shows estimates of this specification on the full sample of movies (column 1), and on subsamples split into approximately equal terciles for Hispanic / Latino, Black, Asian, and other non-white households as the minority, with purchases by non-Hispanic or white households as the baseline. For each of those minority groups, we split the sample by low-, mid- and high-

merciless of minority representation in product categories (columns 2–4).

Table S14 shows analogical results for sum of male and female household members. These results are complicated by the fact that gender is reported for individuals, but purchases are aggregated at the household level. This conflates purchases driven by female and male consumers, and pushes the association towards the mean. Still, consistent with our previous findings, these results find a null overall association between startup products and female purchases, but strong and positive in the lowest third of product categories ($P < .01$), weaker and positive in mid-third of categories ($P.04$), and null in high-representation.

S4 Inventor and director gender interactions

We expand on results presented in [7] by estimating the association between female-oriented biomedical patenting and startup – inventor gender interaction:

$$\begin{aligned} \text{Logit}(Pr(Female_i = 1)) =& \beta_0 + \beta_1 \cdot Startup_i \\ &+ \gamma_2 \cdot \hat{Inventors}_i + \delta \cdot Startup_i \cdot \hat{Inventors}_i \\ &+ \theta \cdot Year \times \hat{Subcategory}_i + \delta\lambda Year \times \hat{TeamSize}_i + \varepsilon_i \end{aligned} \tag{8}$$

where $Female_i$ indicates if a patent is female-oriented, $\hat{Inventors}_i$ is a vector of values all female, majority female (more than half but not all), minority female (more than none, but less than half), and no female (baseline). Table S15 presents the results. Compared to [7], the sample is restricted startup data availability [13], but is consistent with their findings even on this restricted sample.

We also include female director and startup interaction in movies terciles regressions (equation 5):

$$\begin{aligned} \text{Logit}(Pr(Minority_i = 1)) = \alpha + \beta_1 \cdot Startup_i \times DirectorFemale_i + \\ \beta_2 \cdot Startup_i + \beta_1 \cdot DirectorFemale_i + \\ \gamma \cdot Year \times \hat{Genre}_i + \varepsilon_i \end{aligned} \tag{9}$$

where $DirectorFemale_i$ is 1 if the movies director is female, using the same methodology as actor gender identification. Tables S17 and S17 show the results, consistent with the main specification.

S5 Income

To investigate the interactions of income, we expand on our main aggregate specification (equation 7) by spiting the counts of product i by race or ethnic minority dummy j and further by low, mid and high income k:

$$\begin{aligned} Log(NPurchases_{i,j,k}) = \alpha + \beta_1 \cdot Minority_j + \\ \beta_2 \cdot Minority_j \cdot Startup_i + \beta_3 \cdot Income_k + \\ \beta_4 \cdot Minority_j \cdot Income_k + \beta_5 \cdot Startup_i \cdot Income_k + \\ \beta_6 \cdot Minority_j \cdot Startup_i \cdot Income_k + \gamma \cdot \bar{product}_i + \varepsilon_{i,j,k} \end{aligned} \tag{10}$$

Table S18 shows these results for households differentiated between low income (below 25 thousand USD per year, 13 percent of purchases), mid-income (baseline), and high-income (above 99 thousand, 19 percent).

The results clearly show that startups introduce products that receive 15 percent ($P < .01$) more purchases from high income households, and 3.4 percent ($P < .01$) fewer from low-income households, compared to incumbents. This association may explain the earlier relatively lower (but still positive) association between startups and Black households purchases, and relatively high positive association between startups and Asian household purchases: Black and Asian households have the lowest and highest average income by race, respectively (18). Consistent with this interpretation, controlling for income increases the race and ethnicity estimates, with the exception of Asian households. This shows that the interaction with income mostly moderates, rather than exacerbates, the startup association with minority purchases.

S6 Quality

To analyze quality, we regress startup female interaction on log 10 years patent citations (table S19), minority startup interaction on nutri-score in retail (table S20, where lower is better).

In those domains, startups produce significantly higher-quality new products: 28 percent more citations in patents ($P < .01$, table S17) and 0.98 points lower nutri-score ($P < .01$, table S21), consistent with prior literature [TODO: reference]. This higher quality of startup products persists in retail when we distinguish products popular among minority consumers and lowers but remains positive when we distinguish female-oriented patents. This represents clear evidence that along technical quality, startups produce higher quality products for minorities than incumbents.

References

[1] P. Moser, How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs, *American Economic Review* **95**, 1214–1236 (2005).

[2] S. Hercberg, M. Touvier, J. Salas-Salvado, The Nutri-Score nutrition label: A public health tool based on rigorous scientific evidence aiming to improve the nutritional status of the population, *International Journal for Vitamin and Nutrition Research* **92**, 147–157 (2022).

[3] K. K. F. Law, J. Tan, Diversity Tokenism, *Journal of Accounting Research* pp. 1475–679X.70019 (2025).

[4] K. AlNuaimi, *et al.*, Enriching Datasets with Demographics through Large Language Models: What's in a Name? (2024). ArXiv:2409.11491 [cs].

[5] A.-C. Ramón, M. Tran, D. Hunt, Hollywood Diversity Report 2023, *Tech. rep.*, UCLA (2023).

[6] B. Gaut, Empathy and Identification in Cinema, *Midwest Studies In Philosophy* **34**, 136–157 (2010).

[7] R. Koning, S. Samila, J.-P. Ferguson, Who do we invent for? Patents by women focus more on women's health, but few women get to invent, *Science* **372**, 1345–1348 (2021).

[8] M. Ewens, M. Marx, Firm Age and Invention: An Open-Access Dataset, *Tech. rep.* (2024).

[9] C. M. Woods, B. Applebaum, Y. Green, D. L. Kallgren, E. Kappeler, Women's health: 30 years of progress in the U.S. Department of Health and Human Services., *Public health reports (Washington, D.C. : 1974)* **130**, 123–127 (2015). Place: United States.

[10] R. B. Merkatz, S. W. Junod, Historical background of changes in FDA policy on the study and evaluation of drugs in women, *Academic Medicine* **69**, 703–7 (1994). Publisher: LWW.

[11] V. Michelman, L. Msall, Sex, Drugs, and R&D: Missing Innovation from Regulating Female Enrollment in Clinical Trials (2024).

[12] B. H. Hall, A. B. Jaffe, M. Trajtenberg, The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools, *Working Paper 8498*, National Bureau of Economic Research (2001). Series: Working Paper Series.

[13] M. Ewens, M. Marx, Entrepreneurial Patents, *Working Paper* (2023).

List of Tables

S1 Patents female v. male association, 1993 split . . . 14
S2 Main difference-in-differences regression results . . . 15
S3 Linear time pre-trend difference-in-differences regression results . . . 16
S4 Difference-in-differences regression results, six years subset . . . 17
S5 Female and startup association by tercile of representation by patent class . . . 18
S6 Female lead and startup association by tercile of representation by genres combination 19
S7 Minority lead and startup association by tercile of representation by genres combination . . . 20
S8 Female popularity and startup association by tercile of representation by genres combination . . . 21
S9 Female and male review counts and startup association by tercile of representation by genres combination . . . 22
S10 Hispanic household and startup association by tercile of representation by product category . . . 23
S11 Black household and startup association by tercile of representation by product category . . . 24
S12 Asian household startup association by tercile of representation by product category 25
S13 Other non-white and startup association by tercile of representation by product categorys . . . 26
S14 Count of female household members and startup association by tercile of representation by product categorys . . . 27
S15 Logistic regression of startup inventor team interactions . . . 28
S16 Female lead and startup association by tercile of representation by genres combination, director gender fixed effects . . . 29
S17 Female lead and startup association by tercile of representation by genres combination, director gender fixed effects . . . 30
S18 Retail regression interactions by low, mid and high income . . . 31
S19 Biomedical patents quality . . . 32
S20 Retail quality . . . 33
S21 Movies quality, female ratings . . . 34
S22 Movies quality, female lead . . . 35
S23 Movies quality, minority lead . . . 36
S24 Retail geography interaction . . . 37

Dependent Variable:	Female		
Sample	Full sample	Pre1993	Post1992
Model:	(1)	(2)	(3)
Variables			
Startup	0.098	0.499**	0.032
	(0.086)	(0.223)	(0.092)
Fixed-effects			
Inventor gender	Yes	Yes	Yes
NBER subcategory-Year	Yes	Yes	Yes
NBER subcategory	Yes	Yes	Yes
Fit statistics			
Observations	9,051	1,722	7,329
Squared Correlation	0.02286	0.02312	0.01154
Pseudo R^2	0.01698	0.01758	0.00864
BIC	13,133.1	2,491.4	10,624.4

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S1: Patents female v. male association, 1993 split
Logistic regression estimates.

Dependent Variable:	Female			
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.612***	0.596***	0.394***	0.315***
	(0.152)	(0.153)	(0.076)	(0.076)
Post1992	0.731***	0.549***	0.264***	0.145
	(0.042)	(0.159)	(0.035)	(0.099)
Startup × Post1992	-0.505***	-0.497***	-0.093	-0.083
	(0.163)	(0.164)	(0.083)	(0.084)
Drugs			-0.415***	-0.058
			(0.050)	(0.078)
Post1992 × Drugs			0.456***	0.447***
			(0.054)	(0.054)
Startup × Drugs			0.224	0.293*
			(0.170)	(0.171)
Startup × Post1992 × Drugs			-0.408**	-0.404**
			(0.183)	(0.184)
Fixed-effects				
Inventor gender	Yes	Yes	Yes	Yes
Year		Yes		Yes
NBER subcategory		Yes		Yes
Fit statistics				
Observations	43,746	43,746	111,961	111,961
Squared Correlation	0.00851	0.00992	0.00496	0.01271
Pseudo R^2	0.01231	0.01434	0.00688	0.01738
BIC	32,070.6	32,282.7	85,512.0	84,911.3

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S2: Main difference-in-differences regression results
Main difference-in-differences (columns 1–3, subset of drugs patents) and triple-difference (columns 4–5, all patents) results, logistic regression with female patents binary as the outcome. Patents from 1981 to 2005, October 1992 to July 22, 1993 omitted. Incumbent pre-1993 (non-drug in columns 4–5) patents as a baseline. Years 1992 and 1993 omitted in fixed effects.

Dependent Variable:	Female			
Model:	(1)	(2)	(3)	(4)
Variables				
Startup × Post1992	-0.497***	-0.829***		
	(0.164)	(0.251)		
Pre-trend × Startup		0.029*		-0.023**
		(0.017)		(0.010)
Startup × Post1992 × Drugs			-0.404**	-1.02***
			(0.184)	(0.285)
Pre-trend × Drugs				-0.001
				(0.006)
Pre-trend × Startup × Drugs				0.052***
				(0.019)
Fixed-effects				
Startup	Yes	Yes	Yes	Yes
Post1992	Yes	Yes	Yes	Yes
Inventor gender	Yes	Yes	Yes	Yes
NBER subcategory	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
Drugs			Yes	Yes
Startup × Post1992			Yes	Yes
Post1992 × Drugs			Yes	Yes
Startup × Drugs			Yes	Yes
Fit statistics				
Observations	43,746	43,746	111,961	111,961
Squared Correlation	0.00992	0.01001	0.01271	0.01282
Pseudo R^2	0.01434	0.01443	0.01738	0.01749
BIC	32,282.7	32,290.6	84,911.3	84,936.8

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S3: Linear time pre-trend difference-in-differences regression results
Difference-in-differences (columns 1–2, subset of drugs patents) and triple-difference (columns 3–4, all patents) results, logistic regression with female patents binary as the outcome. Patents from 1981 to 2005, October 1992 to July 22, 1993 omitted. Incumbent pre-1993 (non-drug in columns 3–4) patents as a baseline. Years 1992 and 1993 omitted in fixed effects. Linear time pre-trend interactions.

Dependent Variable:	Female			
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.355*	0.353*	0.303***	0.211**
	(0.210)	(0.211)	(0.100)	(0.100)
Post1992	0.598***	0.545***	0.148***	0.144
	(0.059)	(0.159)	(0.047)	(0.101)
Startup × Post1992	-0.397*	-0.387*	0.061	0.069
	(0.234)	(0.235)	(0.115)	(0.115)
Drugs			-0.415***	-0.074
			(0.067)	(0.114)
Post1992 × Drugs			0.443***	0.421***
			(0.075)	(0.076)
Startup × Drugs			0.062	0.151
			(0.232)	(0.234)
Startup × Post1992 × Drugs			-0.456*	-0.455*
			(0.260)	(0.261)
Fixed-effects				
Inventor gender	Yes	Yes	Yes	Yes
Year		Yes		Yes
NBER subcategory		Yes		Yes
Fit statistics				
Observations	19,288	19,288	46,925	46,925
Squared Correlation	0.00749	0.00825	0.00496	0.01227
Pseudo R^2	0.01090	0.01209	0.00688	0.01711
BIC	13,407.2	13,509.6	34,447.9	34,223.2

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S4: Difference-in-differences regression results, six years subset
Difference-in-differences (columns 1–2, subset of drugs patents) and triple-difference (columns 3–4, all patents) results, logistic regression with female patents binary as the outcome. Patents from 1987 to 1998, October 1992 to July 22, 1993 omitted. Incumbent pre-1993 (non-drug in columns 3–4) patents as a baseline. Years 1992 and 1993 omitted in fixed effects. Linear time pre-trend interactions.

Dependent Variable:	Female patent			
Class tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.063	0.148*	-0.050	0.050
	(0.042)	(0.078)	(0.070)	(0.091)
Fixed-effects				
Inventors gender	Yes	Yes	Yes	Yes
Year-Subcategory	Yes	Yes	Yes	Yes
Fit statistics				
Observations	23,083	8,271	7,586	7,086
Squared Correlation	0.01928	0.02189	0.02208	0.03229
Pseudo R^2	0.01418	0.01861	0.01643	0.02672
BIC	32,956.2	10,881.9	11,465.6	9,264.0

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S5: Female and startup association by tercile of representation by patent class
Logistic regression estimates of startup–female association, biomedical patents 1976–2015, subset of exclusively male- or female-oriented patents.

A) Dependent Variable:	Female Lead			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	-0.075**	0.249***	-0.082	-0.161**
	(0.038)	(0.083)	(0.063)	(0.066)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	38,532	11,815	12,388	12,824
Squared Correlation	0.10458	0.05743	0.05107	0.09160
Pseudo R^2	0.08393	0.05524	0.04045	0.07170
BIC	56,598.7	16,449.2	21,334.5	20,954.0

B) Dependent Variable:	Female Lead			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
StartupTMDB	-0.090***	0.122***	0.0007	-0.259***
	(0.021)	(0.045)	(0.037)	(0.034)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	70,006	22,187	22,853	23,279
Squared Correlation	0.09630	0.04290	0.04986	0.07649
Pseudo R^2	0.07758	0.04237	0.04041	0.05974
BIC	95,084.1	27,735.9	35,351.0	36,066.4

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S6: Female lead and startup association by tercile of representation by genres combination

A) Dependent Variable:	Minority Lead			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.333***	0.430***	0.196**	0.269***
	(0.045)	(0.109)	(0.082)	(0.069)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	34,784	9,490	11,185	11,679
Squared Correlation	0.06795	0.07699	0.06682	0.07902
Pseudo R^2	0.06649	0.08247	0.06368	0.06729
BIC	42,598.4	12,299.8	14,974.0	18,131.4

B) Dependent Variable:	Minority Lead			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
StartupTMDB	0.450***	0.433***	0.379***	0.396***
	(0.024)	(0.056)	(0.041)	(0.039)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	55,372	16,232	17,962	18,575
Squared Correlation	0.07168	0.06890	0.05968	0.07634
Pseudo R^2	0.06648	0.07376	0.05698	0.06129
BIC	67,344.1	19,649.1	23,558.4	27,443.6

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S7: Minority lead and startup association by tercile of representation by genres combination

A) Dependent Variable:	Female reviews > median			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.007	0.585***	-0.162	-0.201
	(0.071)	(0.150)	(0.112)	(0.132)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	12,858	3,952	4,274	4,018
Squared Correlation	0.17199	0.09305	0.09336	0.04909
Pseudo R^2	0.13712	0.09046	0.07068	0.04459
BIC	18,708.4	5,583.7	6,966.5	5,510.0

B) Dependent Variable:	Female reviews > median			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
StartupTMDB	-0.049	0.147*	-0.005	-0.285***
	(0.040)	(0.081)	(0.066)	(0.075)
Fixed-effects				
Year-Genre	Yes	Yes	Yes	Yes
Fit statistics				
Observations	18,727	5,936	6,163	6,015
Squared Correlation	0.16117	0.08732	0.08095	0.05883
Pseudo R^2	0.12816	0.08356	0.06108	0.05363
BIC	26,664.6	8,050.3	9,972.9	8,099.0

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S8: Female popularity and startup association by tercile of representation by genres combination

A) Dependent Variable:	N Reviews			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Female	-1.44***	-1.90***	-1.42***	-0.967***
	(0.007)	(0.008)	(0.010)	(0.012)
Startup × Female	0.020	0.106***	-0.037	-0.078*
	(0.022)	(0.033)	(0.032)	(0.045)
Fixed-effects				
Movie	Yes	Yes	Yes	Yes
Fit statistics				
Observations	29,522	9,841	9,841	9,840
Squared Correlation	0.95672	0.97989	0.98611	0.94872
Pseudo R^2	0.23092	0.26154	0.26364	0.22887
BIC	581,720.6	186,466.9	182,149.6	191,750.1
Over-dispersion	9.2682	17.483	15.354	10.907

B) Dependent Variable:	N Reviews			
Genre tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Female	-1.43***	-1.89***	-1.39***	-1.01***
	(0.006)	(0.008)	(0.009)	(0.010)
StartupTMDB × Female	0.024*	0.095***	0.027	-0.053**
	(0.013)	(0.018)	(0.019)	(0.025)
Fixed-effects				
Movie	Yes	Yes	Yes	Yes
Fit statistics				
Observations	41,148	13,716	13,716	13,716
Squared Correlation	0.95775	0.97996	0.98441	0.94835
Pseudo R^2	0.25103	0.27722	0.28723	0.24757
BIC	771,110.9	253,024.0	239,321.0	253,190.3
Over-dispersion	9.4703	17.637	16.331	10.507

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S9: Female and male review counts and startup association by tercile of representation by genres combination

Dependent Variable:	N Purchases			
Category tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Hispanic	-2.71***	-2.89***	-2.68***	-2.46***
	(0.0010)	(0.001)	(0.002)	(0.002)
Startup × Hispanic	0.138***	0.178***	0.090***	0.090***
	(0.004)	(0.007)	(0.006)	(0.007)
Fixed-effects				
Product	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,832,412	1,277,471	1,277,471	1,277,470
Squared Correlation	0.98999	0.99342	0.99806	0.99461
Pseudo R^2	0.43933	0.40819	0.46665	0.45825
BIC	45,452,895.6	15,363,702.3	13,718,427.8	14,153,750.6
Over-dispersion	22.053	20.137	29.019	27.289

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S10: Hispanic household and startup association by tercile of representation by product category

Dependent Variable:	N Purchases			
Category tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Black	-2.27***	-2.68***	-2.24***	-1.87***
	(0.001)	(0.002)	(0.002)	(0.002)
Startup × Black	0.042***	0.087***	0.009	0.056***
	(0.004)	(0.008)	(0.006)	(0.007)
Fixed-effects				
Product	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,760,044	1,257,222	1,249,387	1,253,434
Squared Correlation	0.97736	0.98599	0.99154	0.98980
Pseudo R^2	0.40134	0.40127	0.43868	0.39516
BIC	45,846,536.4	14,847,215.1	13,955,097.7	14,685,341.2
Over-dispersion	9.2432	12.205	15.453	8.8889

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S11: Black household and startup association by tercile of representation by product category

Dependent Variable:	N Purchases			
Category tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Asian	-3.47***	-3.76***	-3.42***	-3.08***
	(0.001)	(0.002)	(0.002)	(0.002)
Startup × Asian	0.344***	0.355***	0.284***	0.113***
	(0.006)	(0.012)	(0.011)	(0.008)
Fixed-effects				
Product	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,704,146	1,245,337	1,236,443	1,222,366
Squared Correlation	0.95389	0.98687	0.95441	0.97329
Pseudo R^2	0.44380	0.43503	0.44917	0.46506
BIC	42,640,066.0	14,011,659.7	13,448,798.8	12,990,800.8
Over-dispersion	18.935	25.430	25.622	17.933

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S12: Asian household startup association by tercile of representation by product category

Dependent Variable:	N Purchases			
Category tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
Other	-2.87***	-3.03***	-2.82***	-2.60***
	(0.0010)	(0.001)	(0.002)	(0.002)
Startup × Other	0.165***	0.197***	0.115***	0.036***
	(0.004)	(0.007)	(0.008)	(0.007)
Fixed-effects				
Product	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,724,320	1,254,634	1,244,331	1,225,355
Squared Correlation	0.99070	0.99663	0.99479	0.99688
Pseudo R^2	0.45612	0.42591	0.47803	0.47894
BIC	43,227,763.2	14,822,388.3	13,379,560.3	12,872,473.8
Over-dispersion	36.460	35.594	50.143	46.410

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S13: Other non-white and startup association by tercile of representation by product categorys

Dependent Variable:	N Purchases			
Category tercile	Full sample	Low tercile	Middle tercile	High tercile
Model:	(1)	(2)	(3)	(4)
Variables				
FemaleN	0.109***	0.057***	0.116***	0.192***
	(0.0002)	(0.0003)	(0.0003)	(0.0004)
Startup × FemaleN	0.0007	0.009***	0.003**	0.002
	(0.0009)	(0.001)	(0.001)	(0.002)
Fixed-effects				
Product	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,832,412	1,277,471	1,277,471	1,277,470
Squared Correlation	0.99968	0.99983	0.99991	0.99947
Pseudo R^2	0.50585	0.49505	0.50810	0.51529
BIC	49,114,801.8	16,049,081.6	16,037,497.9	14,777,557.4
Over-dispersion	526.49	470.61	997.50	848.39

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S14: Count of female household members and startup association by tercile of representation by product categorys

Dependent Variable:	Female			
Model:	(1)	(2)	(3)	(4)
Variables				
Startup	0.224***		0.230***	0.190***
	(0.024)		(0.024)	(0.027)
Inventors all female		0.336***	0.338***	0.295***
		(0.045)	(0.045)	(0.049)
Inventors majority female		0.225***	0.231***	0.216***
		(0.027)	(0.027)	(0.029)
Inventors minority female		0.089***	0.094***	0.079***
		(0.025)	(0.025)	(0.025)
Startup × Inventors all female				0.331**
				(0.130)
Startup × Inventors majority female				0.147
				(0.090)
Startup × Inventors minority female				0.169**
				(0.075)
Fixed-effects				
NBER subcategory-Year	Yes	Yes	Yes	Yes
Team Size-Year	Yes	Yes	Yes	Yes
Fit statistics				
Observations	155,873	155,873	155,873	155,873
Squared Correlation	0.01636	0.01653	0.01721	0.01732
Pseudo R^2	0.02236	0.02261	0.02336	0.02346
BIC	124,904.4	124,898.3	124,819.5	124,843.1

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S15: Logistic regression of startup inventor team interactions

Dependent Variable:	Female Lead							
Genre tercile	Full sample		Low tercile		Middle tercile		High tercile	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Startup	-0.075**	-0.070*	0.202**	0.213**	-0.064	-0.089	-0.151**	-0.125*
	(0.038)	(0.041)	(0.083)	(0.089)	(0.064)	(0.068)	(0.066)	(0.070)
Female director		1.35***		1.76***		1.30***		1.12***
		(0.050)		(0.112)		(0.081)		(0.080)
Startup × Female director		-0.140		-0.155		0.199		-0.345
		(0.157)		(0.349)		(0.249)		(0.254)
Fixed-effects								
Year-Genre	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	38,532	37,357	11,804	11,151	12,405	11,939	12,820	12,772
Squared Correlation	0.10458	0.12968	0.05669	0.09282	0.05264	0.08024	0.09115	0.11014
Pseudo R^2	0.08393	0.10272	0.05457	0.08185	0.04185	0.06188	0.07134	0.08607
BIC	56,598.7	54,410.0	16,493.3	15,579.0	21,347.4	20,462.8	20,924.1	20,618.8

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S16: Female lead and startup association by tercile of representation by genres combination, director gender fixed effects

Dependent Variable:	Female reviews > median							
Genre tercile	Full sample		Low tercile		Middle tercile		High tercile	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Startup	0.007	0.033	0.587***	0.634***	-0.166	-0.151	-0.207	-0.176
	(0.071)	(0.075)	(0.150)	(0.154)	(0.113)	(0.118)	(0.130)	(0.140)
Female director		1.36***		1.37***		1.10***		1.35***
		(0.092)		(0.208)		(0.149)		(0.172)
Startup × Female director		-0.495*		-0.742		-0.274		-0.629
		(0.266)		(0.771)		(0.438)		(0.412)
Fixed-effects								
Year-Genre	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	12,858	12,841	3,952	3,947	4,274	4,268	4,018	4,012
Squared Correlation	0.17199	0.19110	0.09375	0.10774	0.09397	0.10886	0.04846	0.06793
Pseudo R^2	0.13712	0.15399	0.09087	0.10265	0.07132	0.08337	0.04408	0.06438
BIC	18,708.4	18,406.2	5,596.1	5,559.4	6,959.8	6,897.3	5,545.1	5,463.6

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S17: Female lead and startup association by tercile of representation by genres combination, director gender fixed effects

Dependent Variables:	N Purchases	N Hispanic		N Black		N Asian		N Other non-white	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Startup × High income	0.142***		0.146***		0.133***		0.149***		0.144***
	(0.003)		(0.003)		(0.003)		(0.003)		(0.003)
Startup × Low income	-0.035***		-0.034***		-0.011***		-0.032***		-0.028***
	(0.003)		(0.003)		(0.004)		(0.004)		(0.004)
Minority group		-2.71***		-2.27***		-3.47***		-2.87***	
		(0.0010)		(0.001)		(0.001)		(0.0010)	
Startup × Minority group		0.138***	0.139***	0.042***	0.055***	0.344***	0.317***	0.165***	0.171***
		(0.004)	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)
Startup × Minority group × High income			-0.054***		0.019***		-0.048***		-0.051***
			(0.007)		(0.007)		(0.008)		(0.008)
Startup × Minority group × Low income			-0.056***		-0.105***		0.092***		-0.022**
			(0.009)		(0.008)		(0.015)		(0.010)
Fixed-effects									
High income	Yes		Yes		Yes		Yes		Yes
Low income	Yes		Yes		Yes		Yes		Yes
Product	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Minority group			Yes		Yes		Yes		Yes
High income × Minority group			Yes		Yes		Yes		Yes
Low income × Minority group			Yes		Yes		Yes		Yes
Fit statistics									
Observations	5,748,618	3,832,412	11,497,236	3,760,044	11,280,132	3,704,146	11,112,438	3,724,320	11,172,960
Squared Correlation	0.99740	0.98999	0.98430	0.97736	0.97451	0.95389	0.94540	0.99070	0.98450
Pseudo R^2	0.38044	0.43933	0.38722	0.40134	0.36883	0.44380	0.39661	0.45612	0.39207
BIC	55,824,900.5	45,452,895.6	67,603,518.3	45,846,536.4	67,790,498.6	42,640,066.0	62,080,887.0	43,227,763.2	64,224,284.5
Over-dispersion	7.2969	22.053	4.2569	9.2432	3.2664	18.935	4.0619	36.460	4.4921

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S18: Retail regression interactions by low, mid and high income

Dependent Variable:	Citations, 10y log	
Model:	(1)	(2)
Variables		
Female	-0.029	-0.013
	(0.021)	(0.018)
Startup	0.585***	0.288***
	(0.041)	(0.036)
Female × Startup	-0.144**	-0.117**
	(0.059)	(0.051)
Fixed-effects		
Inventor gender	Yes	Yes
Year	Yes	Yes
CPC subclass		Yes
Fit statistics		
Observations	18,570	18,570
R^2	0.06539	0.30874
Within R^2	0.01670	0.00471

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S19: Biomedical patents quality

Dependent Variable:	Nutri-Score	
Model:	(1)	(2)
Variables		
Minority	-0.426***	-0.002
	(0.041)	(0.025)
Startup	-1.42***	-0.984***
	(0.118)	(0.070)
Minority × Startup	-0.210	0.025
	(0.153)	(0.090)
Fixed-effects		
Year	Yes	Yes
Module		Yes
Fit statistics		
Observations	206,458	206,458
R^2	0.00522	0.65775
Within R^2	0.00270	0.00226

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S20: Retail quality
(The lower the better.)

Dependent Variables:	Ratings mean	Female ratings mean	Male ratings mean	Ratings mean	Female ratings mean	Male ratings mean
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Startup	-0.419***	-0.612***	-0.614***	-0.314***	-0.457***	-0.484***
	(0.058)	(0.073)	(0.069)	(0.047)	(0.061)	(0.053)
Female reviews > median	0.408***	0.537***	0.246***	0.227***	0.390***	0.037
	(0.030)	(0.035)	(0.030)	(0.023)	(0.024)	(0.025)
Startup × Female reviews > median	0.251***	0.261**	0.203**	0.215***	0.193**	0.153**
	(0.087)	(0.101)	(0.084)	(0.071)	(0.082)	(0.074)
Fixed-effects						
release_year	Yes	Yes	Yes	Yes	Yes	Yes
Genres (vec.)				Yes	Yes	Yes
Fit statistics						
Observations	14,785	14,785	14,785	14,785	14,785	14,785
R^2	0.04821	0.07180	0.04441	0.32257	0.32401	0.33015
Within R^2	0.03969	0.05633	0.02372	0.01553	0.03203	0.01175

Clustered (release_year) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S21: Movies quality, female ratings

Dependent Variables:	Ratings mean	Female ratings mean	Male ratings mean	Ratings mean	Female ratings mean	Male ratings mean
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Startup	-0.343***	-0.484***	-0.501***	-0.262***	-0.369***	-0.405***
	(0.032)	(0.033)	(0.034)	(0.030)	(0.029)	(0.030)
Female Lead	-0.168***	-0.120***	-0.231***	-0.185***	-0.125***	-0.234***
	(0.018)	(0.019)	(0.017)	(0.015)	(0.018)	(0.018)
Startup × Female Lead	0.110*	0.069	0.098	0.123**	0.065	0.100*
	(0.056)	(0.070)	(0.059)	(0.050)	(0.064)	(0.053)
Fixed-effects						
release_year	Yes	Yes	Yes	Yes	Yes	Yes
Genres (vec.)				Yes	Yes	Yes
Fit statistics						
Observations	22,145	22,141	22,145	22,145	22,141	22,145
R^2	0.05517	0.06963	0.09846	0.30447	0.29140	0.33104
Within R^2	0.01088	0.01284	0.02028	0.01072	0.01002	0.01926

Clustered (release_year) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S22: Movies quality, female lead

Dependent Variable:	Ratings mean	
Model:	(1)	(2)
Variables		
Startup	-0.342***	-0.257***
	(0.032)	(0.026)
Minority Lead	-0.087***	-0.082***
	(0.027)	(0.024)
Startup × Minority Lead	0.159*	0.143**
	(0.087)	(0.067)
Fixed-effects		
release_year	Yes	Yes
Genres (vec.)		Yes
Fit statistics		
Observations	21,534	21,534
R^2	0.05136	0.29832
Within R^2	0.00735	0.00549

Clustered (release_year) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S23: Movies quality, minority lead

Dependent Variable:	Black		Asian		Other non-white		Hispanic	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Minority	-2.27***	-2.21***	-3.47***	-3.24***	-2.87***	-2.82***	-2.71***	-2.76***
	(0.001)	(0.005)	(0.001)	(0.007)	(0.0010)	(0.005)	(0.0010)	(0.005)
Startup × Minority	0.042***	0.028***	0.344***	0.309***	0.165***	0.156***	0.138***	0.149***
	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)
Minority × City products (log)		-0.007***		-0.026***		-0.006***		0.006***
		(0.0006)		(0.0008)		(0.0005)		(0.0005)
Fixed-effects								
Product	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	3,760,044	3,704,768	3,704,146	3,649,666	3,724,320	3,669,468	3,832,412	3,775,534
Squared Correlation	0.97736	0.97705	0.95389	0.95202	0.99070	0.99056	0.98999	0.99005
Pseudo R^2	0.40134	0.40113	0.44380	0.44423	0.45612	0.45635	0.43933	0.43975
BIC	45,846,536.4	45,181,838.8	42,640,066.0	41,996,203.4	43,227,763.2	42,581,508.6	45,452,895.6	44,762,455.0
Over-dispersion	9.2432	9.2485	18.935	19.314	36.460	36.802	22.053	22.369

Clustered (Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table S24: Retail geography interaction