Commercializing Contrarian Ideas: Evidence From AI Contests*

Luca Gius
MIT Sloan

March 1, 2025

Abstract

Most contrarian entrepreneurs – those who challenge conventional "best practices" – struggle to attract resources, yet a few achieve outsized successes once their ideas gain traction. One explanation is that contrarian ideas are simply riskier: they fail more often, but the few that overcome deeper technical hurdles generate exceptional breakthroughs. Another explanation emphasizes commercialization barriers: investors avoid unproven alternatives but rush in once it becomes clear that a contrarian approach outperforms the status quo. To provide evidence on these views, I examine hundreds of AI contests where either a contrarian or mainstream method is revealed to be best. Using a difference-in-differences design that compares the likelihood of founding a startup between winners and close runner-ups, I find that contrarian victors benefit disproportionately from a public win – suggesting ex-ante validation is crucial for commercialization. Importantly, even narrow victories yield large gains for contrarian contestants because contests do more than reveal intrinsic quality – they trigger herding among investors, who disproportionately bid for the newly validated approach. As such, skewed returns for contrarians, in part, reflect that investors support too few heterodox ideas and strongly reward successful contrarians. Finally, I show that mainstream researchers adopt these validated contrarian methods only after more decisive proof, allowing contrarian entrepreneurs to secure resources without facing immediate competition.

*Correspondence to lgius@mit.edu. Website: lucagius.com. I am especially grateful to Pierre Azoulay, Fiona Murray and Scott Stern for their invaluable feedback and advice. I thank Daniel Keum, Soomi Kim, Danielle Li, Christina Nguyen, Georg Rilinger, Ezra Zuckerman and various seminar participants for helpful comments and suggestions. All errors remain my own.

1 Introduction

In 1903, the Wright Brothers managed the first piloted flight, but struggled for five years to convince funders they had surpassed Samuel Langley's heavily sponsored, mainstream approach. Only dramatic public demonstrations in 1908 secured enough resources for commercialization (Howard [1998]; Crouch [2003]). Their story illustrates two stylized facts about contrarian entrepreneurs: defying conventional wisdom can result in remarkable success (Schumpeter [1934]; Thiel and Masters [2014]), yet contrarians must jump through extra hoops – compared to mainstream rivals – in order to attract resources for commercialization (Stevenson [1983]).

Why do contrarians struggle to attract resources, despite the fact that a surprising number of the most groundbreaking and remunerative ideas in history started with a contrarian insight? One possibility is that contrarian approaches are simply higher-risk [Fleming, 2001; Cabral, 2003], so investors support only the few, high-quality contrarian ideas where most of the returns are concentrated. The contrarians' struggle to attract resources (and skewed returns when funded) is a "feature" of rational capital markets screening out cranky ideas.

A complementary explanation is that contrarian ideas face commercialization barriers: because they challenge widely held assumptions, contrarian approaches (i) are ex-ante believe to be wrong (or worse than the mainstream approach) (ii) if proven right, generate disproportionate knowledge spillovers that first-time investors cannot fully appropriate (e.g., other investors can support similar ideas). As such, investors have reduced incentives to shoulder the risk and "go first," and only invest *after* a method has been publicly validated. But opportunities to publicly demonstrate one's idea are scarce, especially if entrepreneurs do not possess the necessary resources to implement it to begin with. This means that investors (i) generally underinvest in contrarian ideas but (ii) fiercely bid for those (few) contrarian ideas that were publicly validated [Terviö, 2009]. Since credible, public demonstrations are scarce, in this scenario many contrarians do not get a "fair shot" in the marketplace [Rosenberg, 1992], and the disproportionate value captured by the few, "lucky" contrarians reflect rents and not (only) their vision.

Distinguishing these scenarios is crucial to assessing whether resource scarcity for contrarian ideas is ultimately efficient or reflects avoidable barriers. Yet three empirical hurdles complicate such an investigation. First, it is difficult to define "contrarian" in a comparable way: does an idea truly challenge the status quo, is it novel yet aligned with prevailing practices, or does it address a fundamentally different question? Second, many contrarian ideas remain "left untried," so we primarily see ex-post accounts of rare, high-profile successes. Studying the commercialization gap requires observing a broader range of approaches at risk of being funded. Finally, to learn whether investors systematically "overreact" to public validation events, we need to compare similar contrarian ideas – those that just barely fail versus those that just barely succeed in the demonstration – and see how a razor-thin margin shifts commercial outcomes.

The field of empirical AI research is well-suited to meeting these challenges. First, AI methods fall into distinct "families" of algorithms (e.g., support vector machines vs. neural networks), each trying to solve the same data-analytics tasks. This clean partition allows us to label whichever family is currently best-performing as mainstream and treat others as contrarian. Second, AI is a prime example of "use-inspired" basic research [Stokes, 1997], where researchers often publish methods and then found startups around them—letting us track all ap-

proaches (successful or not) that could plausibly be commercialized. Finally, AI relies on public, head-to-head contests that compare every algorithm family on the same task under standardized conditions. At any point, exactly one family holds the "state-of-the-art" title, while the others lag behind. These competitions serve as "moments of clarity" for investors, objectively revealing which approach performs better. Critically, they also let us distinguish between clear-cut wins (in which a contrarian approach decisively outperforms the mainstream) and razor-thin races—thereby testing whether even a marginal contrarian victory triggers a disproportionate investor response.

I collect data on 12,662 researchers participating in 983 evaluation challenges, some of which are time-bound contests, while others allow for continuous, ongoing evaluation. For each participant, I collect information on the papers describing their challenge submissions (along with their forward citations) and on any startup they founded.

First, I show that researchers who perform best in an evaluation challenge gain significant academic recognition and are disproportionately likely to start a company afterwards. Compared to the runner-up, the winning result in a challenge receives 121% more academic citations and 61% more patent citations, and by year four after the contest, the winner is 58% more likely to have founded a startup.

Next, I distinguish between contrarian and mainstream participants based on whether their algorithm's "family" is already the state-of-the-art prior to the contest. In practice, mainstream participants rely on the same family that generated the best-known result so far, while contrarians use methods that previously lagged behind the mainstream.[1]

Turning to the core question – how public demonstrations affect contrarian commercialization – I find that contrarian winners benefit far more from a contest victory than mainstream winners. Winning not only makes contrarian winners more likely to start a company, but that effect is 7 times higher than for mainstream winners. I find strong evidence in support of the notion that contrarian approaches need validation because they are inherently more uncertain. The "high-risk, high-reward" nature of contrarian approaches is reflected by the fact that they rarely win, but when they do, they make an impact: they surpass the prior state-of-the-art by a greater margin (40% more), break records that have stood 82% longer, and their achievements take 37% longer to be surpassed compared to mainstream algorithms.

Crucially, these differences persist even among close races, where a contrarian method barely edges out the mainstream. The fact that razor-thin victories produce such a dramatic shift in commercialization outcomes indicates that any public demonstration of viability – however slight – can "break the impasse" for contrarian approaches and trigger outsize investor interest. This result provides evidence for the informational-frictions mechanism, suggesting that proportion of funded contrarians is inefficiently low.

Finally, analyzing startups founded within a 10-year window around each contest reveals that contrarian winners are valued 3.3 times more than mainstream winners and 20 times more than the average AI startup. This valuation gap remains even after controlling for the extent of each victory, validating the notion that investors flock to contrarian winners beyond what their performance alone would justify.

[1] Details in Section 5.1.

This paper contributes to several literatures. First, it builds on a notable body of strategy literature that has long argued that differing beliefs can unveil entrepreneurial opportunities (Barney [1986]; Shane and Venkataraman [2000]; Van den Steen [2004]; Klepper and Thompson [2010]) that were overlooked by established experts and incumbents (Henderson and Clark [1990]; Christensen [1997]; Tripsas and Gavetti [2000]). However, the strategic challenges associated with commercializing a contrarian insight have received comparatively less attention. An informal, yet influential body of work (Thiel and Masters [2014]) has mostly construed contrarian insights as arbitrage opportunities to protect. A more recent theoretical literature has instead emphasized that capturing value from an unorthodox insight is contingent on matching it with the correct entrepreneurial choices (Felin and Zenger [2017]; Wuebker et al. [2023]; Agrawal et al. [2024]). In line with this, my results suggest that, even when a contrarian insight is correct, it alone may not be sufficient to spur entrepreneurship when external resources are needed (Stevenson [1983]).

Second, it contributes to a burgeoning literature on entrepreneurial experimentation. The key insight from this literature is that both society in general (Rosenberg [1992]) and entrepreneurs specifically (Camuffo et al. [2020]) benefit from rigorously testing contrasting entrepreneurial theories. Past work has considered the implications of constraints on experimentation (Kerr et al. [2014]), with a particular emphasis on costs (Ewens et al. [2018]) and commitments (Gans et al. [2019]). In contrast, this paper highlights the role of institutions like innovation contests (Brunt et al. [2012]; Boudreau et al. [2011]; Murray et al. [2012]) in facilitating the comparison of different approaches, even before commercialization. Without such comparisons, I find that contrarian ideas often remain underexplored. This represents a missed opportunity for society, as the findings suggest that contrarian ideas, when validated through these contests, tend to be particularly valuable.

To conclude, these findings provide novel evidence that, while scientific and technological progress usually occurs through incremental steps within the mainstream paradigm, major breakthroughs often come from overlooked and second-tier research camps (Kuhn [1962]; Dosi [1982]; Tushman and Anderson [1986]). Most work in this area has focused on a single sector undergoing a single technological change (Eggers and Park [2018]). My setting allows me to investigate technological change across multiple tasks and industries (ranging from cancer screening software to chatbots), to pinpoint the precise moment when technological discontinuities happen (the time of the contest), to measure the advancement using a standardized, objective metric (the percentage improvement over the current state-of-the-art), and to trace it back to a specific paper pair. My results align with the view that contrarian methods can be "high-reward"—more likely to lead to breakthroughs if successful—but also "risky," as they are unpopular, have a higher failure rate, and fall under the radar of journals and grant agencies (Wang et al. [2017]; Azoulay et al. [2023]).

The paper is structured as follows: in Section 2, I present the theoretical framework. In Section 3, I present the setting. In Section 4 and 5, I explain the data sources, how I construct the main variables, and the research design. In Section 6, I showcase the results and conclude with the strategic implications in Section 7.

2 Liability of contrarian beliefs

2.1 Defining Contrarians

A contrarian belief is a deeply held conviction that "very few people agree with" (Thiel and Masters [2014]). Contrarians believe in a controversial theory (Felin and Zenger [2017]) that directly contradicts a prevailing view. This means that for a contrarian to be right, the majority must be wrong.

The definition builds on a central premise of the entrepreneurship literature: opportunities arise because individuals hold contrasting beliefs about the value of resources and the best ways to capitalize on them[2] (Kirzner [1997]), particularly when they must make subjective judgments about an uncertain future (Knight [1921]). If everyone shared an entrepreneur's conjectures, the ensuing competition would erode potential profits, making the opportunity unattractive (Shane and Venkataraman [2000]). Therefore, entrepreneurs typically pursue opportunities they are uniquely optimistic about (Agrawal et al. [2024]).

Not all disagreements form the basis of a contrarian belief. Some disagreements are minor, concerning ancillary aspects of a value proposition or reflecting a general lack of consensus on the optimal use of a resource or the best strategy to deploy it (Felin et al. [2020]). Other disagreements are consequential, but arise from the uneven distribution of information, knowledge, or expertise across society (Hayek [1945]): by definition, however, simple differences in information cannot give rise to a *controversy*, where reasonable individuals "agree to disagree" on an opportunity (Aumann [1976]). Instead, I am interested in radical, public discrepancies in beliefs between a minority and a majority that interpret the same evidence differently (Morris [1995]; Van den Steen [2004]).

Contrarian beliefs can manifest in various ways. Entrepreneurs can radically disagree with others on what resources are valuable (Barney [1986]), the direction of an existing venture (Van den Steen [2005]), on the potential of a new product (Klepper and Thompson [2010]) or how to interpret an experiment (Agrawal et al. [2021]). In the remainder of the theoretical section, I will not take a stance. However, in the empirical section I will focus on (potential) entrepreneurs invested in competing research and technological regimes (Dosi [1982]; Anderson and Tushman [1990]; Utterback [1994]; Adner and Kapoor [2016]) that battle for technological dominance (Suarez [2004]). I will build on the notion that variation in technical approaches partly reflects heterogeneity in beliefs and cognitive representations (Tripsas and Gavetti [2000]; Eggers and Kaplan [2013]) on the best approach to solve the same technical task. As such, contrarians will be those believing in an unorthodox, lagging approach, and tirelessly "attempt[ing] to get their own technological frame to predominate in the industry" (Kaplan and Tripsas [2008]).

2.2 Entrenched Disadvantage of Contrarian Beliefs

When most people favor a conventional solution, there is typically a sound reason: more often than not, contrarians are proven wrong. This is because differences in prior beliefs—while

[2]There is an influential finance literature built around a similar insight: differences in beliefs about the fundamental value of a stock drive trading activity, with far-reaching implications on asset prices, the emergence of speculative bubbles, and trading volume. See Hong and Stein [2007] for a review.

sometimes rooted in rational considerations—can be traced back to behavioral biases, such as overconfidence (Van den Steen [2001]). As such, some contrarians are driven to doomed endeavors because they irrationally overestimate their chances of success (Moskowitz and Vissing-Jørgensen [2002]; Åstebro et al. [2014]).

Contrarian insights that turn out to be right, however, have long been recognized to drive progress (Schumpeter [1934]), and there is significant anecdotal evidence of extraordinary innovators who strenuously fought to challenge widely accepted hypotheses and achieved major breakthroughs (Karikó [2023]). The reason contrarians can create value is intuitive: if progress is stalled due to a false belief, overturning that belief can significantly advance the frontier. More subtly, contrarians are also uniquely positioned to capture the value they create. The fact that contrarian and mainstream views are antithetical – when one is right, the other is wrong – generates an asymmetry that enables the contrarian to gain a head start over the mainstream rivals (Lieberman and Montgomery [1988]; Fosfuri et al. [2013]), for example by moving first in a new technological phase (Foster [1986]).

How can contrarians turn this lead into entrepreneurial rents? An informal literature has construed contrarian insights as an arbitrage opportunity that entrepreneurs need to protect (Thiel and Masters [2014]). According to this view, a contrarian insight is a "secret" that catalyzes a "conspiracy to change the world." To capture value, contrarians should quietly exploit the discrepancy with prevailing opinion, and avoid taking actions that warn the competition about the arbitrage opportunity. The view of contrarians as mere arbitrageurs, however, is potentially too simplistic (Zuckerman [2012]). Even when contrarians are "right," realizing value depends on matching the theory with the right downstream choices (Wuebker et al. [2023]). Successful commercialization depends on attracting resources beyond the control of the entrepreneur (Stevenson [1983]), including access to complementary assets (Teece [1986]), funding (Gompers and Lerner [2001]), and human capital (Wright et al. [2001]). But to access these resources, contrarians must present their case convincingly to resource-holders, much like short-sellers must argue their position to influence market sentiment (Zuckerman [2012]).

Given that contrarians are often wrong, persuading skeptical resource-holders to back a contrarian approach over a mainstream one requires compelling evidence. The critical uncertainty of a contrarian insight centers on whether it outperforms the conventional theory, rather than other aspects of the underlying opportunity (Wuebker et al. [2023]). Shifting the prevailing viewpoint requires proof that directly addresses this core inconsistency. Without such evidence, contrarians risk being dismissed as fools, regardless of whether they are ultimately right or wrong (Aldrich and Fiol [1994]; Zuckerman [1999]; Benner and Zenger [2016]; Cattani et al. [2017]).[3] Even more challenging is persuading any single investor, as the prevailing skepticism causes them to second-guess themselves; observing that others behave as if the contrarian is wrong, an investor may hesitate to go against the herd unless they possess exceptionally strong convictions

[3]The legitimacy challenge faced by contrarians is reminiscent of the "liability of newness" (Stinchcombe [1965]). The two phenomena, however, are clearly distinct: not all contrarian ideas are novel, nor are all novel ideas contrarian. By the time the heliocentric view was accepted, for example, it was more than a century old (Kuhn [1956]). And conversely, innovations often refine and extend established paradigms (Kuhn [1962]), can enhance (rather than destroy) the incumbents' existing competences (Tushman and Anderson [1986]), and sometimes even recombine knowledge from both conventional and unconventional camps (Uzzi et al. [2013]). In the empirical analysis I will incorporate novelty and contrariness separately.

(Banerjee [1992]).

To clarify the nature of this discount, in the Appendix I develop a simple model where an investor needs to decide whether to invest in one of two ventures pursuing the same idea – one with a mainstream approach, the other with a contrarian approach. Effectively, the contrarian venture is the underdog in a winner-takes-all race against the mainstream alternative. As in classic entrepreneurial finance models (Nanda and Rhodes-Kropf [2016]), the contrarian can conduct an experiment that, if successful, increases the investor's belief that the business is valuable. However, the investor also looks at the decisions of other investors when deciding whether to support the contrarian idea.

The key takeaway is that investors might conform to mainstream investment opportunity even if the entrepreneur produces compelling, private information that the contrarian viewpoint is right. Herding, in other words, produces an entrenched advantage for the mainstream hypothesis. One way to break this dynamic is for the contrarian to produce a salient, public demonstration that shows that the contrarian is right, while the mainstream viewpoint is wrong.

2.3 Overcoming Skepticism through Public Contests

Empirically assessing the liability of contrariness—the discount that contrarians face even when they are right—requires inferring the commercialization traction that a contrarian perspective would achieve if it were mainstream. Ideally, we would observe the same approach in two different scenarios: one where the approach is considered contrarian, and another where it is regarded as mainstream. The extent to which the same approach attracts more resources when transitioning from contrarian to mainstream quantifies the friction.

One way to approximate this ideal experiment is to "shock" a contrarian idea with the chance to go mainstream. In the Appendix, I show that public contests represent an opportunity[4] to do so: contrarians can leverage them as testbeds to publicly demonstrate that their approach outperforms the orthodox one. If contests represent exogenous opportunities for a contrarian to be proven right, then analyzing how winning a contest impacts a contrarian entrepreneur's ability to attract resources and initiate commercialization quantifies the benefit of "going mainstream." Stated differently, it can help quantify the discount associated *with the same approach* depending on whether it is considered contrarian or not.[5]

The notion that contests can be informative entrepreneurial experiments combines insights from two distinct but complementary literatures. First, it has long been recognized that R&D managers can benefit from parallel testing multiple, alternative approaches (Abernathy and Rosenbloom [1969]; McGrath [2001]; Eggers and Green [2012]) while keeping other experimental conditions fixed. Past work has shown that one way to do so is to organize contests to find extreme value solutions (Boudreau et al. [2011]) from unconventional sources (Jeppesen and

[4] As an alternative, entrepreneurs can potentially isolate the effectiveness of the contrarian and the mainstream strategy by testing both as in (Agrawal et al. [2021]). This approach, however, is not only costly but outright impossible if contrarians have surpassed the commitment-free learning phase and have foregone mainstream strategic options (Gans et al. [2019]).

[5] Winning a contest can also participants using approaches already considered mainstream, for example by assuaging information frictions (Howell [2020]). I will isolate the beneficial impact of "going mainstream" by comparing the effect of winning for participants who use approaches that are contrarian vs. mainstream prior to the contest.

Lakhani [2010]). Second, a rich stream of literature has recognized the potential for contests to legitimize participating organizations: victories in contests positively impact commercialization and performance (Rao [1994]; Wade et al. [2006]; Greenberg et al. [2024]), for example by increasing a firm's reputation (Graffin and Ward [2010]) and its ability to attract key resources (Howell [2020]), or by persuading potential entrepreneurs to begin operations (Sine et al. [2007]). That being said, legitimacy-building does not always translate into firm success (Zuzul and Edmondson [2017]; Sands [2024]).

Contests also allow me to study the reaction of potential competitors to the "moment of clarity", and how that affects the value captured by contrarians. The results of the contest could upend the current status ordering of organizations (Rao [1994]) and technical approaches (Suarez [2004]), ultimately undermining the incumbents. By contrast, the clarity provided by contests could generate a massive knowledge spillover (Bloom et al. [2013], Shelef et al. [2024]) that incumbents could seize on (Cohen and Levinthal [1990]), ultimately exposing the winners to imitation, and destroying their first-mover advantage (Lieberman and Montgomery [1988]; Fosfuri et al. [2013]). The response of incumbents to a discontinuity is complex (Eggers and Park [2018]). Prior research highlights that first-mover advantages are not guaranteed, especially in rapidly evolving industries where technological shifts can erode early leads (Christensen et al. [1998]; Suarez and Lanzolla [2007]). While first-mover profits are typically linked to maintaining superior technological capabilities (Franco et al. [2009]) and safeguarding technological leadership through mechanisms like patents or secrecy (Lieberman and Montgomery [1988]; Agarwal and Gort [2001]), the visibility of a public contest and the subsequent dissemination of technical knowledge may hinder the ability to sustain such advantages.

Overall, the extent to which contrarians capture value after being publicly proven right is ultimately an open empirical question that depends on the importance of convincing skeptics and winning the technology selection process (David [1985]; Schilling [2002]), and how quickly incumbents adapt to non-incremental innovations (Anderson and Tushman [1990]; Henderson and Clark [1990]; Raffaelli et al. [2019]). Finding that contrarian insights that went mainstream are disproportionately valuable would also indicate that resource-holders are overly dismissive of contrarian approaches in general. The liability of contrariness, as such, could not be fully justified on the grounds that contrarians are typically wrong; rather, it would reflect an excessive skepticism that prevents resource-holders from recognizing and supporting valuable innovations. This would suggest that resource allocation processes might be inefficiently biased against contrarian strategies, leading to missed opportunities for innovation and growth.

3 Setting

3.1 Ideal Setting

An ideal empirical setting requires three key elements: (i) a body of potential entrepreneurs who disagree on the best strategy to pursue the same opportunity; (ii) a sudden "moment of clarity" that settles the controversy, sometimes in favor of the contrarian approach; (iii) the ability to track whether an approach gets commercialized (and how successfully).

First, the ideal setting would provide a window into the multiple, alternative approaches through which an idea *could be* pursued, some of which can be identified as contrarian and others as conventional. The challenge is not only to identify sets of potential entrepreneurs pursuing the same opportunity, for example a single technological innovation or market need (Shane [2000]). Crucially, the potential entrepreneurs should have competing theories on the best way to capitalize on the opportunity.

Second, the controversy surrounding the contrarian approach should be settled after a sudden "jolt" of clarity (Cattani et al. [2017]) that is orthogonal to a contrarian's commercial trajectory. The challenge is that competing theories might be fundamentally incommensurable and hard to falsify (Kuhn [1962]; Feyerabend [1962]). As such, a contrarian paradigm might overtake the status quo only after a prolonged, agonizing process, during which anomalies accumulate and the old ways reach their performance limit (Fleming [2001]; Sahal [1981]). Alternatively, some contrarian approaches are endogenously legitimized only through mass market success (Greenstein [2024]). In other words, not all controversies are settled in a sudden way that can be leveraged to explore whether the perceived contrariness of an approach hinders entrepreneurship.

Third, in the ideal setting I need to track whether an approach gets commercialized and how successfully. Past research on the significance of contrarian entrepreneurship has mainly depended on highly selected, impressionistic narratives of ultimately successful commercialization stories (Isenberg [2013]; Gallagher [2017]; Thiel and Masters [2014]). However, this ad-hoc approach risks falling prey to survivorship bias, as it overlooks the many contrarian ventures that are unsuccessful or fail to get started.

I address these challenges by focusing on AI contests. AI empirical research is organized in contrasting academic camps, with different researchers invested in rival algorithm families. Periodically, the community organizes evaluation challenges to compare the competing methods and track overall progress in the field. Winners usually leverage methods from the same old "state-of-the-art" algorithm family, but sometimes, a contrarian algorithm family surprisingly takes over. Finally, many AI researchers launch their own ventures to commercialize their ideas: as such, I can track how "going mainstream" affects commercialization by tracking the contestants' entrepreneurial activity before and after the sudden "moment of clarity."

In the next section, I will first illustrate the setting through the lens of Geoffrey Hinton's success in the 2012 ImageNet contest. Then, I will describe more in detail the features that make it empirically attractive.

3.2 AI Contests as an Empirical Setting

3.2.1 The case of Geoffrey Hinton and neural networks

In 2018, Geoffrey Hinton shared the Turing Award together with Yoshua Bengio and Yann LeCun (ACM [2018]). The prize was awarded for their contributions to machine learning, a subfield of AI research that studies how to develop algorithms that allow machines to automatically learn patterns from data (Michalski et al. [1984]).

The three researchers had achieved major conceptual and engineering breakthroughs that demonstrated the advantages of (deep) artificial neural networks, a family of algorithms introduced in the 1950s that had largely been sidelined by the scientific community. As noted by the ACM award committee, the three were part of a "small group who remained committed to this approach" in the early 2000s despite skepticism (ACM [2018]).

Three important factors contributed to the resurgence of neural networks. First, important algorithmic innovations solved the problem of training networks in high-dimensional parametric spaces (e.g., Hinton et al. [2006]). Second, researchers began to repurpose GPUs to speed up neural network training through parallelization (e.g., Oh and Jung [2004]).[6] Third, big datasets scraped from the internet (e.g., Torralba et al. [2008]) could be used to train bigger and deeper networks.

However – as recounted by AI luminary Jitendra Malik – these steps forward had proven insufficient to convince the computer vision community to consider neural networks competitive with the mainstream alternatives (Malik [2017]). Malik, instead, persuaded Hinton that he needed credible results on the datasets used by the community. As a result, Hinton and his team took part in the ImageNet contest in 2012, halving the error rate relative to support vector machine algorithms, the prior winners and mainstream method of the time (Russakovsky et al. [2015]). The results took the entire community by surprise, as later recounted by the contest organizer, Fei-Fei Li (Li [2023]). According to the ACM award committee, this was the key watershed moment that "reshaped the computer vision field."

As of 2024, the paper underlying that victory has been cited more than 130,000 times (Krizhevsky et al. [2012]). Immediately after the contest, the team founded DNNresearch, a startup that was quickly acquired by Google after a bidding war (Techcrunch [2013]).

3.2.2 Identifying contrarian and mainstream algorithm families

The debate over the performance of neural networks versus support vector machines in image recognition tasks illustrates a broader phenomenon in the machine learning field: various academic groups are invested in competing algorithm families, each based on distinct statistical models representing different theories for pattern extraction. These research programs progress in parallel, with each community having its own workshops, conferences, and publication venues (Birhane et al. [2022]), yet they are constantly compared to each other experimentally.

At first, the persistence of decades-long disagreements about the mathematical and statistical properties of what is essentially a best function approximator (Jordan and Mitchell [2015])

[6] Originally developed for videogames, graphical processing units were repurposed – to the surprise of most – to train deep neural networks (Hooker [2020]). That being said, the neural network community had long envisioned that better hardware was needed to fully exploit deep learning (Hinton and Anderson [1989]).

might seem puzzling. However, this controversy cannot be resolved theoretically due to the impossibility of determining a universally optimal algorithm for all tasks in advance (Wolpert and Macready [1997]). Consequently, machine learning has evolved into an experimental science, where focused theoretical work on algorithm development and improvement is paired with empirical experiments to assess performance for each specific task (Langley [1988]). The popularity and status of a research program are directly linked to the performance of its algorithms (Simon [1995]): the best-performing "state-of-the-art" is the yardstick everyone else is compared against.

As such, the setting provides two useful features to identify controversies over the best strategy to pursue an opportunity. First, there are well-defined, contrasting approaches to tackling the same task. Second, there is a natural way to identify the mainstream side—i.e., the one holding the "state-of-the-art" performance—and, by contrast, the contrarians who lag behind.[7]

3.2.3 Leveraging AI contests as moments of clarity

What does it mean for an algorithm to perform better than all others? The machine learning community has addressed this question by developing shared benchmark tasks, i.e., standardized datasets that allow different algorithms to be tested and ranked using a common metric (Hanks et al. [1993]).[8] Testbeds offer three key advantages. First, they enable the community to track progress. Second, they identify which researchers and methods push the frontier, facilitating the meritocratic allocation of academic credit (Merton [1968]) and directing research efforts toward the best-performing methods.[9] Third, they improve replicability (Price [1986]) and, as such, promote the cumulation of knowledge (Simon [1995]).

Algorithms are evaluated through two types of testbeds (or *evaluation challenges*): ongoing benchmarks and academic contests (Stolovitzky et al. [2023]).

Benchmarks refer to ongoing evaluations performed on a publicly available dataset using a predefined evaluation protocol (Prechelt [1994]). The authors of the benchmark propose a metric for judging performance, a baseline model for reference, and rules for participation. Typically, researchers interested in contributing to the benchmark can download the dataset (including the training set and a holdout set) and the code to reproduce the baseline model. Then, they train their algorithm on the training set and test it on an unseen portion of data that is kept

[7] Another key advantage of focusing on academic controversies is the paper trail they create. This trail not only reveals the author's stance in the debate—evident through the algorithms employed and the methodological references cited—but also sheds light on the nature of the breakthrough that eventually resolves the controversy. For example, the resolution might involve the introduction of a novel algorithm or an exceptional performance that significantly advances the existing frontier—critical information that is codified in the paper responsible for the breakthrough.

[8] How the community converges on an appropriate dataset and metric is beyond the scope of the paper. Past work suggests that the choice can be contentious: first, it entails difficult trade-offs between realism (is the task a meaningful milestone towards general artificial intelligence?) and experimental feasibility (Hanks et al. [1993]). Second, different benchmark choices steer innovation in different directions (Wu [2023]). As such, concerns have been raised that shared evaluations might bias research towards "canonical" tasks through peer review and publishing norms (Dehghani et al. [2021]).

[9] Paradoxically, there are growing concerns that the academic community might overemphasize state-of-the-art achievements. For instance, in 2020, an influential NLP conference advised reviewers not to reject papers solely because their results didn't surpass the state-of-the-art (`https://2020.emnlp.org/blog/2020-05-17-write-good-reviews`).

separate. Finally, they publish their model's performance—alongside the baseline and previous state-of-the-art results—and conduct ablation studies to identify the algorithmic components driving the results.

The significant advantage of benchmarks is that they allow methods to be continuously evaluated without imposing a fixed submission deadline. This reduces barrier to entry and encourages broad participation (Bischl et al. [2021]). However, their "open" nature presents two disadvantages. First, in the long run, public benchmarks face overfitting issues, where progress is achieved by exploiting the idiosyncrasies of the underlying dataset rather than the actual strength of the tested algorithms (Blum and Hardt [2015]; Recht et al. [2019]; Niven and Kao [2019]). Second, relying on self-reporting by submitters exposes the community to publication bias and potentially raises concerns about cheating, data contamination, cherrypicking, and other "tricks" (Teney et al. [2020]; Leech et al. [2024]).[10]

As an alternative, the academic community periodically organizes time-bound contests like the ImageNet competition (Balderas et al. [2024]). The organizers typically collect a new dataset and advertise the contest, enticing participants with prizes[11] and the opportunity to present at a prestigious conference. During the contest – which usually lasts no longer than 1-2 months– the participants are given a portion of the dataset to train their algorithms. Then, they submit their predictions to the organizers who check the solution and judge the accuracy using a predefined metric. The prize is typically awarded to the best performing teams provided that they beat the baseline solution introduced by the organizers.

Two key features of contests make them less vulnerable to the drawbacks of ongoing benchmarks. First, submissions are not evaluated using the same dataset available to participants. Instead, organizers typically run the algorithms on a hidden test set, mitigating concerns about cheating and overfitting (Donoho [2017]). Second, after the contest, organizers usually produce a post-challenge paper that provides a detailed, unbiased description of the top solutions (Marot et al. [2023]), thereby reducing self-reporting bias. As a result, time-bound contests more closely resemble the gold standard of a registered experiment.[12]

Academic AI contests are reminiscent of innovation prizes (Brunt et al. [2012]), grand challenges (Murray et al. [2012]) and software contests (Boudreau et al. [2011]). All of these involve competitions with the overarching goal of driving technological innovation (Boudreau and Lakhani [2016]). The AI contests I analyze, however, are a distinctly academic phenomenon, organized by academics (rather than the government or private businesses) for academics, incentivized by the prospect of kudos and publishing in prestigious venues (instead of monetary rewards or job offers). Finally, their purpose is to advance progress in well-defined tasks that the academic community is actively working on, as opposed to tackling grand social challenges or outsourcing private R&D problems.

[10]To mitigate these concerns, several benchmarks have been recently introduced together with companion websites that automatically test researchers' submissions on a secret portion of the dataset and rank them via black box software on a public leaderboard (e.g., Wang et al. [2018]). This, however, does not prevent other sources of benchmark leakage (Zhou et al. [2023]).

[11]Outside of a few cases (such as the Netflix Prize, Bennett and Lanning [2007]), prizes typically consist is small monetary rewards (typically a few thousand dollars) or hardware (e.g., GPUs).

[12]That being said, contests have neither been immune from criticism (Maier-Hein et al. [2018]) nor from cheating concerns (https://www.technologyreview.com/2015/06/04/72951/why-and-how-baidu-cheated-an-artificial-intelligence-test).

In summary, this setting provides two distinct forms of evaluation challenges to resolve a controversy: ongoing benchmarks and time-bound contests. On one hand, time-bound contests come closer to the "gold standard" of a registered experiment, where the alternative approaches are tested in a horse race under controlled conditions. On the other hand, ongoing benchmarks give researchers greater control over the timing and extent of disclosure, but at the cost of less reliable comparisons. In my analysis, I will investigate the clarifying role of both in fostering entrepreneurship.

3.2.4 Tracking commercialization

Finally, academic entrepreneurship plays a central role in commercializing AI technological innovations (Shane [2000]; Perkmann et al. [2021]). The surge in commercial interest in AI has led to hundreds of university spin-offs. Recent evidence suggests that at least 6% of founding teams include a professor or academic research associate (Chattopadhyay et al. [2024]), to the point that some worry about the impact of brain drain on AI departments (Gofman and Jin [2024]).

Part of the vibrant academic entrepreneurship activity can be traced back to the fact that AI is a burgeoning field sitting squarely in Pasteur's Quadrant (Stokes [1997]), so that the same applied models that push the frontiers of basic research can also be applied in industry (Cockburn et al. [2018]). Entrepreneurship is also driven by the scarcity of advanced technical skills and knowledge related to AI (Iansiti and Lakhani [2020]). As in the case of Hinton, academics pursue entrepreneurship to capture value from their skills, with the potential to achieve product market success or be acqui-hired (Ahmed et al. [2023]; Jurowetzki et al. [2021]; Kim [2024]).

Additionally, the decision to capture value from AI innovations through entrepreneurship—rather than, for example, patent licensing—can be attributed to the AI competitive landscape (Gans et al. [2002]). First, early entrants and pioneers have largely avoided trying to capture value from their innovations via patents.[13] Although AI patenting increased significantly in the late 2010s (Miric et al. [2023]), artificial intelligence patents are largely unenforced, untested in court, and typically accumulated as a protective measure (Choi and Gerlach [2017]). Second, digital innovations such as cloud computing have substantially reduced fixed startup costs and increased the scalability of AI software businesses since the mid-2000s (Ewens et al. [2018]).[14]

Overall, the prevalence of academic entrepreneurship enables me to directly track how winning a contest affects a researcher's transition into entrepreneurship, rather than relying on indirect measures of commercialization, such as patent citations to the winning researcher's methods.

[13] https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html

[14] This trend, however, has reversed starting from 2019 with the rise of large language models. As a consequence, startups and entrepreneurs increasingly need to grapple with the cost of computing power and shortage of GPUs needed to train a model (Ahmed et al. [2023]; Wired [2023]). The capital needed to found an AI startup has indeed substantially grown over time, with the average early-stage AI fundraising rounds now typically surpassing $10-20m (Builtworlds [2024]).

4 Data Sources and Descriptive Statistics

To empirically test how a successful public evaluation affect contrarian entrepreneurship, I rely on various data sources. First, I assembled a dataset of 575 time-bound contests sourced from publication venues, dedicated platforms, and AI conferences. Second, I collected data on 408 ongoing benchmarks hosted on Papers with Code.[15] Third, I collected information on the publications of the contestants from DBLP (a popular computer science bibliography database) and Semantic Scholar. Finally, I manually collected data on the startups founded by the participants using LinkedIn, Crunchbase, and Pitchbook. Figure 1 provides an overview of this process.

4.1 AI contests and Open Benchmarks

The dataset comprises 575 contests[16] held between 1997 and 2022. The results of 70% of these contests were published in journals or conference proceedings and were retrieved through keyword searches. The remaining were hosted on specialized academic platforms[17] or conducted during dedicated sessions at major AI conferences.[18] For each, I collected data on (i) the winning and runner-up teams, including the identity of each team member, overall team performance, and the paper describing the team's results; (ii) the type of task; (iii) the overall number of participants; and (iv) the baseline method used by the organizers.

I also collected data on all benchmarks hosted on Papers with Code, a centralized, community-contributed machine learning repository of papers, datasets, and benchmark results. Information on each paper – such as the datasets and models used, the tasks it tackles, and its performance – is automatically extracted (Kardas et al. [2020]), manually validated by the community, and used to populate benchmark leaderboards. After merging the data with DBLP and discarding benchmarks with too few submissions,[19] the remaining data comprise 408 leaderboards introduced between 2007 and 2020. Unlike the contests, I collected information on all submissions, including papers that temporarily establish new best results (the "winners") and those that do not achieve a new state-of-the-art record (the "runner-ups" that will serve as a control group).[20]

Tables B1 and B2 present descriptive statistics for contests and benchmarks, respectively, while Figure C1 illustrates the distribution of task types. Computer vision tasks are predominant in both types of evaluations. However, there are notable differences: natural language processing is overrepresented in benchmarks, reflecting the subfield's longstanding tradition of hosting these types of evaluation challenges (Schlangen [2020]). In contrast, medical tasks are more frequent in contests due to the sensitive nature of the data.[21]

[15] paperswithcode.com

[16] Each contest focuses on a distinct task and has its own prize, participants, and evaluation metric. Some are organized as independent tracks within a joint evaluation campaign.

[17] E.g., https://codalab.org

[18] E.g., https://neurips.cc/Conferences/2022/CompetitionTrack

[19] The machine learning community is disproportionately concentrated on relatively few benchmarks (Koch et al. [2021]; Martínez-Plumed et al. [2021]): I focus on the 5% with at least 20 submissions, but the following results do not depend on using different thresholds.

[20] I exclude from the analysis the paper introducing the benchmark, i.e., the first submission.

[21] Reluctance to openly share medical data on public benchmarks can be attributed to liability risks and regulatory constraints, as well as the relatively high up-front investments required to produce the data. Contests, instead, allow organizers to closely monitor who uses the data and prevent wider circulation beyond the selected

4.2 Submission-paper Pairs

Papers with Code links virtually all benchmark submissions to a publication: in 97% of the cases I was able to match a submission to a DBLP publication record. In the case of contests, 59% of the submissions had an associated paper, which I identified because it was cited by the organizers when describing the contest results, published in the contest workshop proceedings, or from the contestants' publication list.

Overall, there are 4,717 papers with one or more results in an evaluation challenge.

For each paper, I collect its citations in academic papers (sourced from Semantic Scholar, Kinney et al. [2023]) and in patents (sourced from Reliance on Science, Marx and Fuegi [2020]; Marx and Fuegi [2022]). I also rely on Semantic Scholar to retrieve each focal paper's methodological references, which are the works referenced when describing the methods employed. To do so, I exploit the fact that Semantic Scholar categorizes citation intents using the context in which a reference is cited (Cohan et al. [2019]).

Overall, these data sources allow me not only to track paper and patent citations of "winning" and "runner-up" papers after the contest or benchmark submission, but also to analyze the change in popularity of their methodological references before and after the public evaluation.

4.3 Information on Participants: Academic Career and Entrepreneurship

I gathered data on each participant's academic career using DBLP.[22] I supplemented this data with detailed information on conference best paper awards, sourced from Jeff Huang's comprehensive collection.[23] Then, I manually collected data on the participants' startup activity from LinkedIn, Crunchbase, and Pitchbook.

For contests, 3,109 researchers contributed to at least one of the 1,145 winning or runner-up submissions. In the case of benchmarks, 10,288 researchers contributed to at least one of the 13,756 submissions.[24] Overall, I collected information on a total of 12,662 participants: their descriptive statistics are in Table B3. The median researcher has written 10 computer science articles prior to their first participation (of which 20% are preprints and 32% acknowledge any funding) and has been research-active for 6 years. Only 2% of the researchers ever receive a best paper award. To conclude, 4% of the researchers founded a startup at some point in their careers.

4.4 Information on Start-ups

To measure commercialization success, I collected data on 477 startups the participants founded between five years before and five years after their first submission. Information includes founding year, amount of venture capital raised, last valuation, exits (including IPO or

participants (Karargyris et al. [2023]).

[22] For ongoing benchmarks, every submitter has a DBLP profile. This is because I exclude the 3% of submissions where a publication record on Papers with Code cannot be linked to a DBLP record. In time-bound contests, about 10% of participants do not have a DBLP profile, primarily due to challenges in accurately disambiguating their names.

[23] https://jeffhuang.com/best_paper_awards

[24] Some papers contribute to multiple leaderboards, as they tackle multiple tasks, use different datasets, or introduce various models at a time. In the empirical section I describe how I take that into account in the analysis.

acquisition), whether the startup has any patents, has received funding from corporations or the government, and number of competitors at entry (Table B4).

The median startup in the sample was founded in 2017, has raised less than $1m in venture capital, has no granted patent, and enters a market with 862 other active competitors (i.e., firms with the same Pitchbook keywords). These figures, however, mask a significant skewness (Scherer and Harhoff [2000]) driven by a minority of highly successful ventures. Notably, 17% of the startups are acquired (on average as soon as 4 years after founding), and 8% achieve an exceptional valuation (i.e., achieve unicorn status or go public)[25] by 2024. These outcomes are notable: compared to the universe of AI startups in Pitchbook – of which only 1.1% to 1.9% become unicorns or go public – the startups in the sample are 7 times more likely to be successful.

[25]Traditionally, going public has been a critical milestone of success for companies that are not acquired. This metric, however, does not capture the increasingly important phenomenon of unicorns, i.e. private companies with valuations exceeding $1 billion (Brown and Wiles [2015]) that increasingly choose to remain private to avoid the costs, disclosure requirements and regulatory scrutiny associated with public markets (Erdogan et al. [2016]; Ewens and Farre-Mensa [2022]; Ewens et al. [2024]). As such, solely relying on IPOs as a metric of success would introduce both measurement error (i.e., fail to take into account successful startups that choose to stay private) and exacerbate truncation concerns (as highly valuable companies increasingly take longer to go public).

5 Main Variables and Research Design

In this section, I describe how I separately measure whether a submission is contrarian, novel, and/or breakthrough. While teasing apart these distinct dimensions requires nuance (Kovacs et al. [2019]; Capponi et al. [2022]), I can leverage unique aspects from my setting.

5.1 Measuring Contrarianism

I classify a participant's method as contrarian if it belongs to an algorithmic family that had never been state-of-the-art before. I determine this using two complementary methods.

For all 575 contests, I manually classify the models of the winners, runner-ups, and the state-of-the-art prior to the contest (i.e., prior winners and the organizers' baseline) into one of 20 algorithmic families. These families (e.g., neural networks and kernel methods), established before my sample period, provide a natural partition of existing models into competing research paradigms. To build the taxonomy, I rely on various sources, including textbooks (Murphy [2022]), official ACM[26] and IEEE taxonomies[27], practitioner-contributed ontologies[28] and widely used machine learning libraries (Scikit [2024]). A full list of the algorithm families, including examples of individual algorithms for each family, is available in Table C1.[29] Contrarian participants are those who use or introduce models associated with an algorithm family that lags behind the mainstream state-of-the-art family prior to the contest: according to this definition, only 13% of the submissions are contrarian (Table B1).

This classification, while requiring only a description of the methods, is inherently manual and time-consuming. As an alternative, I classify a participant's methods as contrarian or mainstream based on the degree of overlap between the methodological references mentioned in their paper and the prior state-of-the-art.[30] Intuitively, the fewer methodological references a paper has in common with the prior state of the art, the more distant the methods and more likely that the two build on different algorithmic families. This second measure is only available when the methods are described in standalone papers with references. When both measures are available, they are highly correlated and yield very similar results. Since all benchmarks submissions are associated with a paper, I rely on the reference-based measure for analyzing this subset of the data.

5.2 Measuring Novelty

As highlighted in Section 2.2, novel ideas often face the same legitimacy challenges as contrarian ideas, and both can benefit from public evaluations. This overlap can make it difficult to isolate the effects of winning an evaluation challenge on contrarians, particularly if contrarian

[26] https://dl.acm.org/ccs

[27] https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-taxonomy.pdf

[28] https://paperswithcode.com/methods

[29] The comprehensive classification of all algorithms used in the time-bound contests, categorized into the 20 algorithmic families, is available upon request.

[30] I discretize this continuous measure so that the resulting share of contrarians reflects the proportion observed according to the first measure.

contributions are also likely to be novel.[31]

Usefully, the setting allows to distinguish between these two dimensions. To determine if a contribution introduced a new model, I use Semantic Scholar, which (i) assigns topics to each computer science paper based on its title and abstract, and (ii) leverages the context in which each topic is mentioned to produce a brief description and identify the paper that likely introduced it. For each contest and benchmark submission, I check whether it introduced a new topic. If it did, I determine whether the topic refers to a model or algorithm based on its description. I define a submission as having introduced an algorithmic innovation (i.e., being novel) if its associated paper has introduced a new topic describing a model. Finally, I manually reviewed each newly introduced model and classified it as a seminal innovation if it has a dedicated Wikipedia page.

Overall, 59% of the papers describing one or more submissions have introduced a novel machine learning model or component, and 4% introduced a seminal algorithm.

5.3 Measuring Progress

Since contest submissions are judged using a single objective metric, I can measure the performance gap between winners and second-place contributions (or previous state-of-the-art results). This measure quantifies the extent to which a winning contribution has pushed the frontier forward (Sherry and Thompson [2021]).

Figure C2 shows the distribution of the percentage improvements for all winners and new state-of-the-art results. The distribution is very skewed, with 25% of the winning results improving by less than 1% relative to the frontier. By contrast, Hinton's team won the 2012 ImageNet contest with a classification error 41.5% lower than the closest competitor, which corresponds to the 93$^{\text{rd}}$ highest percentile.

5.4 Contrarian vs. Novel vs. Breakthrough

Figure C3 shows the relative proportions of winning submissions that are contrarian, novel, and breakthrough (achieving $> 10\%$ improvement over the existing state-of-the-art). The Venn diagram highlights that these three categories do not perfectly coincide. For example, Alphafold 2 (Jumper et al. [2021]; Tunyasuvunakool et al. [2021]) won the 2020 protein folding CASP assessment dramatically, with a score 2.7x higher than the runner-up. According to Semantic Scholar, the paper introduced a novel topic, "Alphafold 2," described as a machine learning method. However, Alphafold 2 is part of the same algorithm family as Alphafold, the previous winner (Senior et al. [2020]). Thus, Alphafold 2 is breakthrough and novel, but mainstream. In contrast, the winning submission to the Netflix Prize (Koren et al. [2009]) did not introduce a specific algorithmic innovation and won the contest by a tiebreaker.[32] However, the methodology was based on matrix factorization techniques that became popular only after the Netflix Prize results demonstrated their superiority to classical nearest-neighbor techniques, which were used

[31] This could be the case, for instance, if the degree of novelty in mainstream solutions is be constrained by convention (Uzzi et al. [2013]).

[32] https://www.thrillist.com/entertainment/nation/the-netflix-prize

by Cinematch, the contest baseline set by Netflix (Bennett and Lanning [2007]). As such, the winning submission is contrarian, but neither novel nor breakthrough.

5.5 Research Design

In the main specification, I leverage the timing of a successful submission to an evaluation challenge to understand how a public demonstration affects the probability of founding a startup. To achieve this, I employ the following event study design:

$$Y_{it} = \sum_{j \neq -1} \tau_j D_{i,t-j} \times \text{Win}_i + \beta_i + \gamma_c + \alpha_t + \varepsilon_{it} \tag{1}$$

Where:

- Y_{it} is 1 if the researcher has founded a startup.
- β_i, γ_c are researcher and contest fixed effects.
- $D_{i,t-j}$ is an indicator for being j years away[33] from the public demonstration.
- Win_i is 1 if the researcher has won the contest or established a new state-of-the-art in an open benchmark, and 0 otherwise. To capture heterogeneous effect for contrarian and mainstream winners, I further disaggregate Win_i to separately account for either.

The control group in time-bound contests consists of runner-ups. For ongoing benchmarks, it includes researchers who submitted in the same year and to the same benchmark as the winners but did not achieve a state-of-the-art result. This setup enables me to compare quasi-"twin" contributions (Bikard [2020]) – manuscripts published (almost) simultaneously and addressing identical machine learning task – where one has won the contest and the other has not. As such, τ_j identifies the effect of winning on the likelihood of founding a startup j years after the contest.[34] To avoid "forbidden" comparisons (Roth et al. [2023]), I run a stacked regression with first-time winners as treated units and first-time, never-winners as controls (Cengiz et al. [2019]).

The research design eliminates concerns about differences in levels between winners and runner-ups affecting the estimates: simple differences in entrepreneurial propensity, knowledge, network, or access to resources between winners and runner-ups would influence startup activity both before and after the contest and, thus, do not bias the difference-in-differences estimates. The identification of τ_i, however, relies on two key assumptions: parallel trends and no anticipation. First, in the absence of the contest, winners and runner-ups would have evolved similarly. Second, contestants should not alter their behavior in anticipation of a future victory.

Threats to identification arise from either assumption. While differences in pre-trends between treated and control groups can be empirically ruled out by checking for the absence of

[33] In the main specification, I focus on a (-5, 5) time window around the contest.

[34] I augment this specification by applying a standard difference-in-difference approach, comparing the differences between winning and runner-up researchers both at the end of the time window and just before the contest. Additionally, to address concerns about right-censoring and potential biases arising from entrepreneurship being an absorbing state (see Farbmacher and Tauchmann [2023]), I also employ a difference-in-differences Cox proportional hazards model to examine the effect of winning on time-to-entrepreneurship (e.g., Mastrobuoni and Pinotti [2015]).

pre-trends before the evaluation, the timing of contests could correlate with unobserved technological shocks (e.g., GPUs) that could favor certain researchers, influencing both their likelihood of winning and their startup activity. Selection into treatment also poses a potential issue: entrepreneurs might intentionally delay commercialization until after winning, anticipating that success could boost their chances. If those likely to win are already more inclined to start a company, even without the contest, this selection bias could lead to an overestimation of the effect of winning on entrepreneurial propensity.

However, two key factors mitigate these concerns. First, the timing of episodic contests is generally exogenous to the trajectory of individual researchers, as these events are organized by independent committees with their own agendas and timelines. In other words, contests often present themselves as unexpected opportunities, with participants becoming aware of them only a few months in advance.[35] Second, many races are close, with winners and runner-ups achieving nearly identical performance. Previous research has shown that the gap between first and second place is often so narrow that rankings could shift if evaluated with a different subset of the data (Varoquaux and Cheplygina [2022]) or an alternative metric (Maier-Hein et al. [2018]). Therefore, by focusing on challenges where luck likely influenced the outcome—i.e., by employing a regression discontinuity design to isolate the effect of winning—I can minimize endogeneity concerns.

To achieve this, I employ the following difference-in-discontinuity design (Grembi et al. [2016]):

$$\begin{aligned} Y = {} & \delta_0 + \delta_1 \text{Gap} + \text{Win} \times (\gamma_0 + \gamma_1 \text{Gap}) \\ & + \text{Post} \times [\alpha_0 + \alpha_1 \text{Gap} + \text{Win} \times (\beta_W + \beta_1 \text{Gap})] \\ & + \text{Controls}, \quad \text{for Gap} \in [-h, h] \end{aligned} \tag{2}$$

[36]

This specification combines the advantages of a difference-in-difference with those of a regression discontinuity design. Specifically, I identify the causal effect of winning on Y by exploiting both the timing of the challenge submission and the fact that winning is a discontinuous function of Gap, which represents the scaled distance between winners and runner-ups. In other words, if there is no other factor – aside from the challenge itself – that explains the sudden increase in Y as Gap transitions from negative to positive, then the coefficient β_W captures the causal effect of winning on entrepreneurship.

[35] This is not necessarily the case for submissions to open benchmarks, whose timing and disclosure level is at the discretion of the individual researchers.

6 Results

This section is organized as follows. Sections 6.1 and 6.2 show that evaluation challenges are consequential, both in terms of shifting the perception of ideas in academic circles and by creating entrepreneurial momentum. Section 6.3 explores the differences between contrarian and mainstream researchers prior to the challenge and their performance during the evaluation campaign. Section 6.4 highlights how winning disproportionately promotes contrarian entrepreneurship. Finally, Section 6.5 examines the differences between startups founded by contrarian and mainstream winners.

6.1 Impact of Winning a Challenge on Citations

I begin by examining whether a success in an evaluation challenge – whether a contest or an ongoing benchmark – leads to increased academic recognition. To explore this, I analyze differences in citations and awards accrued by papers describing winning versus runner-up submissions.

First, I consider whether winning contributions are more likely to be published. While nearly all benchmark submissions are associated with a paper pair (i.e., unpublished contributions remain largely unobserved), column 1 of Table 1 compares the likelihood that a winning submission in a time-bound contest is published in a standalone paper, relative to runner-ups. I find that winners are 12.1 percentage points (23%) more likely to be published than runner-ups. Therefore, the differences in academic recognition between winners and runner-ups, discussed below, likely understate the true gap between these two groups.

In columns 2-4, I analyze all papers describing a submission to a contest or benchmark. I find that winning contributions, or those that establish a new state-of-the-art, are 0.3 percentage points (8x) more likely to receive a best paper award, receive on average 121% more paper citations, and are cited 61% more in patents than runner-ups.

Part of the difference in recognition between papers that establish a new state-of-the-art and those that do not may stem from the vitality of different subfields and machine learning tasks. For instance, the community might prioritize tasks experiencing rapid progress and frequent breakthroughs. To account for this, I include contest fixed effects in columns 5-7. I focus on "racing pairs" (Hill and Stein [2024]) – that is, comparing winning and losing contributions from the same challenge – to isolate the effect of winning on academic recognition. I find that winners receive, on average, 70% more paper citations, 23% more patent citations, and are 0.6 percentage points (121%) more likely to receive a best paper award than runner-ups.

These results hold even after controlling for novelty and the extent to which winning contributions advance the research frontier. While a paper's novelty strongly predicts future recognition – novel contributions receive on average 814% more paper citations, 281% more patent citations, and are 3.6 times more likely to receive an award – winning still boosts academic recognition relative to similarly novel runner-ups. In columns 8-10, I show that winning contributions are cited more frequently in both papers and patents, even when controlling for the performance gap between winners and runner-ups.

Overall, these findings suggest that evaluation challenges effectively spotlight the best-

performing methods. Winners gain significant academic and commercial recognition, benefiting from a "state-of-the-art" premium that, according to the most conservative estimates, is more than double the priority premium found in structural biology (Hill and Stein [2024]).

6.2 Impact of Winning a Challenge on Academic Entrepreneurship

In this section, I shift to a researcher-level analysis and ask whether a successful public evaluation fosters entrepreneurship. The main specification is the difference-in-difference event study design described in Eq. 1. Figure 2 plots the difference in the propensity to found a startup between winners and runner-ups, τ_j, before and after the contest or benchmark submission. Relative to runner-ups, winners become more likely to found an AI startup immediately after the demonstration: after four years, winning makes them 2.7 pp. more likely to have transitioned to entrepreneurship. I find no evidence of pre-trends: winners and runner-ups have a similar trajectory prior to the contest, and they start to diverge immediately afterward.

To further unpack this result, Table 2 reports the coefficient of a difference-in-difference that parallels the above event study. Researchers are observed in two periods, one year before and four years after the evaluation; winning results in a 2.3 pp. increase in the likelihood to found a startup by the fourth year after the challenge submission relative to the control group. The findings are robust to controlling for the novelty of the method (column 2), which does not predict transition into entrepreneurship.

To further validate causality, I leverage the fact that many races are close to see if even winners who performed similarly to runner-ups are also disproportionately likely to become entrepreneurs. Column 3 shows the estimates from the difference-in-discontinuity design described in Eq. 2, while Figure C4 shows the results of a regression discontinuity design including only researchers that did not found a startup prior to the evaluation challenge. In both cases, I find that winners of close races are 1.8-1.9 pp. more likely to become entrepreneurs after the contest.

To conclude, in column 4 I find that the bulk of the effect is concentrated among time-bound contests, with the effect being 3x higher than in ongoing benchmarks. Taking stock, the findings suggest that the significance of AI evaluation challenges – and particularly contests – extends beyond academia. The 1.8-2.7 percentage points increase in entrepreneurship following a victory is economically significant, representing an increase of around 40-60% relative to the counterfactual.

6.3 High Risk, High Gain of Contrarian Research

This subsection draws a contrast between researchers using contrarian and mainstream methods, highlighting both how they differ before the public evaluation, and how they perform during it.

To begin with, in Figure C5 I analyze differences prior to the first contest or open benchmark submission. First, I find that contrarians have produced, other things equal, substantially fewer publications (19% less) and the publications they have produced are disproportionately preprints (5pp. or 23% more). The discrepancy cannot be fully explained by the fact that contrarians have less research experience (one year less on average). Instead, it is consistent with the notion of contrarians as lower-status outsiders who struggle to get published (Simcoe

and Waguespack [2011]) and to attract attention from the rest of the academic community (Azoulay et al. [2014]).[37] Second, I find that contrarians' publications are 3 pp. (i.e., 10%) less likely to cite funding sources in the acknowledgements,consistently with a growing literature on how outsiders are shunned by funding organizations (Wang et al. [2017]) because of excessive conservatism (Lane et al. [2022]).

Next, I focus on differences in mainstream and contrarian submissions to the public evaluation (Table C2). In Columns 1-4, I show that the contrarian contributions are no more likely than the mainstream ones to introduce new methods (including seminal ones). In Columns 5-8 I show that contrarians perform worse than mainstream contestants on average, even after controlling for the fact that few participants are contrarian in the first place: on average, they rank between four to eleven positions lower in the leaderboard than mainstream contributions and are 2.7 pp. to 12.8 pp. less likely to win.[38] This is consistent with prior evidence that new state-of-the-art results usually come from the same communities (Martínez-Plumed et al. [2021]), and can rationalize higher rejection rates if contrarian research ideas are perceived to be "risky" and more likely to underperform (Azoulay and Li [2021]; Myers et al. [2023]). Interestingly, however, in Table C3 I find that when they do win, contrarians push the research frontier further (by 4.4 pp., i.e. 40% more than mainstream winners), break more longstanding records (i.e., records 82% older), and set records that will withstand the test of time (lasting 37% longer). These findings align with the long-standing, but undertested notion that contrarians disproportionately contribute to major breakthroughs (Dosi [1982]). Overall, they suggest that while contrarian methods may often underperform, they hold the potential for groundbreaking, long-lasting advancements when successful.

6.4 Impact of Winning a Contest on Contrarian Entrepreneurship

Next, I analyze the effect of winning separately for researchers with mainstream and contrarian methods: I find that a success boost entrepreneurship 70% more for contrarian winners compared to mainstream winners (Table 3, column 1).

The discrepancy is robust to controls for novelty (column 2), which explains little of the propensity to become entrepreneurs. When focusing on close races through a differences-in-discontinuity design (Eq. 2), I find that the difference between contrarian and mainstream winners increases significantly (12 pp.). In other words, much of the boost for contrarians is driven by demonstrating that they can surpass mainstream methods (i.e., a reshuffling in the ordinal ranking between methods) rather than by the strength of their performance (i.e., the magnitude of the cardinal improvement).

In column 4, I analyze contests and ongoing benchmarks separately, finding that contrarian

[37] Indeed, there are many examples of seminal machine learning techniques that were rejected from mainstream journals and circulated as preprints, such as the dropout technique – rejected from NIPS in 2012 and only accepted by JMLR in 2014 (Srivastava et al. [2014]) – and SIFT – rejected twice in 1997, and only published in 1999 as a poster and finally in 2004 in a journal (Lowe [2004]).

[38] The more conservative estimates include challenge fixed effects. In time-bound contests, including these fixed effects effectively compares whether contrarians are less likely to win conditional on having at least reached second place within each challenge. Since contrarians are disproportionately less likely to reach second place to begin with (as judged from the ongoing benchmarks), this approach results in more conservative estimates of their likelihood to win.

winners especially benefit from winning a time-bound challenge: there, the effect is 7.25 times higher for contrarian winners than mainstream winners. Figure 3 reproduces the event study design described in Eq. 1 separately for contrarian and mainstream winners in contests: both contrarian and mainstream winners have a similar trajectory as runner-ups prior to the contest. Immediately after the contest, however, the probability of founding a startup increases for both, but particularly so for contrarians: after four years, mainstream winners are 3 pp. more likely to become entrepreneurs than runner-ups, while contrarian winners are 21 pp. more likely to found a startup than the same control group.

This discrepancy is consistent with the notion that time-bound contests shed more clarity on the benefit of a contrarian approach, both because they are a more trustworthy demonstration and because they draw a sharper contrast with mainstream methods. In Table C4, column 1, I find that a new state-of-the-art result on an open benchmark only fosters contrarian entrepreneurship if the underlying benchmark data are made widely available to the community through a centralized repository, making the result easy to validate and replicate. In column 2, I find that contrarian state-of-the-art results that are "short-lived" – i.e., quickly surpassed by other researchers – are disproportionately less likely to result in a commercialization event.

Finally, I find that losing contrarians enjoy a positive spillover effect when defeated by a contrarian algorithm, whereas no spillover effect occurs for mainstream runner-ups or when a mainstream algorithm wins (Figure C6). This provides further evidence that winning certifies the entire research program, not just individual winners. This is consistent with prior work (Reschke et al. [2018]) on the nuanced effect of prizes on the attention received by the recipients' "neighbors." Competition for attention mostly results in a negative spillover effect, unless the scientific space is emerging – in which case the prize endorses the entire camp.

6.5 Contrarian Winners and Startup Outcomes

In this section I analyze the outcomes of startups founded around the time of the evaluation challenge, including venture capital raised, their valuation, being acquired, and going public (Table 4).

First, consider the amount of venture capital raised (column 1). Contrarian winners raise five times as much as startups founded by mainstream winners or runner-ups. Conditional on raising, contrarian winners do so at around 3.3x the valuation of the startups founded by other participants, whereas the difference between mainstream winners and runner-ups is small and insignificant (Column 2). Similar large differences are observed when focusing on late-stage success and exit: startups founded by contrarian winners are 4.5 times more likely to achieve an IPO or reach unicorn status compared to those founded by losers, and 2.3 times more likely than those founded by mainstream winners (column 3). In column 4, I find that winners are 50% more likely to be acquired than startups founded by runner-ups, although there is little difference between contrarian and mainstream winners. Taken together, these discrepancies do not reflect a low quality of startups founded by runner-up or mainstream founders. Instead, the results suggest that contrarians – while relatively less likely to engage in entrepreneurship when they lose – found disproportionately more successful startups when they win: in comparison to the average AI startup tracked by Pitchbook, the valuation of contrarian winners is around 20

times higher.[39]

What factors drive these differences in outcomes? In column 6, I rule out intellectual property rights: startups founded by contrarian winners are no more likely than those founded by other contestants to have a patent. Intriguingly, I also find that they enter markets with three to four times as many incumbents as those entered by runners-up and mainstream winners (column 7). I can also rule out incumbent inattention. The results of the public demonstration hardly fly under the radar: I find that winners in general, and contrarians in particular, are immediately recognized through best paper awards (Figure C7). Moreover, despite entering a crowded field without superior intellectual property, contrarian winners prompt a swift reaction from incumbents: startups founded by contrarian winners are, on average, acquired 50% sooner after their founding than mainstream winners and runners-up (column 5).[40]

The evidence instead suggests that while the public demonstration draws the attention of the research community and validates the contrarian approach, it does not translate into an immediate adoption of the new state-of-the-art techniques. To show this, in Figure 4 I analyze the citation trajectory of contrarian and mainstream winning submissions (relative to runner-up submissions). After the evaluation, the number of citations accruing to mainstream winners grows faster than those accruing to the contrarian winners: the gap hovers at around 30%-50% and only closes after 10 years. This pattern is even clearer when examining the citations to the methods referenced by the winners. In Figure C8, I plot the citation trajectory of the methods used by winners before and after the evaluation (using the methods used by runner-ups as a control group). The methods used by mainstream winners become immediately more popular after the contest. Instead, the methods used by contrarian winners are growingly *unpopular* prior to the contest: this downward trend starts to slow down only after the success and reverses 4 to 5 years after the contest. This result is consistent with prior studies showing that scientists are slower to incorporate in their work research produced by outsiders (Cole [1970]) and face high switching costs associated with changing research direction (Myers [2020]). For example, (Malik [2017]) recounted that the computer vision community took one to two years to fully replicate Hinton's result and test neural networks on datasets other than Imagenet.[41] Moreover, incorporating new practices into an incumbent's mature and highly complex machine learning system requires significant time and careful design (Sculley et al. [2015]). To sum up, the comparatively slower adoption of contrarian approaches turned "best practices" explains why the contrarian winners seem to retain a first-mover advantage even in the absence of superior patent protection, prompting incumbents to update their absorptive capacity (Cohen and Levinthal [1990]) by quickly acquiring the contrarian startups (Arora et al. [2002]).

To conclude, the disproportionate success of startups founded by contrarian winners suggests that contrarian ideas, when given the opportunity to overcome commercialization barriers, can

[39] Based on personal calculations from Pitchbook data.

[40] The government, instead, is 33% more likely to support mainstream than contrarian winners through grants.

[41] "Everyone was impressed by the results, but there was debate about their generality. Would the success on whole image classification problems extend to more tasks such as object detection? Was the finding a very fragile one, or was it a robust one that other groups would be able to replicate? Stochastic gradient descent (SGD) can only find local minima, so what is the guarantee the minima we find will be good? In the true spirit of science, many of us, skeptics and believers, went back to our laboratories to explore these questions. Within a year or two, the evidence was quite clear." Malik [2017]

generate substantial value. If this success reflects the untapped potential of other contrarian ideas that currently lack such opportunities, then reducing these barriers could enable society to benefit more from unconventional innovations. In other words, the obstacles faced by contrarians may be unnecessarily restrictive, and easing them could unlock welfare gains.

7 Discussion and Conclusion

Research in innovation and entrepreneurial strategy has long recognized the outsized role contrarians play in driving innovation through entrepreneurship (Schumpeter [1934]). Dramatic accounts of heroic outsiders have underscored both the potential significance of contrarians and the skepticism they face even when they are right (Karikó [2023]). Yet, contrarianism is challenging to study because it is hard to track controversies about entrepreneurial opportunities and determine how easier it would be to commercialize a contrarian idea if it was perceived as mainstream. This paper presents some of the first empirical evidence on the phenomenon.

First, I conceptualize contrarians as agents with unusual beliefs (Agrawal et al. [2024]) who have a theory of value that contradicts the mainstream view. As such, contrarians suffer from an entrenched disadvantage: the only experiment that can persuade investors is a test that proves the mainstream hypothesis wrong. I test this framework by leveraging AI contests as "moments of clarity" that can potentially prove a contrarian right.

My results suggest that public demonstrations are key stepping stones for contrarians to initiate commercialization. Stated differently, in the absence of validation, contrarians face substantial frictions that make them less likely to initiate commercialization than when they "go mainstream." I also corroborate two long-standing hypotheses about contrarians: they are often wrong, but when they are right, their insights are disproportionately impactful.

The disproportionate commercial success of contrarian startups founded after winning a contest indicates that once contrarian ideas overcome commercialization barriers and gain mainstream acceptance, they can generate substantial value. This suggests that the barriers contrarians face prior to (or in absence of) a contest may be unnecessarily restrictive. If contrarian ideas are capable of producing outsized returns when given the opportunity, then resource-holders may be excessively dismissive of contrarian approaches. The liability of contrariness would not be fully justified by the notion that contrarians are typically wrong; rather, it could reflect an excessive skepticism that prevents valuable innovations from receiving support. Reducing these barriers could enable society to capture greater value from unconventional innovations.

The evidence challenges an informal literature that has suggested contrarians to embrace skepticism, keep their insights secret, and avoid competition (Thiel and Masters [2014]). My results suggest instead that contrarians retain a competitive advantage despite the public disclosure and even when entering crowded fields without superior intellectual property assets. Why do incumbents fail to capitalize on the free and salient knowledge spillover generated by the contest? Managerial inattention (Eggers and Kaplan [2009]) is unlikely, as contrarian successes are well documented and publicly recognized through prominent paper awards. Concerns about cannibalization or commitments to an existing customer base (Christensen and Rosenbloom [1995]; Christensen and Bower [1996]) can also be plausibly ruled out, given that the contrarian approach is objectively superior to the previous state-of-the-art while targeting the same task. This conclusion is further supported by the fact that incumbents quickly acquire contrarian startups, reacting through markets for ideas, rather than directly adapting lessons from the contest. Instead, the evidence suggests that incumbents and other researchers are slow to assimilate the new knowledge (Eggers and Park [2018]), despite the attention the contest attracts. Citations to winning contrarians accumulate at a more gradual pace compared to mainstream winners; a

sign that the community takes time to react to the methodological discontinuity.

A practical implication of this study is that skepticism is the key barrier that prevents contrarians from attracting resources (Stevenson [1983]) – entrepreneurs should prioritize persuading resource-holders rather than worrying about generating a competitive response. To do so, the results emphasize the importance of designing an experiment that is credible and rigorously tests whether the contrarian approach is superior to the conventional one. Policymakers, in principle, could play an active role in organizing benchmarking contests to test contrasting entrepreneurial approaches, leveraging their extensive experience in organizing innovation prizes (Williams [2012]).

The main limitation of the study is that it is circumscribed to a specific field that values commensurability and has developed institutions that contrarians can leverage to be proven right. It is reasonable to assume that this may not be the case in many settings. As such, entire fields can potentially remain stuck in unproductive paradigms – "bubbles" that contrarians are unable to burst (Zuckerman [2012]). What strategy contrarians should adopt in such circumstances remains a question for future research. Second, while the analysis suggests that the barriers faced by contrarians may reduce overall welfare, it does not quantify the societal costs of missed opportunities or identify the optimal level of support for contrarian ideas that would maximize welfare. These and other questions could be addressed in future work.

References

Abernathy, W. J. and Rosenbloom, R. S. (1969). Parallel Strategies in Development Projects. *Management Science*, 15(10):B–486–B–505.

ACM (2018). Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award. *https://awards.acm.org/about/2018-turing*.

Adner, R. and Kapoor, R. (2016). Innovation ecosystems and the pace of substitution: Re-examining technology S-curves. *Strategic Management Journal*, 37(4):625–648.

Agarwal, R. and Gort, M. (2001). First-Mover Advantage and the Speed of Competitive Entry, 1887–1986. *The Journal of Law and Economics*, 44(1):161–177.

Agrawal, A., Camuffo, A., Gambardella, A., Gans, J. S., Scott, E. L., and Stern, S. (2024). Bayesian Entrepreneurship. *Working Paper*.

Agrawal, A., Gans, J. S., and Stern, S. (2021). Enabling Entrepreneurial Choice. *Management Science*, 67(9):5510–5524.

Ahmed, N., Wahed, M., and Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635):884–886.

Aldrich, H. E. and Fiol, C. M. (1994). Fools Rush in? The Institutional Context of Industry Creation. *The Academy of Management Review*, 19(4):645.

Anderson, P. and Tushman, M. (1990). Technological Discontinuities and Dominant Designs: A Cyclical Model of Technological Change. *Administrative Science Quarterly*, 35(4):604–633.

Arora, A., Fosfuri, A., and Gambardella, A. (2002). *Markets for technology: the economics of innovation and corporate strategy*. MIT Press, Cambridge, Mass.

Aumann, R. J. (1976). Agreeing to Disagree. *The Annals of Statistics*, 4(6).

Azoulay, P., Greenblatt, W., Levy, R., and Maity, S. K. (2023). Does Grant Peer Review Penalize Scientific Risk Taking? Evidence from the NIH. *Working Paper, MIT*.

Azoulay, P. and Li, D. (2021). SCIENTIFIC GRANT FUNDING.

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

Balderas, H. J. E., Guyon, I., Howard, A., Reade, W., and Treguer, S. (2024). Challenge design roadmap. arXiv:2401.13693 [cs].

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817.

Barney, J. B. (1986). Strategic Factor Markets: Expectations, Luck, and Business Strategy. *Management Science*, 32(10):1231–1241.

Benner, M. J. and Zenger, T. (2016). The Lemons Problem in Markets for Strategy. *Strategy Science*, 1(2):71–89.

Bennett, J. and Lanning, S. (2007). The Netflix Prize. *Proceedings of KDD cup*.

Bikard, M. (2020). Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8):1528–1543.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2022). The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, Seoul Republic of Korea. ACM.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N., and Vanschoren, J. (2021). OpenML Benchmarking Suites. arXiv:1708.03731 [cs, stat].

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying Technology Spillovers and Product Market Rivalry. *Econometrica*, 81(4):1347–1393.

Blum, A. and Hardt, M. (2015). The Ladder: A Reliable Leaderboard for Machine Learning Competitions. arXiv:1502.04585 [cs].

Boudreau, K. J., Lacetera, N., and Lakhani, K. R. (2011). Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis. *Management Science*, 57(5):843–863.

Boudreau, K. J. and Lakhani, K. R. (2016). Innovation Experiments: Researching Technical Advance, Knowledge Production, and the Design of Supporting Institutions. *Innovation Policy and the Economy*, 16:135–167.

Brown, K. C. and Wiles, K. W. (2015). In Search of Unicorns: Private IPOs and the Changing Markets for Private Equity Investments and Corporate Control. *Journal of Applied Corporate Finance*, 27(3):34–48.

Brunt, L., Lerner, J., and Nicholas, T. (2012). Inducement Prizes and Innovation. *The Journal of Industrial Economics*, 60(4):657–696.

Builtworlds (2024). Inside an AI Startup and the Need for Large, Early-Stage Funding Rounds.

Cabral, L. M. B. (2003). R&d competition when firms choose variance. *Journal of Economics & Management Strategy*, 12(1):139–150.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal: Promoting communications on statistics and Stata*, 14(4):909–946.

Camuffo, A., Cordova, A., Gambardella, A., and Spina, C. (2020). A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial. *Management Science*, 66(2):564–586.

Capponi, G., Martinelli, A., and Nuvolari, A. (2022). Breakthrough innovations and where to find them. *Research Policy*, 51(1):104376.

Cattani, G., Ferriani, S., and Lanza, A. (2017). Deconstructing the Outsider Puzzle: The Legitimation Journey of Novelty. *Organization Science*, 28(6):965–992.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The Effect of Minimum Wages on Low-Wage Jobs*. *The Quarterly Journal of Economics*, 134(3):1405–1454.

Chamley, C. P. (2003). *Rational Herds: Economic Models of Social Learning*. Cambridge University Press.

Chattopadhyay, S., Honore, F., and Won, S. (2024). Academic Entrepreneurs and Market Opportunities in Artificial Intelligence: A Problem-solving Perspective. *Working Paper*.

Choi, J. P. and Gerlach, H. (2017). A Theory of Patent Portfolios. *American Economic Journal: Microeconomics*, 9(1):315–351.

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

Christensen, C. M. and Bower, J. L. (1996). Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal*, 17(3):197–218.

Christensen, C. M. and Rosenbloom, R. S. (1995). Explaining the attacker's advantage: Technological paradigms, organizational dynamics, and the value network. *Research Policy*, 24(2):233–257.

Christensen, C. M., Suárez, F. F., and Utterback, J. M. (1998). Strategies for Survival in Fast-Changing Industries. *Management Science*, 44(12-part-2):S207–S220.

Cockburn, I., Henderson, R., and Stern, S. (2018). The Impact of Artificial Intelligence on Innovation. Technical Report w24449, National Bureau of Economic Research, Cambridge, MA.

Cohan, A., Ammar, W., Van Zuylen, M., and Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35(1):128.

Cole, S. (1970). Professional Standing and the Reception of Scientific Discoveries. *American Journal of*

Sociology, 76(2):286–306.

Crouch, T. D. (2003). *The bishop's boys: a life of Wilbur and Orville Wright*. Norton, New York London, reissued as a norton paperback edition.

David, P. (1985). Clio and the Economics of QWERTY. *American Economic Review*, 75(2):332–337.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. (2021). The Benchmark Lottery. arXiv:2107.07002 [cs].

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Dosi, G. (1982). Technological paradigms and technological trajectories. *Research Policy*, 11(3):147–162.

Eggers, J. and Green, E. (2012). Choosing Not to Choose: A Behavioral Perspective on Parallel Search. *Academy of Management Proceedings*, 2012(1):17579.

Eggers, J. P. and Kaplan, S. (2009). Cognition and Renewal: Comparing CEO and Organizational Effects on Incumbent Adaptation to Technical Change. *Organization Science*, 20(2):461–477.

Eggers, J. P. and Kaplan, S. (2013). Cognition and Capabilities: A Multi-Level Perspective. *Academy of Management Annals*, 7(1):295–340.

Eggers, J. P. and Park, K. F. (2018). Incumbent Adaptation to Technological Change: The Past, Present, and Future of Research on Heterogeneous Incumbent Response. *Academy of Management Annals*, 12(1):357–389.

Erdogan, B., Kant, R., Miller, A., and Sprague, K. (2016). Grow fast or die slow: Why unicorns are staying private.

Ewens, M. and Farre-Mensa, J. (2022). Private or Public Equity? The Evolving Entrepreneurial Finance Landscape. *Annual Review of Financial Economics*, 14(1):271–293.

Ewens, M., Nanda, R., and Rhodes-Kropf, M. (2018). Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics*, 128(3):422–442.

Ewens, M., Xiao, K., and Xu, T. (2024). Regulatory costs of being public: Evidence from bunching estimation. *Journal of Financial Economics*, 153:103775.

Farbmacher, H. and Tauchmann, H. (2023). Linear fixed-effects estimation with nonrepeated outcomes. *Econometric Reviews*, 42(8):635–654.

Felin, T., Gambardella, A., and Zenger, T. R. (2020). Value Lab: A Tool for Entrepreneurial Strategy. *SSRN Electronic Journal*.

Felin, T. and Zenger, T. R. (2017). The Theory-Based View: Economic Actors as Theorists. *Strategy Science*, 2(4):258–271.

Feyerabend, P. (1962). *Explanation, reduction, and empiricism*.

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Management Science*, 47(1):117–132.

Fosfuri, A., Lanzolla, G., and Suarez, F. F. (2013). Entry-Timing Strategies: The Road Ahead. *Long Range Planning*, 46(4-5):300–311.

Foster, R. N. (1986). *Innovation: the attacker's advantage*. Summit Books, New York; NY.

Franco, A. M., Sarkar, M., Agarwal, R., and Echambadi, R. (2009). Swift and Smart: The Moderating Effects of Technological Capabilities on the Market Pioneering–Firm Survival Relationship. *Management Science*, 55(11):1842–1860.

Gallagher, L. (2017). *The Airbnb story: how three ordinary guys disrupted an industry, made billions... and created plenty of controversy*. Houghton Mifflin Harcourt, Boston.

Gans, J. S., Hsu, D. H., and Stern, S. (2002). When Does Start-Up Innovation Spur the Gale of Creative Destruction? *The RAND Journal of Economics*, 33(4):571.

Gans, J. S., Stern, S., and Wu, J. (2019). Foundations of entrepreneurial strategy. *Strategic Management Journal*, 40(5):736–756.

Gofman, M. and Jin, Z. (2024). Artificial Intelligence, Education, and Entrepreneurship. *The Journal of*

Finance, 79(1):631–667.

Gompers, P. A. and Lerner, J. (2001). *The money of invention: how venture capital creates new wealth*. Harvard Business School Press, Boston.

Graffin, S. D. and Ward, A. J. (2010). Certifications and Reputation: Determining the Standard of Desirability Amidst Uncertainty. *Organization Science*, 21(2):331–346.

Greenberg, J., Sands, D. B., Cattani, G., and Porac, J. (2024). Rating systems and increased heterogeneity in firm performance: Evidence from the New York City Restaurant Industry, 1994–2013. *Strategic Management Journal*, 45(1):36–65.

Greenstein, S. (2024). Archetypes of product launch by insiders, outsiders, and visionaries. *Industrial and Corporate Change*, 33(1):216–237.

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *Am. Econ. J. Appl. Econ.*, 8(3):1–30.

Hanks, S., Pollack, M., and Cohen, P. (1993). Benchmarks, Test Beds, Controlled Experimentation, and the Design of Agent Architectures. *AI Magazine*, 4(4).

Hayek, F. A. (1945). The Use of Knowledge in Society. *The American Economic Review*, 35(4):519–530.

Henderson, R. M. and Clark, K. B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly*, 35(1):9.

Hill, R. and Stein, C. (2024). Scooped! Estimating Rewards for Priority in Science.

Hinton, G. E. and Anderson, J. A., editors (1989). *Parallel models of associative memory*. The Cognitive science series. L. Erlbaum, Hillsdale, N.J, updated ed edition.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554.

Hong, H. and Stein, J. C. (2007). Disagreement and the Stock Market. *Journal of Economic Perspectives*, 21(2):109–128.

Hooker, S. (2020). The Hardware Lottery. arXiv:2009.06489 [cs].

Howard, F. (1998). *Wilbur and Orville: a biography of the Wright brothers*. Dover Publications, Mineola, N.Y.

Howell, S. T. (2020). Reducing information frictions in venture capital: The role of new venture competitions. *Journal of Financial Economics*, 136(3):676–694.

Iansiti, M. and Lakhani, K. R. (2020). *Competing in the age of AI: strategy and leadership when algorithms and networks run the world*. Harvard Business Review Press, Boston, MA.

Isenberg, D. J. (2013). *Worthless, impossible, and stupid: how contrarian entrepreneurs create and capture extraordinary value*. Harvard Business Review Press, Boston, Massachusetts.

Jeppesen, L. B. and Lakhani, K. R. (2010). Marginality and Problem-Solving Effectiveness in Broadcast Search. *Organization Science*, 21(5):1016–1033.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

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

Jurowetzki, R., Hain, D., Mateos-Garcia, J., and Stathoulopoulos, K. (2021). The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain? arXiv:2102.01648 [cs].

Kaplan, S. and Tripsas, M. (2008). Thinking about technology: Applying a cognitive lens to technical change. *Research Policy*, 37(5):790–805.

Karargyris, A., Umeton, R., Sheller, M. J., Aristizabal, A., et al. (2023). Federated benchmarking of medical artificial intelligence with MedPerf. *Nature Machine Intelligence*, 5(7):799–810.

Kardas, M., Czapla, P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., and Stojnic, R. (2020). AxCell: Automatic Extraction of Results from Machine Learning Papers. arXiv:2004.14356 [cs, stat].

Karikó, K. (2023). *Breaking through: my life in science*. Crown, New York, first edition edition.

Kerr, W. R., Nanda, R., and Rhodes-Kropf, M. (2014). Entrepreneurship as Experimentation. *Journal of Economic Perspectives*, 28(3):25–48.

Kim, J. D. (2024). Startup Acquisitions as a Hiring Strategy: Turnover Differences Between Acquired and Regular Hires. *Strategy Science*, 9(2):118–134.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S., Gorney, J., Graham, D., Hu, F., Huff, R., King, D., Kohlmeier, S., Kuehl, B., Langan, M., Lin, D., Liu, H., Lo, K., Lochner, J., MacMillan, K., Murray, T., Newell, C., Rao, S., Rohatgi, S., Sayre, P., Shen, Z., Singh, A., Soldaini, L., Subramanian, S., Tanaka, A., Wade, A. D., Wagner, L., Wang, L. L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., Van Zuylen, M., and Weld, D. S. (2023). The Semantic Scholar Open Data Platform. arXiv:2301.10140 [cs].

Kirzner, I. (1997). Entrepreneurial Discovery and the Competitive Market Process: An Austrian Approach. *Journal of Economic Literature*, 35(1):60–85.

Klepper, S. and Thompson, P. (2010). Disagreements and intra-industry spinoffs. *International Journal of Industrial Organization*, 28(5):526–538.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Signalman, Kissimmee, Fla.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. arXiv:2112.01716 [cs, stat].

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37.

Kovacs, A., Marullo, C., Verhoeven, D., and Van Looy, B. (2019). Radical, Disruptive, Discontinuous and Breakthrough Innovation: more of the same? *Academy of Management Proceedings*, 2019(1):14866.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kuhn, T. S. (1956). *The Copernican revolution: planetary astronomy in the development of western thought*. Harvard Univ. Pr, Cambridge, Mass., 19. print edition.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press, Chicago, IL, 3rd ed edition.

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., and Lakhani, K. R. (2022). Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation. *Management Science*, 68(6):4478–4495.

Langley, P. (1988). Machine learning as an experimental science. *Machine Learning*, 3(1):5–8.

Leech, G., Vazquez, J. J., Yagudin, M., Kupper, N., and Aitchison, L. (2024). Questionable practices in machine learning. arXiv:2407.12220 [cs].

Li, F. F. (2023). *The worlds I see: curiosity, exploration, and discovery at the dawn of AI*. Moment of Lift Books ; Flatiron Books, New York, first edition edition.

Lieberman, M. B. and Montgomery, D. B. (1988). First-mover advantages. *Strategic Management Journal*, 9(S1):41–58.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., Van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H.,

Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., Van Der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217.

Malik, J. (2017). Technical Perspective: What led computer vision to deep learning? *Communications of the ACM*, 60(6):82–83.

Marot, A., Rousseau, D., and Xu, Z. (2023). Ai competitions and benchmarks: towards impactful challenges with post-challenge papers, benchmarks and other dissemination actions.

Martínez-Plumed, F., Barredo, P., hÉigeartaigh, S. , and Hernández-Orallo, J. (2021). Research community dynamics behind popular AI benchmarks. *Nature Machine Intelligence*, 3(7):581–589.

Marx, M. and Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594.

Marx, M. and Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392.

Mastrobuoni, G. and Pinotti, P. (2015). Legal status and the criminal activity of immigrants. *American Economic Journal: Applied Economics*, 7(2):175–206.

McGrath, R. G. (2001). Exploratory learning, innovative capacity, and managerial oversight. *Academy of Management Journal*, 44(1):118–131.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors (1984). *Machine Learning An Artificial Intelligence Approach*. Springer, Berlin, 1983 edition. OCLC: 1259418305.

Miric, M., Jia, N., and Huang, K. G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, 44(2):491–519.

Morris, S. (1995). The Common Prior Assumption in Economic Theory. *Economics and Philosophy*, 11(2):227–253.

Moskowitz, T. J. and Vissing-Jørgensen, A. (2002). The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle? *American Economic Review*, 92(4):745–778.

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts.

Murray, F., Stern, S., Campbell, G., and MacCormack, A. (2012). Grand Innovation Prizes: A theoretical, normative, and empirical evaluation. *Research Policy*, 41(10):1779–1792.

Myers, K. (2020). The Elasticity of Science. *American Economic Journal: Applied Economics*, 12(4):103–134.

Myers, K. R., Tham, W. Y., Thursby, J., Thursby, M., Cohodes, N., Lakhani, K., Mural, R., and Xu, Y. (2023). New Facts and Data about Professors and their Research. arXiv:2312.01442 [econ, q-fin].

Nanda, R. and Rhodes-Kropf, M. (2016). Financing Entrepreneurial Experimentation. *Innovation Policy and the Economy*, 16:1–23.

Niven, T. and Kao, H.-Y. (2019). Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Oh, K.-S. and Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314.

Perkmann, M., Salandra, R., Tartari, V., McKelvey, M., and Hughes, A. (2021). Academic engagement: A review of the literature 2011-2019. *Research Policy*, 50(1):104114.

Prechelt, L. (1994). PROBEN 1 - a set of benchmarks and benchmarking rules for neural network training algorithms. *Technical Report*.

Price, K. (1986). Anything you can do, I can do better (No you can't)... *Computer Vision, Graphics, and Image Processing*, 36(2-3):387–391.

Raffaelli, R., Glynn, M. A., and Tushman, M. (2019). Frame flexibility: The role of cognitive and emotional framing in innovation adoption by incumbent firms. *Strategic Management Journal*, 40(7):1013–1039.

Rao, H. (1994). The Social Construction of Reputation: Certification Contests, Legitimation, and the Survival of Organizations in the American Automobile Industry: 1895–1912. *Strategic Management Journal*, 15(S1):29–44.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? arXiv:1902.10811 [cs, stat].

Reschke, B. P., Azoulay, P., and Stuart, T. E. (2018). Status Spillovers: The Effect of Status-conferring Prizes on the Allocation of Attention. *Administrative Science Quarterly*, 63(4):819–847.

Rosenberg, N. (1992). Economic experiments. *Industrial and Corporate Change*, 1(1):181–203.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sahal, D. (1981). *Patterns of technological innovation*. Addison-Wesley Pub. Co., Advanced Book Program/World Science Division, Reading, Mass.

Sands, D. B. (2024). Double-edged stars: Michelin stars, reactivity, and restaurant exits in new york city. *Strategic Management Journal*.

Scherer, F. and Harhoff, D. (2000). Technology policy for a world of skew-distributed outcomes. *Research Policy*, 29(4–5):559–566.

Schilling, M. A. (2002). Technology success and failure in winner–take–all markets: the impact of learning orientation, timing, and network externalities. *Academy of Management Journal*, 45(2):387–398.

Schlangen, D. (2020). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. arXiv:2007.04792 [cs].

Schumpeter, J. A. (1934). *The Theory of Economic Development: An Inquiry into Profits, Capital Credit, Interest, and the Business Cycle*. Harvard University Press.

Scikit (2024). Scikit User Guide.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J., and Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Neurips*.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Shane, S. (2000). Prior Knowledge and the Discovery of Entrepreneurial Opportunities. *Organization Science*, 11(4):448–469.

Shane, S. and Venkataraman, S. (2000). The Promise of Enterpreneurship as a Field of Research. *The Academy of Management Review*, 25(1):217.

Shelef, O., Wuebker, R., and Barney, J. B. (2024). Heisenberg effects in experiments on business ideas. *Academy of Management Review*.

Sherry, Y. and Thompson, N. C. (2021). How Fast Do Algorithms Improve? [Point of View]. *Proceedings of the IEEE*, 109(11):1768–1777.

Simcoe, T. S. and Waguespack, D. M. (2011). Status, Quality, and Attention: What's in a (Missing) Name? *Management Science*, 57(2):274–290.

Simon, H. A. (1995). Artificial intelligence: an empirical science. *Artificial Intelligence*, 77(1):95–127.

Sine, W. D., David, R. J., and Mitsuhashi, H. (2007). From Plan to Plant: Effects of Certification on Operational Start-up in the Emergent Independent Power Sector. *Organization Science*, 18(4):578–594.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15.

Stevenson, H. (1983). A Perspective On Entrepreneurship. *Harvard Business School Working Paper #9-384-131*.

Stinchcombe, A. L. (1965). Social structure and organizations.

Stokes, D. E. (1997). *Pasteur's quadrant: basic science and technological innovation*. Brookings Institution Press, Washington, D.C.

Stolovitzky, G., Saez-Rodriguez, J., Bletz, J., Albrecht, J., Andreoletti, G., Costello, J. C., and Boutros, P. (2023). Ai competitions and benchmarks: The life cycle of challenges and benchmarks.

Suarez, F. F. (2004). Battles for technological dominance: an integrative framework. *Research Policy*, 33(2):271–286.

Suarez, F. F. and Lanzolla, G. (2007). The Role of Environmental Dynamics in Building a First Mover Advantage Theory. *Academy of Management Review*, 32(2):377–392.

Techcrunch (2013). Google Scoops Up Neural Networks Startup DNNresearch To Boost Its Voice And Image Search Tech.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6):285–305.

Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., and Hengel, A. v. d. (2020). On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law. arXiv:2005.09241 [cs].

Terviö, M. (2009). Superstars and mediocrities: Market failure in the discovery of talent1. *Review of Economic Studies*, 76(2):829–850.

Thiel, P. and Masters, B. (2014). *Zero to one: notes on startups, or how to build the future*. Crown Business, New York, NY, 1. ed edition.

Torralba, A., Fergus, R., and Freeman, W. (2008). 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970.

Tripsas, M. and Gavetti, G. (2000). Capabilities, cognition, and inertia: evidence from digital imaging. *Strategic Management Journal*, 21(10-11):1147–1161.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596.

Tushman, M. L. and Anderson, P. (1986). Technological Discontinuities and Organizational Environments. *Administrative Science Quarterly*, 31(3):439.

Utterback, J. M. (1994). *Mastering the Dynamics of Innovation*. Harvard Business School Press.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, 342(6157):468–472.

Van den Steen, E. (2001). *Essays on the Managerial Implications of Differing Priors*. Phd thesis, Stanford University, Stanford, CA.

Van den Steen, E. (2004). Rational Overoptimism (and Other Biases). *American Economic Review*, 94(4):1141–1151.

Van den Steen, E. (2005). Organizational Beliefs and Managerial Vision. *Journal of Law, Economics, and Organization*, 21(1):256–283.

Varoquaux, G. and Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures

and recommendations for the future. *npj Digital Medicine*, 5(1):48.

Wade, J. B., Porac, J. F., Pollock, T. G., and Graffin, S. D. (2006). The Burden of Celebrity: the Impact of Ceo Certification Contests on Ceo Pay and Performance. *Academy of Management Journal*, 49(4):643–660.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Williams, H. (2012). Innovation inducement prizes: Connecting research to policy. *Journal of Policy Analysis and Management*, 31(3):752–776.

Wired (2023). Nvidia Chip Shortages Leave AI Startups Scrambling for Computing Power.

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Wright, P. M., Dunford, B. B., and Snell, S. A. (2001). Human resources and the resource based view of the firm. *Journal of Management*, 27(6):701–721.

Wu, J. (2023). Measurement for Dummies? Exploring the Role of Policy-Driven Measurement in Automotive Safety. *Working Paper*.

Wuebker, R., Zenger, T., and Felin, T. (2023). The theory-based view: Entrepreneurial microfoundations, resources, and choices. *Strategic Management Journal*, 44(12):2922–2949.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. (2023). Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv:2311.01964 [cs].

Zuckerman, E. W. (1999). The Categorical Imperative: Securities Analysts and the Illegitimacy Discount. *American Journal of Sociology*, 104(5):1398–1438.

Zuckerman, E. W. (2012). Construction, Concentration, and (Dis)Continuities in Social Valuations. *Annual Review of Sociology*, 38(1):223–245.

Zuzul, T. and Edmondson, A. C. (2017). The advocacy trap: When legitimacy building inhibits organizational learning. *Academy of Management Discoveries*, 3(3):302–321.

Åstebro, T., Herz, H., Nanda, R., and Weber, R. A. (2014). Seeking the Roots of Entrepreneurship: Insights from Behavioral Economics. *Journal of Economic Perspectives*, 28(3):49–70.

Main Figures and Tables

Figure 1: Overview of Data.

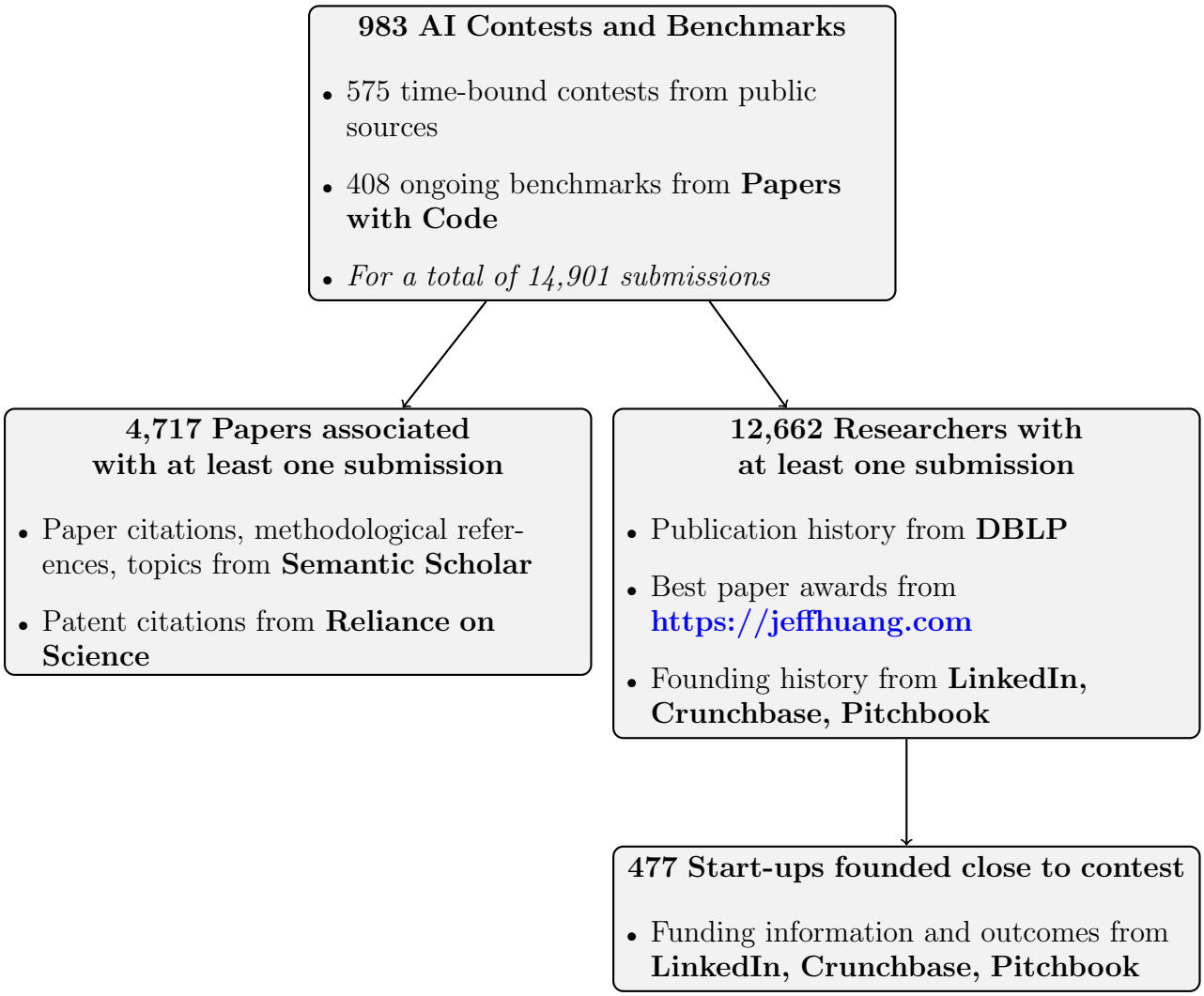
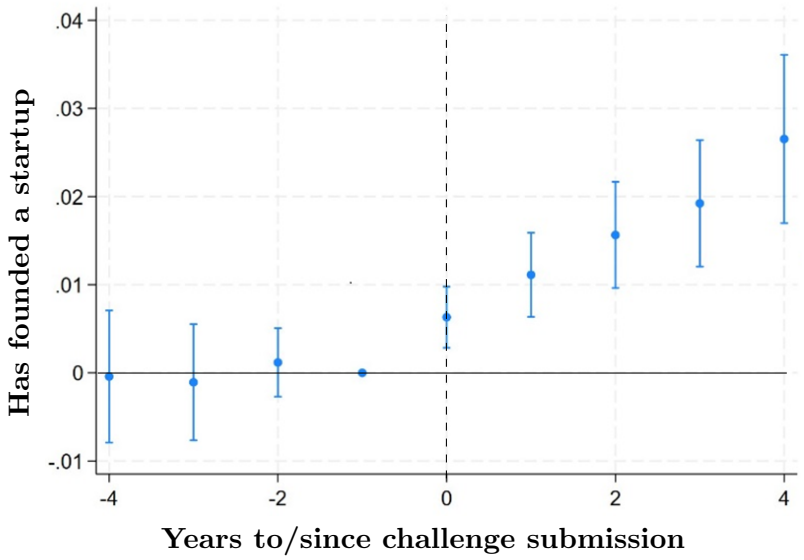


Figure 2: Impact of winning a challenge on founding a start-up

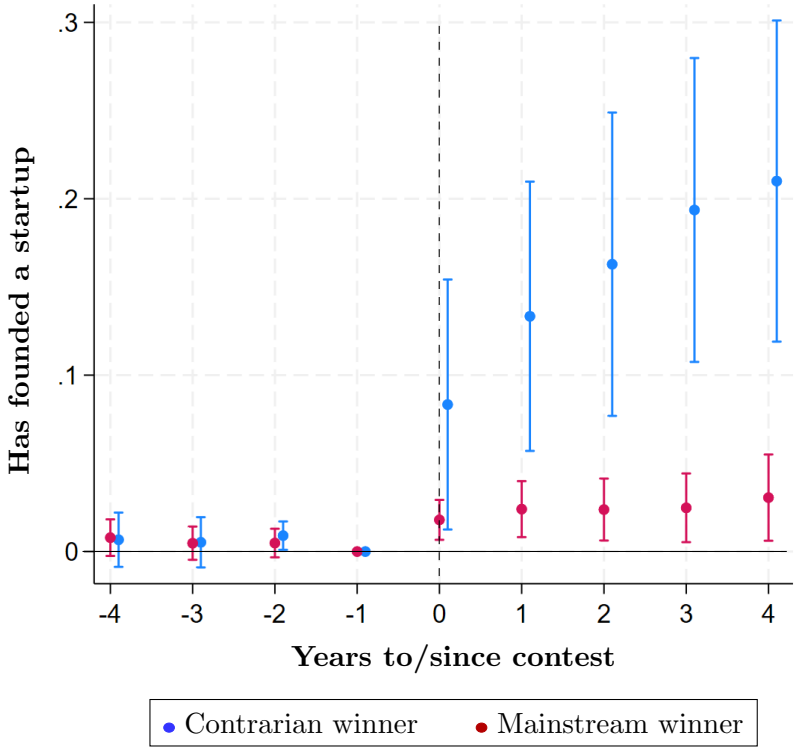


Note: The figure shows the event study coefficients τ_j estimated from the following specification:

$$Y_{it} = \sum_{j \neq -1} \tau_j^w \; D_{i,\, t-j} \quad \times \; \text{Winner}_i + \sum_{j \neq -1} \tau_j^r \; D_{i,\, t-j} \; + \beta_i + \gamma_c + \alpha_t$$

Y_{it} is 1 if researcher i has founded a startup on or before year t and 0 otherwise. The chart plots τ_j^w, the difference in the propensity to have founded a startup for winners relative to runner-ups, before and after the the researcher's first challenge submission (including both time-bound contests and ongoing benchmarks). Controls include researcher i, challenge c , and year t fixed effects. Standard errors are clustered by researcher and challenge.

Figure 3: Impact of winning a contest on founding a start-up: contrarian vs. mainstream approaches

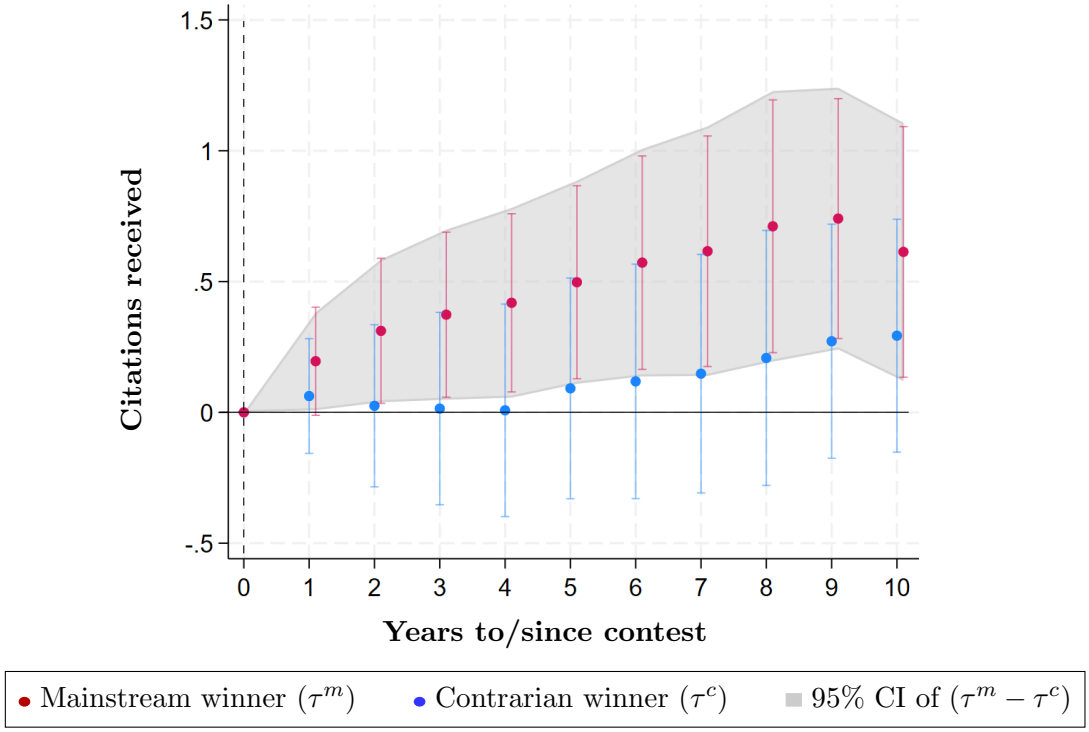


Note: The figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following specification:

$$Y_{it} = \sum_{j\neq -1} \tau_j^m \; D_{i,\; t-j} \times \text{Mainstream winner}_i + \sum_{j\neq -1} \tau_j^c \; D_{i,\; t-j} \times \text{Contrarian winner}_i + \\ + \sum_{j\neq -1} \tau_j^r \; D_{i,\; t-j} \; + \beta_i + \gamma_c + \alpha_t$$

The dependent variable is 1 if researcher i has founded a startup on or before year t and 0 otherwise. The chart plots τ_j^c (τ_j^m), the difference in the propensity to have founded a startup for contrarian (mainstream) winners relative to runner-ups, before and after the researcher's first time-bound contest (excluding ongoing benchmarks). Controls include researcher, contest, and year fixed effects. Standard errors are clustered by researcher and contest.

Figure 4: Impact of winning a challenge on citations: contrarian vs. mainstream approaches



Note: The figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following Poisson quasi–maximum likelihood specification:

$$Y_{it} = \sum_{j>0} \tau_j^m \; D_{i,\,t-j} \;\times\; \text{Mainstream winner}_i + \sum_{j>0} \tau_j^c \; D_{i,\,t-j} \;\times\; \text{Contrarian winner}_i + \\ + \sum_{j>0} \tau_j^r \; D_{i,\,t-j} + \beta_i + \gamma_c + \alpha_t$$

where Y_{it} represents the cumulative number of citations received by the paper describing challenge submission i as of year t. The chart plots τ_j^c (τ_j^m), the difference in cumulative citations accruing to contrarian (mainstream) winning submissions relative to runner-up submissions. The figure shows that contrarian winners are cited (i) as much as runner-up papers in the first 4 years (ii) less than mainstream winners in the first 10 years after the evaluation challenge (including contests and ongoing benchmarks).
The gray area plots the 95% confidence interval for the vertical difference between red and blue dots, $\tau_j^m - \tau_j^c$: the difference is statistically significant whenever the blue dot is outside the gray area (i.e., up until year 7 after the challenge submission).
To mitigate truncation concerns, I only include submissions whose paper-pair has at least 5 years of citation data. I also include challenge, paper, and year fixed effects. Standard errors are clustered by paper and challenge. No pre-trends are available because the papers detailing the winning and runner-up entries did not exist prior to the challenge submission.

Table 1: Impact of Winning on Academic Recognition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Is published	Best paper award	Paper citations	Patent citations	Best paper award	Paper citations	Patent citations	Best paper award	Paper citations	Patent citations
Winner	0.121***	0.003**	0.797***	0.479**	0.006+	0.528***	0.207*	0.004	0.484**	0.386***
	(0.020)	(0.001)	(0.159)	(0.167)	(0.003)	(0.125)	(0.102)	(0.004)	(0.155)	(0.109)
Novel		0.002*	2.634***	2.120***	0.004	2.213***	1.339***	0.004	2.213***	1.355***
		(0.001)	(0.173)	(0.154)	(0.003)	(0.169)	(0.151)	(0.003)	(0.151)	(0.172)
Gap								0.015	-0.088	-0.780*
								(0.016)	(0.206)	(0.326)
Winner × Gap								-0.023	0.554	0.410
								(0.042)	(0.380)	(0.606)
Constant	0.525***	-0.001*	4.761***	0.335*	0.003	5.883***	1.237***	0.006	5.864***	1.055***
	(0.010)	(0.000)	(0.142)	(0.155)	(0.002)	(0.144)	(0.164)	(0.006)	(0.156)	(0.183)
Observations	1,140	4,695	4,717	4,304	14,128	14,220	12,647	14,074	14,166	12,617
Challenge FE	YES	NO	NO	NO	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Units of obs.	Submission	Paper	Paper	Paper	Submission	Submission	Submission	Submission	Submission	Submission
Model	OLS	OLS	QMLE-Poisson	QMLE-Poisson	OLS	QMLE-Poisson	QMLE-Poisson	OLS	QMLE-Poisson	QMLE-Poisson

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. This table shows that winning contributions are more likely to be published and – when published – receive more citations (both in academic papers and patents) and best paper awards. Column (1) only includes submissions to time-bound contests, since virtually all benchmark submissions are published by construction. In columns (2-4), the unit of observation is a paper, with winners defined as papers associated with (at least one) winning submission. Columns (5-10) analyze contest or benchmark submissions with an associated paper pair (where a paper can in principle describe multiple submissions by the same authors to different benchmarks). All columns control for the year in which the contest or benchmark entry was submitted (and the corresponding paper pair was written). Columns (1) and (5-10) also include challenge fixed effects, directly comparing winners to runner-ups within the same challenge. Columns (2-10) control for novelty (which instead is missing for unpublished contributions). Columns (8-10) control for the % difference in performance between the focal submission's result and the minimum result required to win (i.e., the second-place result for winners and the first-place result for runners-up). I winsorize both tails at (5, 95)% levels to remove outliers. Standard errors are clustered by submission and challenge (col. 1), by paper (cols. 2-4), and by paper and challenge (cols 5-10).

Table 2: Impact of Winning on Startup Founding

VARIABLES	(1) Has founded a startup	(2) Has founded a startup	(3) Has founded a startup	(4) Has founded a startup
Winner × Post	0.023***	0.023***	0.018+	0.016**
	(0.005)	(0.003)	(0.010)	(0.005)
Novel × Post		0.001		
		(0.005)		
Winner × Post × Contest				0.037*
				(0.010)
Constant	0.025**	0.027**	0.008	0.023*
	(0.009)	(0.009)	(0.012)	(0.009)
Observations	41,707	41,588	8,490	41,707
Researcher FE	YES	YES	YES	YES
Challenge FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Units of observation	Researcher × Challenge × Pre/Post	Researcher × Challenge × Pre/Post	Researcher × Challenge × Pre/Post	Researcher × Challenge × Pre/Post
Specification	Diff-in-diff	Diff-in-diff	Diff-in-discontinuity	Diff-in-diff

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. The dependent variable is 1 if the researcher has founded a startup on or before the focal year, 0 otherwise. Researchers are observed in two periods: one before the submission, and four years after the submission. Standard errors are clustered by researcher and challenge.
The table presents difference-in-difference estimates of the impact of winning a time-bound contest (or achieving a state-of-the-art result in an ongoing benchmark) on the likelihood of becoming an entrepreneur. Column (2) controls for novelty. Column (3) shows the results of a difference in discontinuity design. The model is specified as

$$Y = \delta_0+\delta_1\text{Gap}+\text{Win}\times(\gamma_0+\gamma_1\text{Gap})+\text{Post}[\alpha_0+\alpha_1\text{Gap}+\text{Win}\times(\beta_W+\beta_1\text{Gap})]+\text{Controls} \qquad \text{for Gap} \in [-h, h]$$

where the running variable Gap represents the (scaled) distance between the focal submission's result and the minimum result required to win (i.e., the second-place result for winners and the first-place result for runners-up). The bandwidth h is selected using the procedure described in Calonico et al. [2014]. β_W is the coefficient of interest and the others are omitted for brevity.
To conclude, column (4) explores heterogeneous effects of winning between contests and ongoing benchmarks.

Table 3: Impact of Winning on Contrarian Entrepreneurship

VARIABLES	(1) Has founded a startup	(2) Has founded a startup	(3) Has founded a startup	(4) Has founded a startup
Mainstream Win × Post	0.021***	0.021***	0.009	0.018**
	(0.005)	(0.005)	(0.009)	(0.006)
Contrarian Win × Post	0.036*	0.035*	0.129*	0.002
	(0.014)	(0.014)	(0.060)	(0.014)
Mainstream Win × Post × Contest				0.019
				(0.015)
Contrarian Win × Post × Contest				0.201***
				(0.047)
Novel × Post		0.001		
		(0.005)		
Constant	0.025**	0.028**	0.023***	0.024**
	(0.009)	(0.009)	(0.003)	(0.009)
Observations	41,500	41,385	8,401	41,500
Researcher FE	YES	YES	YES	YES
Challenge FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Specification	Diff-in-diff	Diff-in-diff	Diff-in-discontinuity	Diff-in-diff

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Dependent variable is 1 if the researcher has founded a startup on or before the focal year, 0 otherwise. Researchers are observed in two periods: one before the submission, and four years after the submission. Standard errors clustered by researcher and challenge. Column (1) presents difference-in-difference estimates of the impact of winning a time-bound contest (or achieving a state-of-the-art result in an ongoing benchmark) on the likelihood of becoming an entrepreneur, separately for contrarian and mainstream winners. Column (2) controls for novelty. Column (3) shows the results of a difference in discontinuity design. The model is specified as

$$Y = \delta_0 + \delta_1 \text{Gap} + \sum_{i \in C,M} \text{Win}_\text{i} \times (\gamma_0^i + \gamma_1^i \text{Gap}) + \text{Post}[\alpha_0 + \alpha_1 \text{Gap} + \sum_{i \in C,M} \text{Win}_\text{i} \times (\beta_W^i + \beta_1^i \text{Gap})] + \text{Controls}$$

where the running variable $\text{Gap} \in [-h, h]$ represents the (scaled) distance between the focal submission's result and the minimum result required to win (i.e., the second-place result for winners and the first-place result for runners-up). The bandwidth h is selected using the procedure described in Calonico et al. [2014]. β_W^C, β_W^M are the causal effect of winning on entrepreneurship respectively for contrarians and mainstreams. The other coefficients are omitted for brevity.
To conclude, column (4) shows differences in the estimates between time-bound contests and ongoing benchmarks.

Table 4: Startups Founded by Contrarian Winners Are Disproportionately Successful

VARIABLES	(1) Venture capital raised	(2) Log Valuation	(3) IPO or unicorn	(4) Acquired	(5) Years btw. founding and acquisition	(6) Has ≥ 1 patents	(7) Log competitors at founding	(8) Startup is backed by government
Contrarian winner	481.294+	1.196*	0.097+	-0.006	-1.759*	-0.002	1.332***	-0.027**
	(289.220)	(0.519)	(0.051)	(0.055)	(0.826)	(0.058)	(0.359)	(0.009)
Runner-up	-40.725	-0.109	-0.015	-0.106**	1.009	-0.022	0.377	-0.020*
	(41.099)	(0.430)	(0.026)	(0.034)	(0.871)	(0.046)	(0.359)	(0.009)
Mainstream winner	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
Constant	100.088**	4.141***	0.078***	0.206***	3.866***	0.280***	6.228***	0.041***
	(36.375)	(0.280)	(0.018)	(0.025)	(0.434)	(0.028)	(0.232)	(0.007)
Observations	440	160	472	472	71	458	221	3,961
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Units of observation	Start-up	Start-up	Start-up	Start-up	Start-up	Start-up	Start-up	Start-up x funding round

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Standard errors are robust in columns (1-7) and clustered by startup in column 8. The table shows that startups founded by contrarian winners are disproportionately successful across several dimensions, including IPO likelihood, valuation, and venture capital raised. Contrarian winners and runner-ups are compared to the omitted baseline of mainstream winners.

Appendix A: Modeling Contrariness, Herding, and Entrepreneurial Experimentation

This section provides a theoretical framework to explain the liability of contrariness. Specifically, the model aims to address the following points: (i) why contrarian entrepreneurs, compared to mainstream ones, face greater difficulty in attracting resources for commercialization, (ii) why public experiments, as opposed to private signals, are particularly effective in validating contrarian hypotheses, and (iii) why winning a contest serves as a more salient public signal than non-competitive public demonstrations.

I proceed as follows: First, I define a contrarian as someone who challenges the conventional viewpoint, i.e., a mainstream hypothesis that all investors conform to. I develop a simple model of herding, building on Bikhchandani et al. [1992] and Chamley [2003], in which a contrarian entrepreneur must persuade at least one investor to support the startup, but each investor's decision is influenced by the actions of others. Under certain conditions, only a salient public signal can disrupt the information cascade and convince an investor that the contrarian hypothesis is valid. Second, I examine the boundary conditions under which public signals are most effective. I argue that the most persuasive signals are those that sharply differentiate the mainstream hypothesis from the contrarian one. Experiments that blur this distinction by conflating multiple aspects of a venture opportunity can dilute the signal's impact, perpetuating herding behavior and reinforcing the mainstream viewpoint's dominance.

Contrariness and Herding

There are two potential states of the world, $s \in S = \{m, c\}$. When $s = m$, the mainstream view is valid; when $s = c$, instead, the contrarian theory is correct.

A contrarian entrepreneur is fully persuaded that $s = c$. In order to put the theory in practice, however, he needs to persuade an investor to support his venture. A finite number of investors are indexed by integer t. They take action in exogenous order: in each period, one and only one agent makes her investment decision. For simplicity, each agent takes an action a_t in $A \in \{m, c\}$, meaning that she invests either in the mainstream idea ($a_t = m$) or in the contrarian one ($a_t = c$). The action yields utility $u(a, s) = \mathbb{I}\{a_t = s\}$, i.e. it is 1 if the action matches the right entrepreneurial theory, 0 otherwise.

Each investor bases her decision on the public actions of the preceding investors as well on a private signal. Her private signal is a symmetric binary signal $\sigma_t \in \{s, \neg s\}$ with precision $q > 1/2 : P(\sigma_t = s|s) = q$. The prior investor decisions $h = (a_1, ..., a_{t-1})$ are used to form a prior, public belief on the likelihood that the contrarian hypothesis is correct: $\mu_t = P(s = c|h_t)$. I assume that, ex-ante, this likelihood is low: $\mu_1 \ll 1/2$.

After observing the private signal, investor t has posterior beliefs μ_t^+ if she observes a signal $\sigma = c$, and beliefs μ_t^- if $\sigma = m$, with $\mu_t^- < \mu_t < \mu_t^+$.[42] The investor supports the contrarian venture if her posterior is higher than $1/2$.

Analyzing the first investor's decision, it is easy to see that even a positive signal will be insufficient to lead to a contrarian investment if $\mu_1^+ < 1/2$, i.e. if $\mu_1 < 1 - q$: if the conventional

[42] $\mu_t^+ = \frac{\mu_t q}{\mu_t q + (1-\mu_t)(1-q)}$; $\mu_t^- = \frac{\mu_t (1-q)}{\mu_t (1-q) + (1-\mu_t) q}$.

viewpoint is that the hypothesis is sufficiently likely to be wrong, even a very precise private signal will not persuade the investor to support the contrarian.

The fact that the private signal is too weak to counterbalance the public belief generates herding. Crucially, herding prevents the public belief from updating: the following investors will only observe the decision of the investor, not her private signal. Given that the first investor would have invested in the mainstream startup whatever signal, $\mu_2 = \mu_1$. This again results in a rejection of the contrarian hypothesis by the second investor no matter the private signal, and so on, in what has been described as an informational cascade (Bikhchandani et al. [1992]).

Proposition 1. *Given $\mu_1 > 0$ and $q > \frac{1}{2}$, if $\mu_1 < 1 - q$, no amount of consecutive, positive private signals will result in an investment in the contrarian venture.*

If the signal observed by the investor is *public*, it influences the decisions of future investors because it becomes part of common knowledge. Therefore, even if the initial belief $\mu_1 < 1 - q$ and the first investor favors the mainstream hypothesis despite a positive signal, after a finite number of consecutive public signals T, the updated belief μ_t will exceed $\frac{1}{2}$. This guarantees that investor T will support the contrarian venture. In other words, public information leads to public learning, breaking the informational cascade and strengthening the belief in the contrarian hypothesis. This is summarized in the next proposition.

Proposition 2. *Given $\mu_1 > 0$ and $q > \frac{1}{2}$, after T consecutive public signals, investors are guaranteed to invest in the contrarian venture.*

Proof.

We start with the belief update formula when signals are public:

$$\mu_{t+1} = \frac{\mu_t q}{\mu_t q + (1 - \mu_t)(1 - q)}$$

To simplify, we use the odds ratio $L_t = \frac{\mu_t}{1 - \mu_t}$. The update rule for the odds ratio becomes:

$$L_{t+1} = L_t \left(\frac{q}{1 - q} \right)$$

This shows that the odds ratio updates multiplicatively by a factor of $\frac{q}{1-q} > 1$ since $q > \frac{1}{2}$. Therefore, after t updates, the odds ratio is:

$$L_t = L_1 \left(\frac{q}{1 - q} \right)^{t-1}$$

We want to find the minimum T such that $\mu_T > \frac{1}{2}$, which is equivalent to $L_T > 1$.

Starting from $L_1 = \frac{\mu_1}{1 - \mu_1}$, we have:

$$L_T = \frac{\mu_1}{1 - \mu_1} \left(\frac{q}{1 - q} \right)^{T-1} > 1$$

Solving for T:

$$\left(\frac{q}{1-q}\right)^{T-1} > \frac{1-\mu_1}{\mu_1}$$

Which yields

$$T > 1 + \frac{\ln\left(\frac{1-\mu_1}{\mu_1}\right)}{\ln\left(\frac{q}{1-q}\right)}$$

Therefore, after T consecutive public signals, where T satisfies the above inequality, the belief μ_T exceeds $\frac{1}{2}$. This guarantees that the investor will choose the contrarian venture.

The role of contests as public experiments

Most conventional model of entrepreneurial experimentation (e.g., Nanda and Rhodes-Kropf [2016]) focus on the overall value of the venture as the key source of uncertainty. In this framework in which ventures' quality is one-dimensional, it is unclear why contests are "special" relative to any other public signal: if the contrarian and mainstream strategies are mutually exclusive, any evidence that increases the chance that the contrarian venture is valuable, at the same time, raises μ, i.e. the belief that the contrarian theory is correct.

As argued by (Agrawal et al. [2021]), however, the value of a business opportunity can depend on multiple aspects. To accommodate this eventuality, I assume that instead of a signal on the *relative* value of the mainstream and contrarian venture, the investor observes a signal on the *overall* quality of the contrarian venture v.[43]

Specifically, I assume that $v(s, i)$ depends both on whether the contrarian strategy is correct ($s = c$) and on the quality of the idea $i \in I = \{g, b\}$. As such, $p_v = \mathfrak{B}(\mu, p_i)$ – the belief that the contrarian venture is valuable – depends on both the likelihood that the contrarian strategy is highly effective (state c) and the idea is good (state g).[44] Importantly, I assume that the same idea can be pursued through a contrarian or mainstream strategy, meaning that persuading an investor depends on raising not only p_v (e.g., above a fixed outside option), but also μ (i.e., making the contrarian venture relatively more appealing than the mainstream one).

As such, the binary signal $\{H, L\}$ on the value v will *conflate* information about s (which matters for the decision to invest in the contrarian venture, instead of the mainstream one) and i (which is the same for both ventures). A good signal H, in other words, bears good news on both the quality of the idea and the quality of the strategy.

[43] This, for example, could be a prototype built by the entrepreneur to reduce uncertainty before fully funding the startup (Nanda and Rhodes-Kropf [2016]).

[44] I assume that $\mathfrak{B}(\mu, p_i) : [0, 1] \times [0, 1] \to [0, 1]$ is a function weakly increasing in both arguments.

Table 5: Experiment with Bidimensional Quality

		Idea Value	
		g	**b**
Strategy Effectiveness	**c**	$\begin{cases} H \text{ with probability } \gamma(c,g) \\ L \text{ with probability } 1-\gamma(c,g) \end{cases}$	$\begin{cases} H \text{ with probability } \gamma(c,b) \\ L \text{ with probability } 1-\gamma(c,b) \end{cases}$
	m	$\begin{cases} H \text{ with probability } \gamma(m,g) \\ L \text{ with probability } 1-\gamma(m,g) \end{cases}$	$\begin{cases} H \text{ with probability } \gamma(m,b) \\ L \text{ with probability } 1-\gamma(m,b) \end{cases}$

Note: The table shows the joint distribution of strategy effectiveness (c or m), idea value (g or b), and experiment outcomes (H or L).

The parameters $\gamma(s,i) = P(H|s,i)$ are restricted so that the posterior on the quality of the idea and the quality of the strategy weakly rise after a positive experiment outcome: $\gamma(c,g) \geq \max[\gamma(m,g),\gamma(c,b)] \geq \min[\gamma(m,g),\gamma(c,b)] \geq \gamma(m,b)$. As in Agrawal et al. [2021], the signal could be relatively more informative about either aspect. For example, if $\gamma(c,b) = \gamma(m,b) = 0$, a positive outcome H would perfectly reveal that the idea is valuable (i.e., the state is (g,c) or (g,m)) but still leave the investor uncertain about the effectiveness of the contrarian strategy. By contrast, if $\gamma(c,g) = \gamma(c,b)$ and $\gamma(m,g) = \gamma(m,b)$ the experiment reveals nothing about idea quality.[45]

The previous setup can be recovered by assuming that $q_g \coloneqq \gamma(c,g) = 1-\gamma(m,g) > 1/2$ and $q_b \coloneqq \gamma(c,b) = 1-\gamma(m,b) > 1/2$. In other words, this ensures that when conditioning on idea quality the experiment is again a symmetric binary signal (whose precision depends on i). Further imposing $\gamma_g = \gamma_b$(i.e., $\gamma(c,g) = \gamma(c,b)$ and $\gamma(m,g) = \gamma(m,b)$) would mean that the signal is informative only about the strategy, i.e. the idea quality is irrelevant. If $\gamma_g \neq \gamma_b$ (i.e., $\gamma(c,g) > \gamma(c,b)$ or $\gamma(m,g) > \gamma(m,b)$), however, the new dimension plays a role.

Proposition 3. *Consider two experiments:*

1. *An experiment where the signal is informative only about the strategy ($q_g = q_b$).*
2. *An experiment where the signal conflates information about strategy and idea quality , i.e., $q_g \neq q_b$.*

Even if both experiments increase the belief p_v by the same amount upon a positive outcome, the first experiment is more effective in convincing the investor to invest in the contrarian venture.

Proof. While both experiments in case of success raise $\mathfrak{B}(\mu;p_i)$ to the same level, the first keeps p_i fixed while the second raises it to $p_{i'}$. As such, given that $\mathfrak{B}$ is weakly increasing in both arguments, it must be that μ increases more in the first scenario than in the second. □

Intuitively, a positive signal H is partially "wasted" on updating the investor's opinion about the idea's quality—information that cannot tip the scale in favor of the contrarian venture over

[45] Another way to interpret this set-up is to consider the second variable as the "trustworthiness of the test" rather than the quality of the idea. Under this interpretation, g = "generous" experiment (i.e., less trustworthy) while b = "burdensome" experiment (i.e., more trustworthy). For example, if $\gamma(c,g) = \gamma(c,b) = \gamma(m,g) = 1$ and $\gamma(m,b) = 0$, the experiment perfectly reveals the strategy effectiveness in the burdensome experiment and is useless in the generous experiment. As such, if the contrarian idea is extremely unbelievable, all a successful experiment does is shift the investor's prior p_i that the experiment is too generous.

the mainstream one since the idea is the same for both strategies. Specifically, the investor's updated belief about the contrarian hypothesis is:

$$\mu_{t+1} = \frac{\mu_t P(H|s=c)}{\mu_t P(H|s=c) + (1-\mu_t)P(H|s=m)},$$

where:

$$P(H|s) = \sum_{i\in\{g,b\}} P(H|s,i)P(i) = \gamma(s,g)p_i + \gamma(s,b)(1-p_i).$$

As such, after a string of positive signals, the odds ratio $L_t = \frac{\mu_t}{1-\mu_t}$ of the contrarian hypothesis grows by a factor of $\frac{P(H|s=c)}{P(H|s=m)} = \frac{\gamma(c,g)p_i + \gamma(c,b)(1-p_i)}{(1-\gamma(c,g))p_i + (1-\gamma(c,b))(1-p_i)}$. If the idea quality were known and fixed after the signal, the growth factor would ex-ante be $p_i\frac{\gamma(c,g)}{1-\gamma(c,g)} + (1-p_i)\frac{\gamma(c,b)}{1-\gamma(c,b)}$ (i.e., a weighted average of the growth factor for each i), which is larger by Jensen's inequality and convexity of $f(x) = x/(1-x)$. This result is summarized in the following proposition.

Proposition 4. *When the signal conflates information about strategy effectiveness and idea quality, the growth rate of the investor's odds ratio in favor of the contrarian hypothesis is slower than when the signal is informative only about the strategy. Consequently, the conflated signal is less effective in convincing the investor to invest in the contrarian venture.*

This shows that the extent to which the signal is dissipated depends on the relative precision with which the experiment identifies idea quality and strategy effectiveness—that is, the distance between q_g and q_b. The more the signal is conflated with idea quality, the less effective it is in updating the investor's belief about the contrarian hypothesis.

As such, the information garbling can entrench the mainstream's advantage, reducing the returns from public experimentation for contrarians and reinforcing the informational cascade. Paradoxically, the returns from public experimentation can be higher for mainstream entrepreneurs, who do not need to further increase the gap relative to the contrarian hypothesis, and instead care about the overall investor's belief p_v.

This issue can be solved by a contest between the two approaches, defined as a single experiment that contemporaneously yields results on both the contrarian and the mainstream strategy, effectively creating a horse race. As such, the experiment is defined as a probability distribution that, for each combination of strategy C, strategy M, and idea quality, outputs a bi-dimensional signal $\sigma = (\sigma_c, \sigma_m) \in \{(W_c, L_m), (L_c, W_m)\}$ – indicating which of the two strategies wins in the contest, and which loses (I rule out ties by assumption). A contrarian victory updates the investor's prior that the state of the world is c, and vice versa. While the idea quality remains unobserved, it is the same for both ventures being tested and is thus effectively conditioned out.

I assume that (i) the two strategies are mutually exclusive and exhaustive, so that (m,c) and (c,m) have probability 0 (ii) $P(\sigma = (W_c, L_m)|s=c) = q_1 > 1/2$ and $P(\sigma = (L_c, W_m)|s = m) = q_2 > 1/2$. [46] Then, the joint distribution of the signals and the values of the strategies

[46] If the contest is unbiased, $P(\sigma = (W_c, L_m)|s=c) = P(\sigma = (L_c, W_m)|s=m) = q$. This would make the

(conditional on idea value) are:

Table 6: Contest between M and C as Experiment

		Strategy M	
		m	**c**
Strategy C	**c**	0	$\begin{cases}(W_c,\ L_M) \text{ with prob. } q_1\\(L_c,\ W_m) \text{ with prob. } 1-q_1\end{cases}$
	m	$\begin{cases}(L_C,\ W_M) \text{ with prob. } q_2\\(W_C,\ L_M) \text{ with prob. } 1-q_2\end{cases}$	0

A contrarian victory (W_c, L_M) raises the contrarian posterior from μ to $\frac{\mu q_1}{\mu q_1+(1-\mu)(1-q_2)}$ as in the original setup. As such, a contest completely resolves the issue of omitted variable bias and fully restores the upside of public experimentation for contrarians.

contest a symmetric binary signal as in the initial setup.

Appendix B: Descriptive Statistics

Table B1: Descriptive Statistics for Time-Bound Contests

Variable	N	Mean	Median	Std. dev.	Min	Max	Description
Number of contest events	354						
Year of contest event	354	2015.0	2016	5.11	1997	2022	Year in which the event was organized.
Number of tracks per event	354	1.62	1	1.42	1	12	Number of distinct contest tracks per event.
Number of contest tracks	575						
Teams per contest track	518	133.9	14	501.9	1	5,169	Teams taking part in the contest.
Winning or runner-up results*	1,145						Number of winning or runner-up results.
Winning result	1,145	0.5	1	0.5	0	1	Result is winning.
Year of result	1,145	2015.0	2016	4.92	1997	2022	Year in which winning or runner-up result was obtained.
Margin of victory°	450	0.10	0.04	0.20	0.00	2.42	% Gap between 1^{st} and 2^{nd} (winsorized 5-95 percentiles).
Result published in paper	1,145	0.59	1	0.49	0	1	Result published in preprint or academic paper.
Paper citations of result	670	2,450.5	53	16,496.5	0	166,054	Number of paper citations of result's paper.
Patent citations of result	670	6.99	0	50.66	0	820	Number of patent citations of result's paper.
Contrarian×	656	0.13	0	0.33	0	1	Result uses method from an algorithmic family different than the one in baseline and prior winners.
Novel algorithm†	661	0.24	0	0.429	0	1	Result introduces a novel component or algorithm to an existing algorithmic family.
Seminal algorithm†	661	0.033	0	0.180	0	1	Result introduces a seminal algorithm to an existing algorithmic family.

*Five contest tracks only reported information on the winners. ° Calculated only for winners. × Only available when there is at least one prior winner or baseline result. † Only available when the submission is associated with a paper pair.

Table B2: Descriptive Statistics for Ongoing Benchmarks

Variable	N	Mean	Median	Std. dev.	Min	Max	Description
Number of datasets	237						
Dataset has public data	220	0.83	1	0.38	0	1	Dataset can be downloaded at paperswithcode.com.
Number of benchmarks per dataset	237	4.8	2	6.73	1	41	Number of benchmarks per dataset.
Number of Benchmarks	408						Number of academic open benchmarks.
Benchmark's First Year	408	2015.2	2015.5	2.06	2007	2020	Year of creation of open benchmark.
Submissions per benchmark	408	33.7	27	23	13	267	Number of submissions per benchmark (at least 2 years after inception).
Number of Submissions	13,756						Number of distinct results submitted to the benchmark (at least 2 years after inception).
Year of Submission	13,756	2019.2	2019	1.83	2011	2022	Year the result was submitted.
State-of-the-art submission	13,756	0.21	0	0.41	0	1	The result is the best at submission time.
Days as state-of-the-art°	2,913	279.3	181	308.7	0	2,336	Conditional on being best submission, how long until other result beats it.
Margin of victory×	2,865	0.11	0.03	0.18	0.00	0.68	% Improvement relative to prior state-of-the-art result (winsorized 5-95 percentiles).
Result published in paper	13,756	0.99	1	0.07	0	1	Result published in preprint or academic paper.
Paper citations of result	13,684	993.0	160	5,444	0	166,054	Number of citations of result's paper.
Patent citations of result	13,684	3.01	0	14.1	0	390	Number of patent citations of result's paper.
Contrarian	10,919	0.17	0	0.37	0	1	Result uses method from algorithmic family different than prior winner or baseline.
Novel algorithm	13,665	0.61	1	0.49	0	1	Result introduces novel component or algorithm to an existing algorithmic family.
Seminal algorithm	13,665	0.04	1	0.19	0	1	Result introduces seminal algorithm to an existing algorithmic family.

° Calculated only for state-of-the art results.
× Only available when there is at least one prior state-of-the-art result.

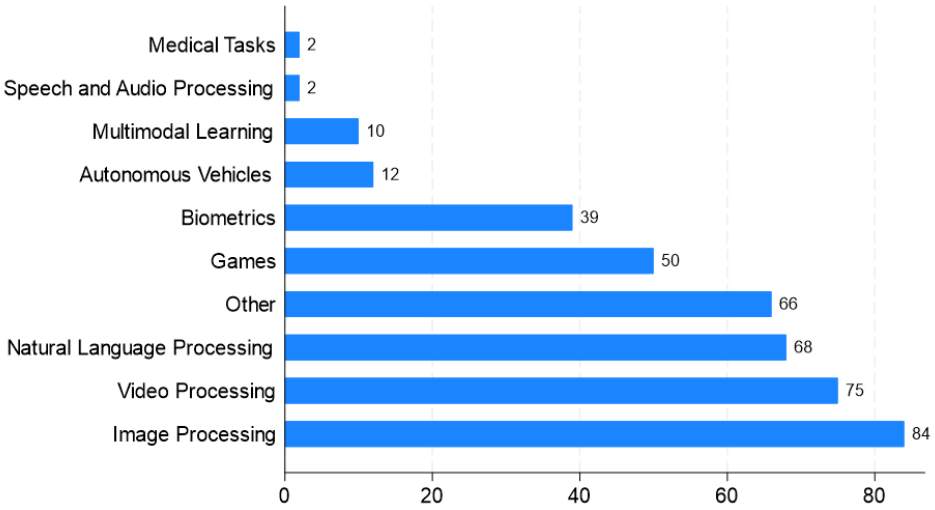
Table B3: Descriptive Statistics for Researchers

Variable	**N**	**Mean**	**Median**	**Std. dev.**	**Min**	**Max**	**Description**
Number of Researchers	12,662						Participants to a contest or open benchmark that have a research profile on the DBLP computer science repository.
Experience at first contest	12,089	8.23	6	8.12	0	50	Years of research activity prior to the first participation in a contest or open benchmark.
Articles before first contest	12,089	39.4	10	82.3	0	1,312	Number of academic works produced prior to the first participation.
% past articles with funding acknowledg.	9,579	0.32	0.26	0.28	0	1	Proportion of academic works prior to the first participation that mention funding in the acknowledgements.
% past articles that are preprints	10,039	0.20	0.09	0.24	0	1	Proportion of academic works prior to the first participation that are unpublished.
Ever received best paper award	12,662	0.02	0	0.15	0	1	The researcher was awarded a best paper award in a 10-year window around the first participation.
Ever entrepreneur	12,662	0.04	0	0.19	0	1	The researcher founded a startup in a 10-year window around the first participation.

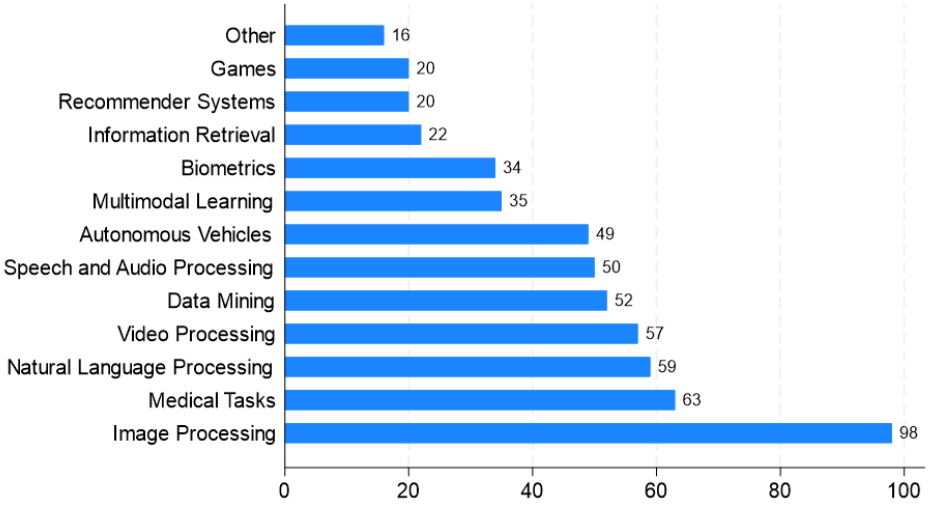
Table B4: Descriptive Statistics for Startups Founded Close to Contest [t-5, t+5]

Variable	**N**	**Mean**	**Median**	**Std. dev.**	**Min**	**Max**	**Description**
Number of Startups	477						Number of startups founded by a participant in a [t-5, t+5] interval across a contest or open benchmark.
Founding year	477	2016.9	2017	4.6	1998	2024	Startup's founding year.
Years since contest	477	0.40	1	2.74	-5	5	Years since submission to contest.
Number of team members co-founders	477	1.12	1	2.37	1	5	Number of team members in the contest that co-founded the startup.
Team has won contest	477	0.62	1	0.48	0	1	Cofounders have won the contest.
Team has won as contrarian	477	0.13	0	0.34	0	1	Cofounders have won the contest as contrarians.
Venture capital raised ($m)	444	148.3	0.7	886.5	0	11,312	Total venture capital raised ($m).
Last valuation ($m)	166	1,456.8	50.5	7,336.5	0.29	86,000	Last known valuation ($m).
IPO or unicorn status	477	0.08	0	0.28	0	1	Startup has reached unicorn status or has IPO'd.
Was acquired	477	0.17	0	0.37	0	1	Startup was acquired.
Years between founding and acquisition	77	3.90	4	2.89	0	14	Years between founding and acquisition, for startups that are acquired.
Number of competitors at entry	231	5,373.9	862	9,689.7	0	38,230	Number of competitors active in the founding year that have the same Pitchbook keywords.
Number of government grants	352	0.32	0	1.18	0	16	Number of governmental grants that support the startup.
Number of corporate-backed rounds	352	1.74	1	3.39	0	35	Number of corporate-backed funding rounds.
Has a granted patent	462	0.27	0	0.45	0	1	Startup has at least 1 granted patent.

Appendix C: Supporting Evidence



(a) Open Benchmarks



(b) Time-Bound Contests

Figure C1: Distribution of Tasks Across Open Benchmarks and Time-Bound Contests

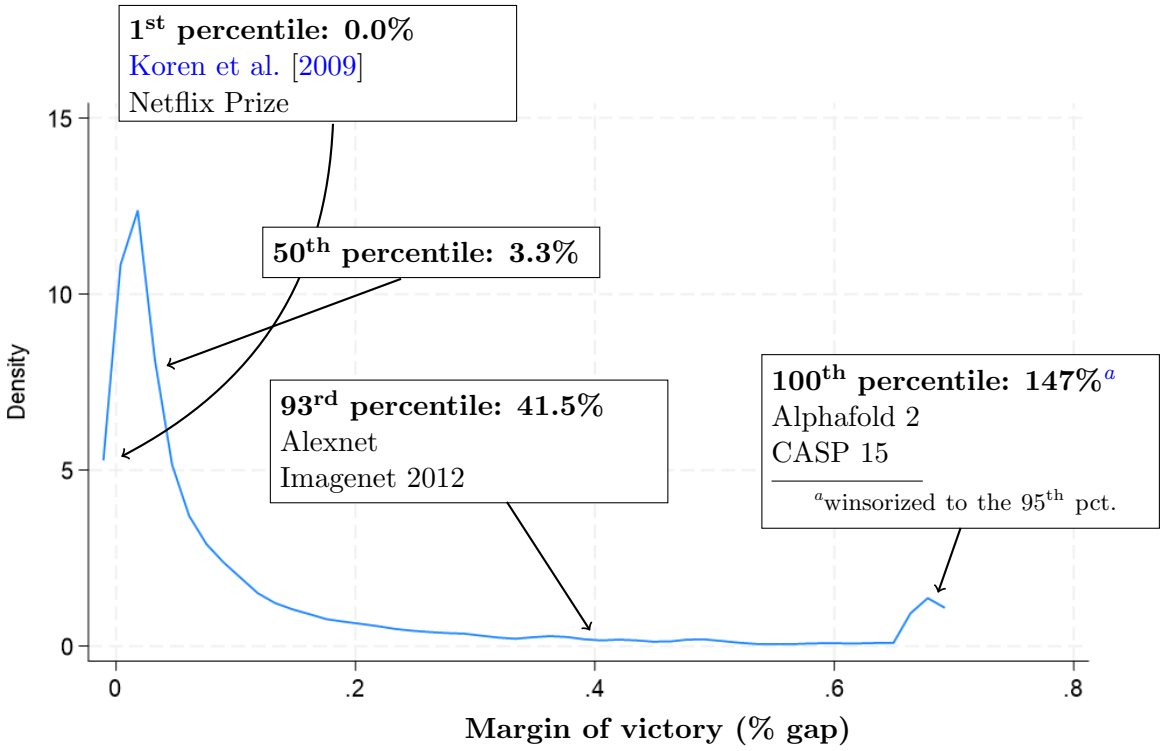


Figure C2: Percentage improvement between best-performers and runner-ups

Note: The figure shows the distribution of the gap between 1st and 2nd place in time-bound contests and between new and previous state-of-the-art contributions in ongoing benchmarks. The analysis is based on 3,315 winning submissions and has been winsorized at the 5% and 95% levels.

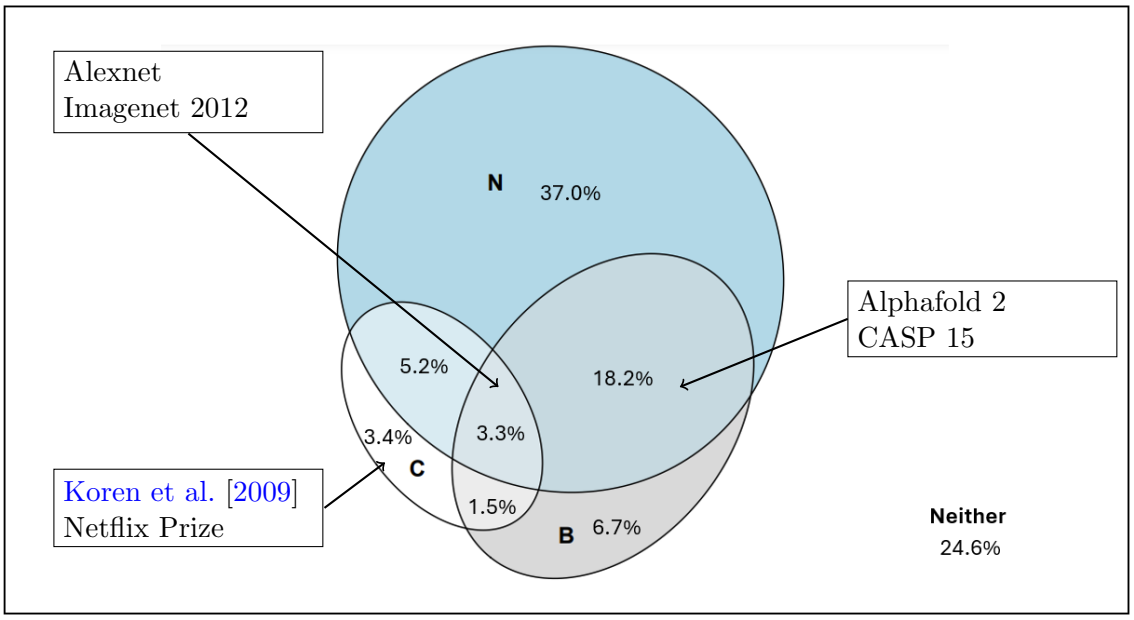
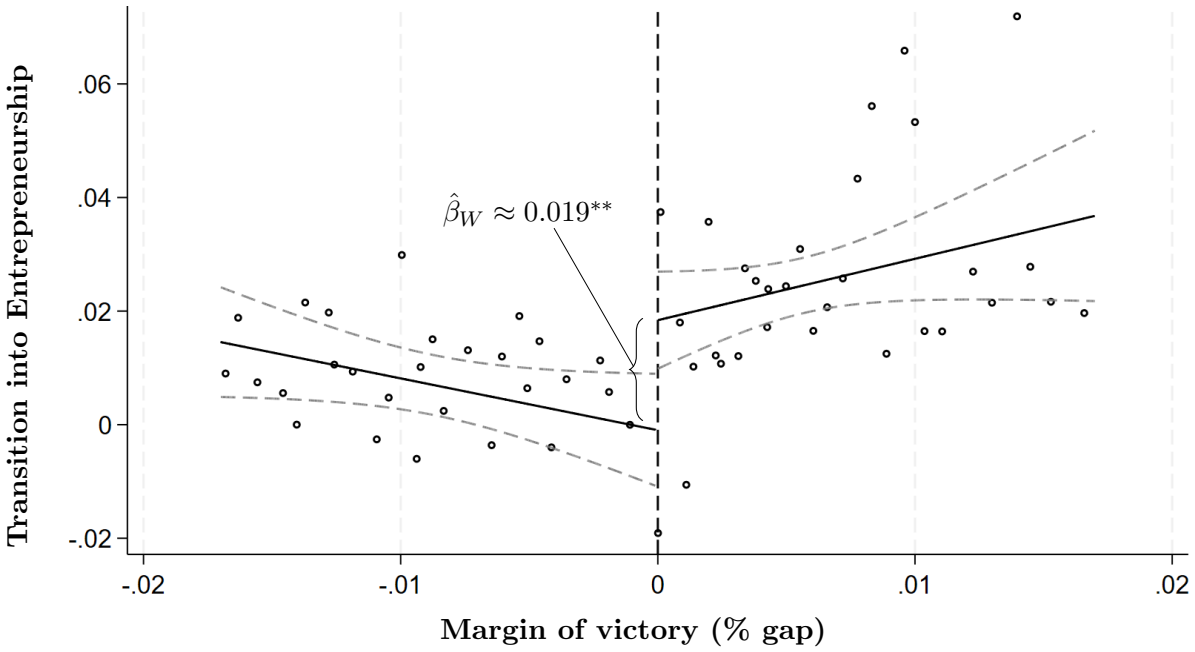


Figure C3: Overlap between Contrarian (**C**), Novel (**N**), and Breakthrough (**B**) winners.

Note: Calculated for 1,538 winners with non-missing information for all variables.



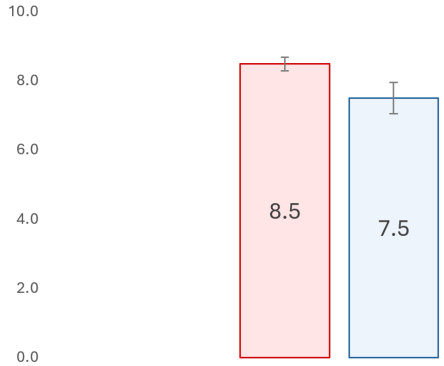
VARIABLES	(1) Founds a startup	(2) Founds a startup	(3) Founds a startup	(4) Founds a startup
Winner	0.011*	0.019**	0.019**	0.058*
	(0.006)	(0.007)	(0.007)	(0.028)
Novel			0.001	-0.011
			(0.004)	(0.043)
Constant	0.011	-0.001	-0.002	0.031
	(0.005)	(0.005)	(0.006)	(0.014)
Observations	6,428	6,410	6,378	1,526
Gap	YES	YES	YES	YES
Gap × Win	YES	YES	YES	-
Challenge FE	NO	YES	YES	YES
Submission Year FE	NO	YES	YES	YES
Exclude entrepreneurs at t-1	YES	YES	YES	YES
Only include contests	NO	NO	NO	YES
Units of observation	Researcher × Cont./Bench.	Researcher × Cont./Bench.	Researcher × Cont./Bench.	Researcher × Contest

Figure C4: Regression Discontinuity Analysis: Impact of Winning on Transition into Entrepreneurship

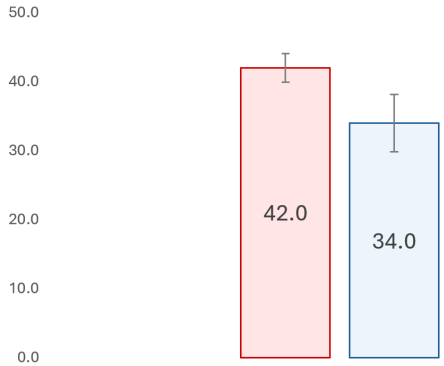
Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Standard errors clustered by researcher and challenge. This table presents the results from a sharp regression discontinuity design analyzing the effect of winning an evaluation challenge (including both contests and benchmarks) on the likelihood of founding a startup by year 4 after the evaluation for researchers who had not founded a company prior to it. The model is specified as the following local linear regression equation:

$$Y = \beta_W \text{Win} + \alpha_1 \text{Gap} \times \text{Win} + \alpha_2 \text{Gap} + \text{Controls} \qquad \text{for Gap} \in [-h, h]$$

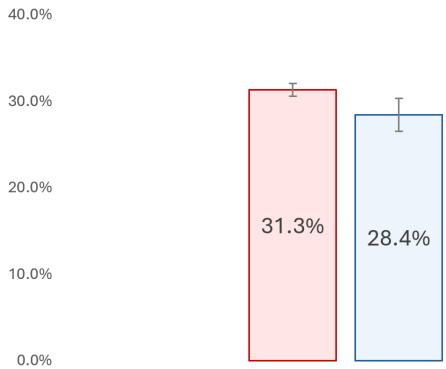
where the running variable Gap represents the (scaled) distance between the focal submission's result and the minimum result required to win (i.e., the second-place result for winners and the first-place result for runners-up). The bandwidth h is selected using the procedure described in Calonico et al. [2014]. β_W captures the winning premium, i.e., the causal effect of winning on entrepreneurship. The figure corresponds to the estimates reported in column (3). Column (4) corresponds to the RDD estimates when focusing on time-bound contests only: given that I analyze only winners and runner-ups in each contest, Gap × Win is omitted.



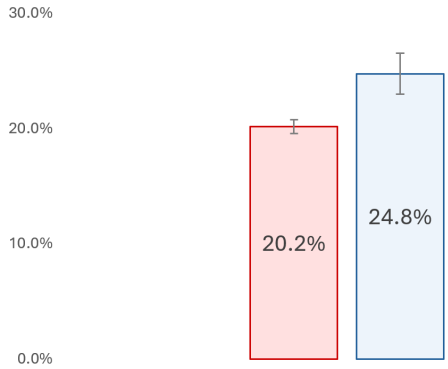
(a) Years between first publication and evaluation challenge (N=8,007 researchers)



(b) Number of Academic Articles before evaluation challenge (N=8,007 researchers)



(c) % Articles with Funding (N=6,472 researchers)

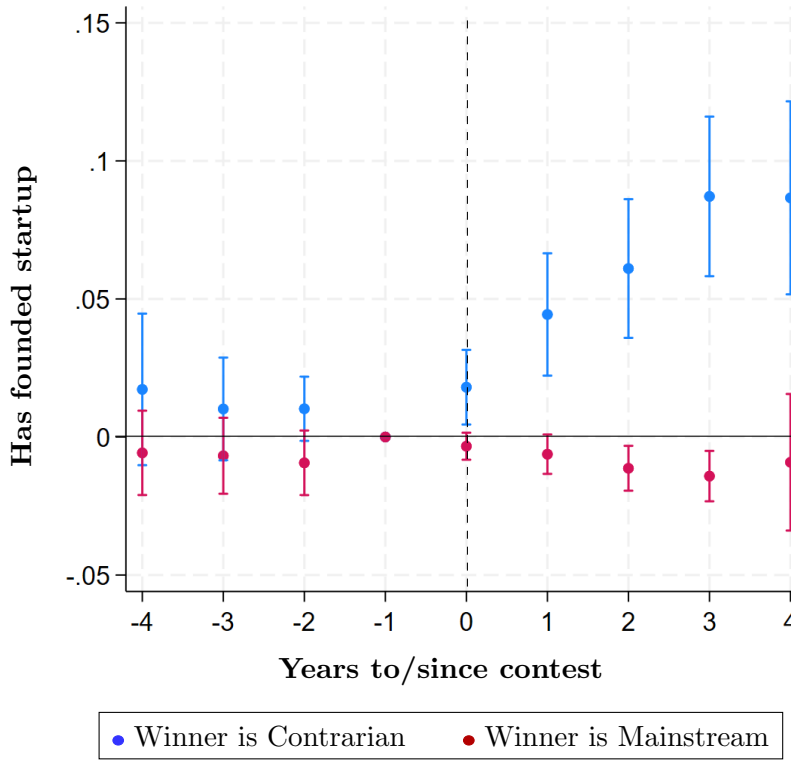


(d) % Articles that are preprints (N=6,766 researchers)

Figure C5: ■ Mainstream vs. ■ Contrarian Researchers prior to the Evaluation Challenge

Note: This figure summarizes the characteristics of researchers prior to participating in an evaluation campaign for the first time, contrasting those who made a contrarian contribution vs. those who made a mainstream contribution. 95% confidence intervals based on robust standard errors.

Figure C6: Spillover effect of a contrarian victory on contrarian vs. mainstream runner-ups

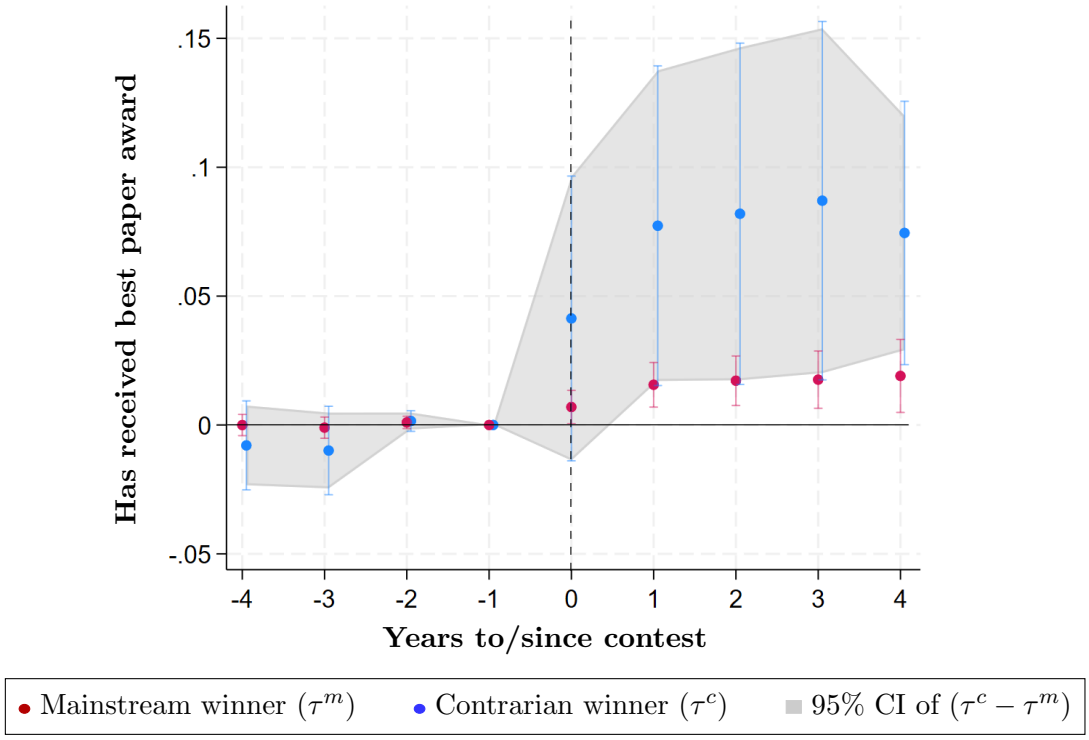


Note: The figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following specification, focusing solely on *runner-up researchers*.

$$
\begin{aligned}
Y_{it} = &\sum_{j\neq -1} \tau_j^m D_{i,\,t-j} \times \text{Contrarian}_i \times \text{Defeated by Mainstream}_i \\
&+ \sum_{j\neq -1} \tau_j^c D_{i,\,t-j} \times \text{Contrarian}_i \times \text{Defeated by Contrarian}_i \\
&+ \sum_{j\neq -1} \tau_j^{r_1} D_{i,\,t-j} \times \text{Defeated by Mainst.}_i + \sum_{j\neq -1} \tau_j^{r_2} D_{i,\,t-j} \times \text{Defeated by Contr.}_i + \beta_i + \gamma_c + \alpha_t
\end{aligned}
$$

where Y_{it} is 1 if the runner-up researcher has founded a startup on or before the focal year, 0 otherwise. τ_j^m estimates the difference in entrepreneurial propensity between contrarian and mainstream runner-ups when the winner of the challenge is mainstream (for ongoing benchmarks, at least one state-of-the-art result submitted in the same focal year was mainstream). τ_j^c, instead, provides the difference between contrarian and mainstream runner-ups when the winner of the challenge is contrarian (for ongoing benchmarks, the state-of-the-art results submitted in the same focal year are all contrarian). The results show that contrarian runner-ups become more likely to found a startup after the challenge submission than mainstream runner-ups, but only if the winner was contrarian. Controls include researcher, challenge, year fixed effects. Standard errors are clustered by researcher and challenge.

Figure C7: Effect of winning on chances of receiving a best paper award

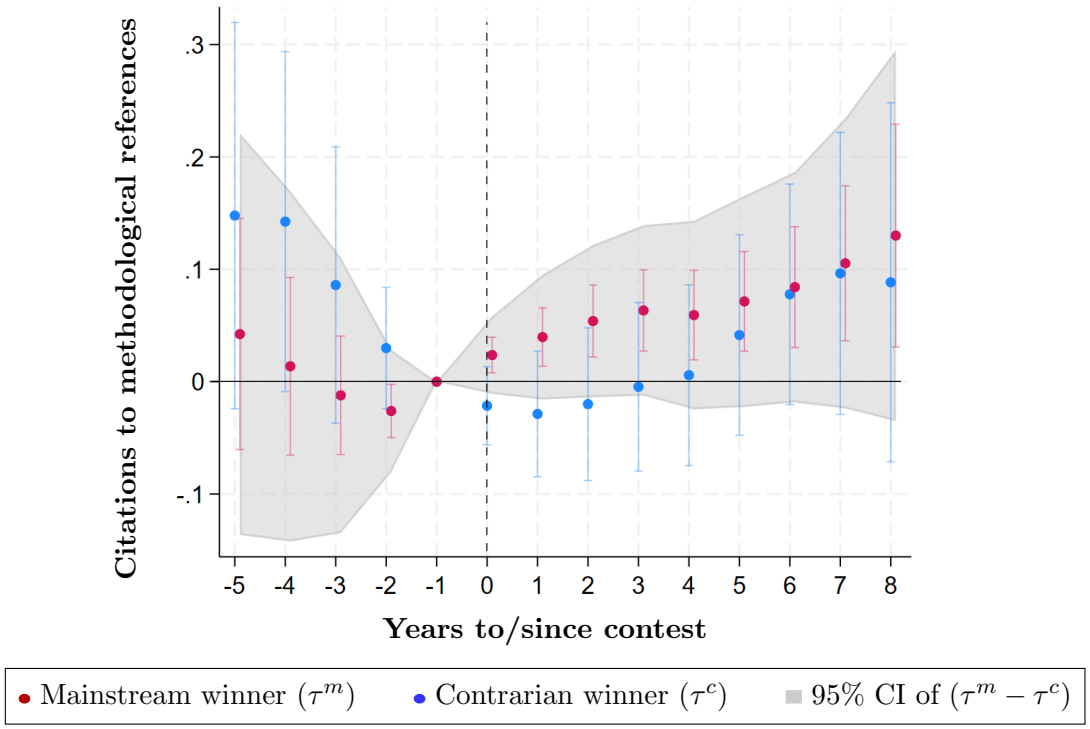


Note: The figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following specification:

$$Y_{it} = \sum_{j \neq -1} \tau_j^m \; D_{i,\, t-j} \times \text{Mainstream winner}_i + \sum_{j \neq -1} \tau_j^c \; D_{i,\, t-j} \times \text{Contrarian winner}_i + \\ + \sum_{j \neq -1} \tau_j^r \; D_{i,\, t-j} \; + \beta_i + \gamma_c + \alpha_t$$

The dependent variable is 1 if the researcher has received a best paper award on or before the focal year, 0 otherwise. τ_j^c (τ_j^m) represent the difference in the propensity to receive a best paper award for contrarian (mainstream) winners relative to runner-ups, before and after the year of the winning submission. The gray area plots the 95% confidence interval for the vertical difference between blue and red dots $\tau_j^c - \tau_j^m$: the effect of winning for contrarians vs. mainstream is statistically different from 0 whenever the red dot is outside the gray area. Controls include researcher, challenge, and year fixed effects. Standard errors are clustered by researcher and challenge.

Figure C8: Change in Citations to the Methods used by Contrarian Winners vs. Mainstream Winners

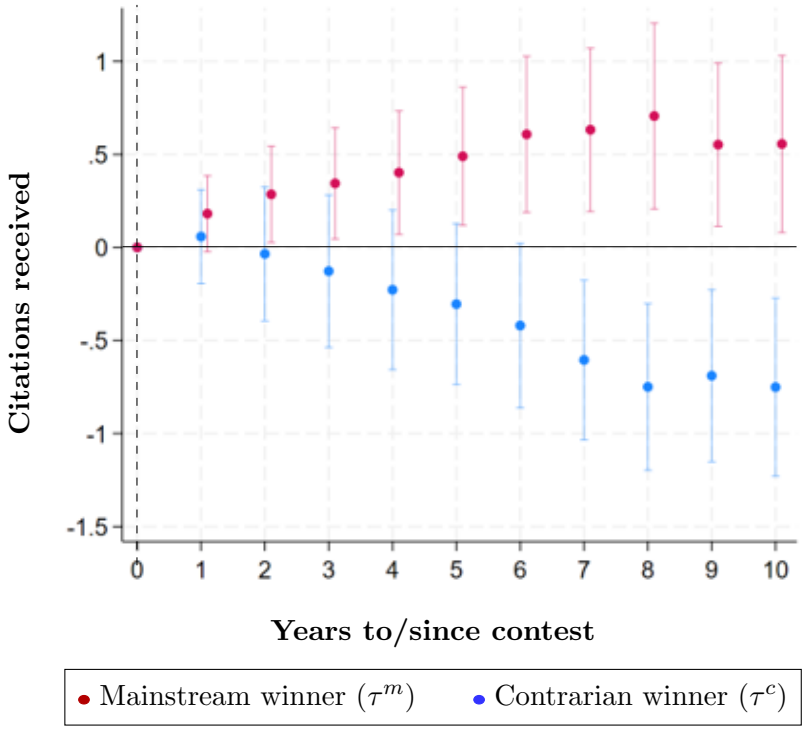


Note: the figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following conditional fixed effects quasi–maximum likelihood Poisson specification:

$$Y_{rt} = \sum_{j \neq -1} \tau_j^m \; D_{r,\, t-j} \;\times\; \text{Mainstream winner}_e + \sum_{j \neq -1} \tau_j^c \; D_{r,\, t-j} \;\times\; \text{Contrarian winner}_e + \\ + \sum_{j \neq -1} \tau_j^r \; D_{r,\, t-j} + \beta_r + \gamma_e + \alpha_t$$

where Y_{rt} is the number of cumulative citations of r, a methodological reference cited by challenge entry e. τ_j^c (τ_j^m) represent the difference in cumulative citations accruing to methods used by contrarian (mainstream) winners relative to runner-ups, before and after the year of the winning submission. The strong negative pre-trends indicate that citations to contrarian methods were declining relative to those used by runner-ups and mainstream winners before the evaluation challenge. However, this downward trajectory begins to reverse after winning the challenge in $j = 0$. After four years ($j = 4$), citations to methods used by contrarian winners surpass those of runner-ups compared to the levels at $j = -1$.
The gray area plots the 95% confidence interval for the vertical difference between red and blue dots $\tau_j^m - \tau_j^c$: the effect of winning for mainstream vs. contrarians is statistically different from 0 whenever the blue dot is outside the gray area. I include fixed effects for the methodological reference r, entry e and year t. Standard errors are clustered by methodological reference.

Figure C9: Impact of winning a challenge on citations: contrarian vs. mainstream approaches in close races



Note: The figure shows the event study coefficients τ_j^m, τ_j^c estimated for the following Poisson quasi–maximum likelihood specification:

$$Y_{it} = \sum_{j>0} \tau_j^m \; D_{i,\,t-j} \;\times\; \text{Mainstream winner}_i + \sum_{j>0} \tau_j^c \; D_{i,\,t-j} \;\times\; \text{Contrarian winner}_i + \\ + \sum_{j>0} \tau_j^r \; D_{i,\,t-j} + \beta_i + \gamma_c + \alpha_t$$

where Y_{it} represents the cumulative number of citations received by the paper describing challenge submission i as of year t. The chart plots τ_j^c (τ_j^m), the difference in cumulative citations accruing to contrarian (mainstream) winning submissions relative to runner-up submissions. The figure only includes winners that won a close race ($< 1\%$ gap with runner-up). It shows that contrarian winners are cited less than both mainstream winners and runner-ups in the first 10 years after the evaluation challenge (including contests and ongoing benchmarks).
To mitigate truncation concerns, I only include submissions whose paper-pair has at least 5 years of citation data. I also include challenge, paper, and year fixed effects. Standard errors are clustered by paper and challenge. No pre-trends are available because the papers detailing the winning and runner-up entries did not exist prior to the challenge submission.

Table C1: Classification of algorithm families

Algorithm family	**Examples of methods and models**
Neural networks	CNN, RNN, Transformers
Regression	Poisson regression, Multivariate adaptive regression splines
Clustering methods	k-means clustering, collaborative filtering
Bayesian methods	Latent Dirichlet allocation, Bayesian net
Symbolic AI	Association Rule Learning, expert systems
k Nearest Neighbors	k Nearest Neighbors, Hart algorithm
Time series methods	GARCH, VAR
Feature engineering methods	bag of words, SIFT, spatial pyramid
Probabilistic graphical models	Hidden Markov model, Conditional Random Field
Dimensionality reduction	Latent semantic analysis, principal component analysis
Automata	probabilistic automata, learning automata
Anomaly detection methods	Local outlier factor, Random simple consensus
Discriminant analysis	Linear Discriminant Analysis, Quadratic Discriminant Analysis
Fuzzy systems	ANFIS, fuzzy logic systems
Metaheuristic Optimization	Genetic Algorithms, Particle Swarm Optimization
Filters	Gabor filter, correlation filters
Reinforcement Learning	Monte-Carlo methods, Q-learning
Kernel methods	Support vector machines, Fisher kernel
Trees	Decision trees, CART
Ensemble methods	gradient boosting machines, bagging

Note: This taxonomy categorizes twenty algorithmic families (first column), with representative models listed in the second column. The complete taxonomy, including all models, is available upon request. For further details on data sources and methodology, please refer to the main text.

Table C2: Relationship between Contrarianism, Novelty, and Performance

VARIABLES	(1) Novel	(2) Novel	(3) Novel & seminal	(4) Novel & seminal	(5) -Rank	(6) -Rank	(7) Winner	(8) Winner
Contrarian	0.017	0.002	-0.024	-0.041+	-10.948*	-3.839*	-0.128***	-0.027
	(0.028)	(0.027)	(0.016)	(0.022)	(4.387)	(1.786)	(0.016)	(0.018)
Novel					2.789**	4.707**	0.069***	0.090***
					(0.870)	(1.380)	(0.015)	(0.014)
Constant	0.619***	0.624***	0.049***	0.052***	-10.343***	-12.797***	0.246***	0.213***
	(0.014)	(0.013)	(0.008)	(0.009)	(1.213)	(0.841)	(0.012)	(0.010)
Observations	11,355	11,256	11,355	11,256	11,355	11,256	11,355	11,256
Challenge FE	NO	YES	NO	YES	NO	YES	NO	YES
Year of submission FE	YES	YES	YES	YES	YES	YES	YES	YES
Units of observation	Paper x Contest	Paper x Contest	Paper x Contest	Paper x Contest	Paper x Contest	Paper x Contest	Paper x Contest	Paper x Contest

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Standard errors clustered by paper and evaluation challenge (including contests and ongoing benchmarks). To construct "-Rank," I take the negative value of the paper's ranking at the time of submission. A higher value represents a better rank (e.g., for winner or state-of-the-art it is -1, for 2nd place it is -2, etc.).

Table C3: Contrarian Winners and Breakthrough Impact

VARIABLES	(1) Performance relative to 2nd best	(2) Days as state of the art	(3) Age of broken record
Contrarian	0.042*	94.446**	203.405***
	(0.016)	(32.539)	(45.982)
Novel	0.023*	30.414+	-4.584
	(0.011)	(16.843)	(18.738)
Constant	0.085***	234.875***	250.540***
	(0.007)	(12.870)	(16.065)
Observations	3,021	2,787	2,795
Only winners	YES	YES	YES
Only ongoing benchmarks	NO	YES	YES
Year of submission FE	YES	YES	YES
Units of observation	Paper x Contest	Paper x Contest	Paper x Contest

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Standard errors clustered by paper and challenge. This table presents regression results on the relationship between contrarianism and novelty and the nature of the breakthrough achieved by contest winners and new state-of-the-art results. The analysis does not include contest FE to avoid excluding contests with only one winner. 'Performance relative to 2nd best' is the % gap between 1st and 2nd place (for contest winners) or new vs. old state-of-the-art (for new best results in open benchmarks). 'Days as state of the art' is the number of days the new best result will remain in place, while 'Age of broken record' is the number of days the old state-of-the-art had been resisting: both can only be constructed for ongoing benchmarks.

Table C4: Winning and contrarian entrepreneurship in ongoing benchmarks

	(1)	(2)
VARIABLES	Has founded a startup	Has founded a startup
Mainstream Win × Post	0.026*	
	(0.012)	
Contrarian Win × Post	-0.014+	-0.004
	(0.008)	(0.014)
Mainstream Win × Post × Public data	-0.013	
	(0.013)	
Contrarian Win × Post × Public data	0.019*	
	(0.008)	
Contrarian Win × Post × (≤ 30 days as SOTA)		-0.024+
		(0.015)
Constant	0.027***	0.046***
	(0.003)	(0.000)
Observations	35,536	15,852
Only ongoing benchmarks	YES	YES
Only new state-of-the-art	NO	YES
Researcher FE	YES	YES
Benchmark FE	YES	NO
Year of submission FE	YES	YES
Units of observation	Researcher x Benchm. x Year	Researcher x Benchm. x Year

Note: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1. Dependent variable is 1 if the researcher has founded a startup on or before the focal year, 0 otherwise. Researchers are observed in two periods: one before the submission, and four years after the submission. Standard errors clustered by researcher and benchmark. This table examines the impact of achieving a new state-of-the-art (SOTA) on an ongoing benchmark on the likelihood of becoming an entrepreneur, separately for contrarians and mainstream participants. Column (1) explores heterogeneous effects based on whether the underlying data are openly available on paperswithcode.com. The omitted baseline are non-SOTA results. Column (2) analyzes the heterogeneous effects of how long the new state-of-the-art remains in place. Since only state-of-the-art results are included, the omitted baseline here are mainstream state-of-the-art results: e.g., "Contrarian winner × Post" captures difference between contrarian winners and mainstream winners.