Article

Quantifying the dynamics of failure across science, startups and security

https://doi.org/10.1038/s41586-019-1725-y

Received: 15 February 2019

Accepted: 27 September 2019

Published online: 30 October 2019

There are amendments to this paper

Yian Yin[1,2,3], Yang Wang[1,2,4], James A. Evans[5,6] & Dashun Wang[1,2,3,4]*

Human achievements are often preceded by repeated attempts that fail, but little is known about the mechanisms that govern the dynamics of failure. Here, building on previous research relating to innovation[1–7], human dynamics[8–11] and learning[12–17], we develop a simple one-parameter model that mimics how successful future attempts build on past efforts. Solving this model analytically suggests that a phase transition separates the dynamics of failure into regions of progression or stagnation and predicts that, near the critical threshold, agents who share similar characteristics and learning strategies may experience fundamentally different outcomes following failures. Above the critical point, agents exploit incremental refinements to systematically advance towards success, whereas below it, they explore disjoint opportunities without a pattern of improvement. The model makes several empirically testable predictions, demonstrating that those who eventually succeed and those who do not may initially appear similar, but can be characterized by fundamentally distinct failure dynamics in terms of the efficiency and quality associated with each subsequent attempt. We collected large-scale data from three disparate domains and traced repeated attempts by investigators to obtain National Institutes of Health (NIH) grants to fund their research, innovators to successfully exit their startup ventures, and terrorist organizations to claim casualties in violent attacks. We find broadly consistent empirical support across all three domains, which systematically verifies each prediction of our model. Together, our findings unveil detectable yet previously unknown early signals that enable us to identify failure dynamics that will lead to ultimate success or failure. Given the ubiquitous nature of failure and the paucity of quantitative approaches to understand it, these results represent an initial step towards the deeper understanding of the complex dynamics underlying failure.

To understand the dynamics of failure, we collected three large-scale datasets (Supplementary Information 1). The first dataset (D_1) contains all R01 grant applications submitted to the NIH (776,721 applications by 139,091 investigators, 1985–2015; Supplementary Information 1.1). For each grant application, we obtained ground-truth information on whether or not it was funded, allowing us to reconstruct individual application histories and their repeated attempts to obtain funding. Our second dataset (D_2) traces start-up investment records from VentureXpert[18] (58,111 startup companies involving 253,579 innovators, 1970–2016; Supplementary Information 1.2). Tracing every startup in which venture capital firms invested, D_2 allows us to reconstruct individual career histories counting successive ventures in which they were involved. Here we follow previous studies in the entrepreneurship literature[19], and classify successful ventures as those that achieved initial public offering (IPO) or high-value mergers and acquisitions, and correspondingly failed attempts as those that failed to obtain such an exit within five years after their first investment by venture capital firms. Going beyond traditional innovation domains, we collected our third dataset (D_3) from the Global Terrorism Database[20] (170,350 terrorist attacks by 3,178 terrorist organizations, 1970–2017; Supplementary Information 1.3). For each organization we trace their attack histories[21,22], and classify success as fatal attacks that killed at least one person, and correspondingly failure as those that failed to claim casualties.

Mechanisms of chance and learning

Chance and learning[13,16] are two primary mechanisms that explain how failures may lead to success. If each attempt has a certain likelihood of success, the probability that multiple attempts all lead to failure decreases exponentially with each trial. The chance model therefore emphasizes the role of luck, suggesting that success eventually arises

[1]Center for Science of Science and Innovation, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [3]McCormick School of Engineering, Northwestern University, Evanston, IL, USA. [4]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [5]Department of Sociology, University of Chicago, Chicago, IL, USA. [6]Santa Fe Institute, Santa Fe, NM, USA. *e-mail: dashun.wang@northwestern.edu

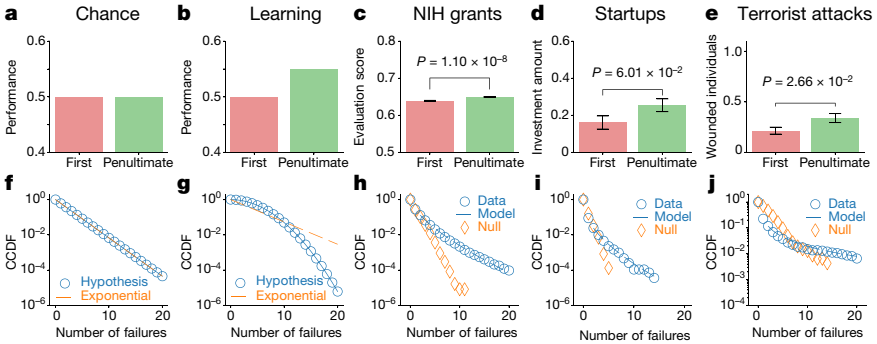


Fig. 1 | Mechanisms of chance and learning. a–j, We compare theoretical predictions and empirical measurements for performance changes (**a**–**e**) as well as the length distribution of failure streaks (**f**–**j**). **a**, **f**, The chance model predicts no performance change (**a**) with a failure streak length that follows an exponential distribution (**f**). **b**, **g**, The learning hypothesis predicts improved performance (**b**) with failure streaks that are shorter than expected by the chance model, corresponding to a faster-than-exponential distribution (**g**). Both hypotheses are contested by empirical patterns observed across the three datasets. To ensure that performance metrics are comparable across data and models, we standardized performance measures according to their underlying distribution (Supplementary Information 5.1). **c**–**e**, We find that failures in real data are associated with improved performance between the first and penultimate attempt. Two-sided Welch's *t*-test; data are mean ± s.e.m. **c**, $n = 4{,}872$ (first), 5,966 (penultimate). **d**, $n = 579$ (first), 548 (penultimate). **e**, $n = 231$ (first), 230 (penultimate). **h**–**j**, At the same time, however, failure streaks are characterized by a fat-tailed length distribution, indicating that failure streaks in real data are longer than expected by chance. For clarity, here we show results for failure streaks for which the length is less than 21 (Supplementary Information 5.2). We further construct a randomized sequence of successes and failures by assigning each attempt to agents at random (Supplementary Information 5.2). We find that failure streak length in the randomized sequence follows an exponential-like distribution, showing clear deviations from the data.

from an accumulation of independent trials. To test this, we compared the performance of the first and penultimate attempt within failure streaks (Supplementary Information 5.1), measured by NIH percentile score for a grant application (D_1), investment size by venture capital firms to a company (D_2) and number of wounded individuals by an attack (D_3). We find that across all three datasets, the penultimate attempt shows systematically better performance than the initial attempt (Fig. 1c–e). These results reject that success is simply driven by chance (Fig. 1a) but lend support to the learning mechanism (Fig. 1b), which suggests that failure may teach valuable lessons that are difficult to learn otherwise[12,13,16]. As such, learning reduces the number of failures required to achieve success, and predicts that failure streaks should follow a narrower length distribution (Fig. 1g) than the exponential distribution predicted by chance (Fig. 1f). However, across all three domains, the length of failure streaks follows a fat-tailed distribution (Fig. 1h–j, Supplementary Information 5.2), indicating that despite improvements in performance, failures are characterized by longer-than-expected streaks before the onset of success. Together, these observations demonstrate that neither chance nor learning alone can explain the empirical patterns that underlie failures, suggesting that more complex dynamics may be at work.

Modelling dynamics of failure

Here we explore the interplay between chance and learning by developing a simple one-parameter model that mimics how future attempts build on previous failures (Fig. 2a, b, Supplementary Information 3.1). We consider that each attempt consists of many independent, unweighted components, with each component i being characterized by an evaluation score x^i (Fig. 2a). For example, components for the submission of an NIH proposal include constructing a biosketch, assembling a budget, writing a data management plan, adding preliminary data and outlining broader impacts. We also note that granting agencies often provide rubrics to grade proposals on specific components.

To formulate a new attempt, one goes through each component, and decides to either create a new version (with probability p) or reuse the best version x^* among the previous k attempts (with probability $1-p$) (Fig. 2b). A new version is assigned a score drawn randomly from a uniform distribution $U[0, 1]$, approximating the percentile of score distributions real systems follow. The decision to create a new version is often not random, but driven by the quality of previous versions. Indeed, given the best version x^*, $1-x^*$ captures the potential to improve it[16]. The higher this potential, the more likely one may create a new version, prompting us to consider a simple relationship, $p = (1-x^*)^\alpha$, with $\alpha > 0$ (Methods, Supplementary Information 3.6). Creating a new version takes one unit of time with no certainty that its score will be higher or lower than the previous one. By contrast, reusing the best version from the past saves time, and allows the component to retain its best score x^*.

Here we explore a single parameter k for our model, measuring the number of previous attempts one considers when formulating a new one (Fig. 2b). Mathematically the dynamical process can be described as: with probability p, $x_n \sim U[0,1]$ or $x_n = x_n^*$ otherwise (with probability $1-p$) where $x_n^* = \max\{x_{n-k}, \cdots, x_{n-1}\}$. We quantify the dynamics of the model by calculating the quality of the nth attempt, $\langle x_n \rangle$, which measures the average score of all components, and the efficiency after that attempt, $\langle t_n \rangle$, which captures the expected proportion of components updated in new versions. Let us first consider the two extreme cases. In the first case, $k = 0$ means that each attempt is independent from previous attempts (Supplementary Information 3.2). Here our model recovers the chance model, predicting that as n increases, both $\langle x_n \rangle$ and $\langle t_n \rangle$ remain constant (Extended Data Fig. 1a, d). That is, without considering past experience, failure does not lead to quality improvement. Nor is it more efficient to try again.

The other extreme ($k \to \infty$) considers all past attempts. The model predicts a temporal scaling in failure dynamics (Supplementary Information 3.3). That is, the time it takes to formulate a new attempt decays with n, asymptotically following a power law (Extended Data Fig. 1e):

$$T_n \equiv \langle t_n \rangle / \langle t_1 \rangle \sim n^{-\gamma} \quad (1)$$

where $\gamma = \gamma_\infty = \alpha/(\alpha+1)$ falls between 0 and 1 and '~' indicates 'asymptotically proportional to'. Besides increased efficiency, new attempts also improve in quality, as the average potential for improvement decays according to $\langle 1 - x_n \rangle \sim n^{-\eta_\infty}$ where $\eta_\infty = \min\{\gamma_\infty, 1-\gamma_\infty\}$ (Extended Data Fig. 1b). Here the model recovers the canonical result from the learning literature[12,15,23–25], commonly known as Wright's law[26]. This is because, as experience accumulates, high-quality versions are preferentially retained, whereas their lower-quality counterparts are more likely to receive updates. As fresh attempts improve in quality (Extended Data

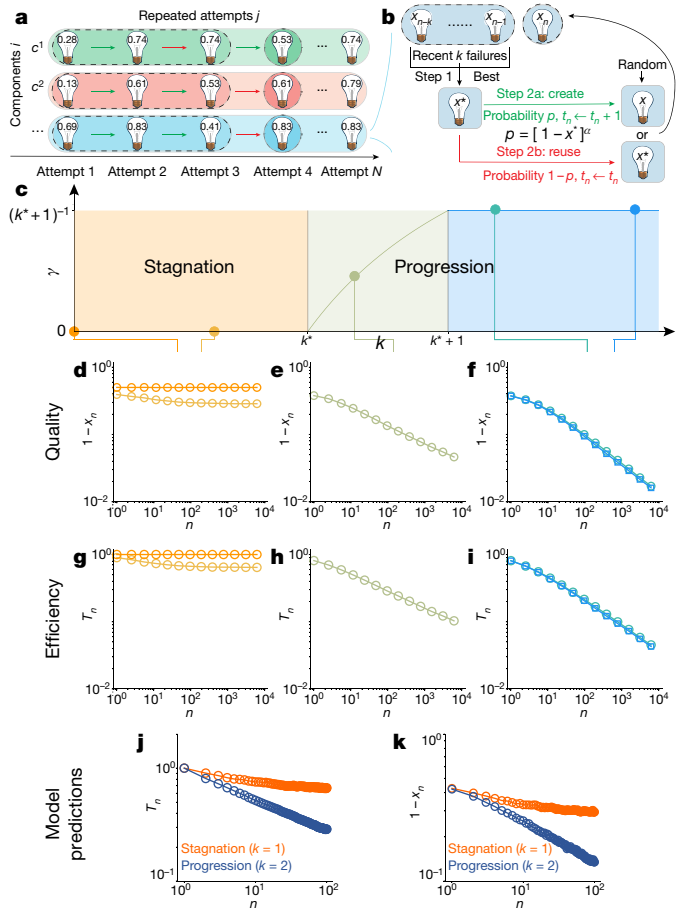


Fig. 2 | The *k* model. a, We treat each attempt as a combination of many independent components (c^i). For attempt j, each component i is characterized by an evaluation score x^i_j, which falls between 0 and 1. The score for a new version is often unknown until attempted, hence a new version is assigned a score, drawn randomly from the range 0–1. **b**, To formulate a new attempt, one can either create a new version (with probability p, green arrow) or reuse an existing version by choosing the best one among past versions x^* (with probability $1-p$, red arrow). $P(x \geq x^*) = 1 - x^*$ captures the potential to improve on prior versions, prompting us to assume $p = (1-x^*)^\alpha$ where $\alpha > 0$ characterizes the propensity of an agent to create new versions given the quality of existing ones. **c**, The analytical solution of the model reveals that the system is separated into three regimes by two critical points k^* and k^*+1. The solid line shows the extended solution space of our analytical results. **d**–**i**, Simulation results from the model ($\alpha = 0.6$) for quality (**d**–**f**) and efficiency (**g**–**i**) trajectories for different k parameters, showing distinct dynamical behaviour in different regimes. All results are based on simulations averaged over 10^4 times. **j**, **k**, A phase transition around k^* predicts the coexistence of two groups that fall in the stagnation and progression regimes, respectively.

Fig. 1b), they reduce the need to start anew, thus increasing the efficiency of future attempts (Extended Data Fig. 1e).

These two limiting cases (Extended Data Fig. 1c, f) might lead one to suspect a gradual emergence of scaling behaviour as we learn from more failures. By contrast, as we increase parameter k, the scaling exponent γ follows a discontinuous pattern (Fig. 2c, Supplementary Information 3.4) and only varies within a narrow interval of $\lfloor k^* \rfloor < k < \lceil k^* \rceil + 1$ where $k^* \equiv 1/\alpha$. Indeed, when k is small ($k < k^*$), the system converges back to the same asymptotic behaviour as $k = 0$ (Fig. 2c, d, g). In this region, k is not large enough to retain a good version once it appears. As a result, while performance might improve slightly in the first few attempts, it quickly saturates. In this region, agents reject previous attempts and flail around for new versions, not processing enough feedback to initiate a pattern of intelligent improvement, prompting us to call it the stagnation region. Once k passes the critical threshold k^*, however, scaling behaviour emerges (Fig. 2c, e, h), indicating that the system enters a region of progression, in which failures lead to continuous improvement in both quality and efficiency. Nevertheless, with a single additional experience considered, the system quickly hits the second critical point k^*+1, beyond which the scaling exponent γ becomes independent of k (Fig. 2c, f, i). This means that once $\lceil k^* \rceil + 1$ number of previous failures is considered, the system is characterized by the same dynamical behaviour as $k \to \infty$, indicating that $\lceil k^* \rceil + 1$ attempts are sufficient to recover the same rate of improvement as considering every failure from the past.

Importantly, the two critical points in our model can be mapped to phase transitions within a canonical ensemble consisting of three energy levels (Extended Data Fig. 1g–j, Methods, Supplementary Information 3.5). Phase transitions indicate that small variations at the microscopic level may lead to fundamentally different macroscopic behaviours. For example, two individuals near the critical point may initially appear identical in their learning strategy or other characteristics, but depending on which region they inhabit, their outcomes following failures could differ considerably (Fig. 2j, k). In the progression region ($k > k^*$), agents exploit rapid refinements to improve through past feedback. By contrast, those in the stagnation region ($k < k^*$) do not seem to profit from failure, as their efforts stall in efficiency and saturate in quality. As such, the phase transitions uncovered in our simple model make four distinct predictions, which we now test directly in the contexts of science, entrepreneurship and security.

Testing model predictions

Not all failures lead to success

Although we tend to focus on examples that eventually succeeded following failures, the stagnation region predicts that there exists a non-negligible fraction of cases that do not succeed following failures. We measure the number of failed cases that did not achieve eventual success in our three datasets, finding not only that members of the unsuccessful group exist, but also that the size of the unsuccessful group is of a similar order of magnitude as the successful group (Fig. 3a–c). Notably, the number of consecutive failures before the last attempt for the unsuccessful group follows a statistically similar distribution from those that lead to success (Fig. 3a–c), suggesting that people who ultimately succeeded did not try more or less than their unsuccessful counterparts.

Early signals for ultimate success or failure

Our model predicts that the successful group is characterized by power-law temporal scaling (Eq. (1)), which is absent for the unsuccessful group (Fig. 2j), predicting that the two groups may follow fundamentally different failure dynamics that are distinguishable at an early stage. To test this prediction, we measure the average inter-event time between two failures T_n as a function of the number of failures (Supplementary Information 5.3). Figure 3d–f shows three important observations. First, for the successful group, T_n decays with n across all three domains, approximately following a power law, as captured by Eq. (1) (Extended Data Fig. 2, Supplementary Information 5.3, Supplementary Table 4). The scaling exponents are within a similar range as those reported in learning curves[15], further supporting the validity of power-law scaling. Although the three datasets are among the largest in their respective domains, agents with a large number of failures are exceedingly rare, limiting the range of n that can be measured empirically. We therefore test whether alternative functions may offer a better fit, finding a power law to be the consistently preferred choice (Supplementary Information 6.2). Second, we found that temporal scaling disappears when we measure the same quantity for the unsuccessful group (Fig. 3d–f), consistent with predictions about the stagnation region. Third, the two groups show distinguishable failure dynamics as early as $n = 2$,

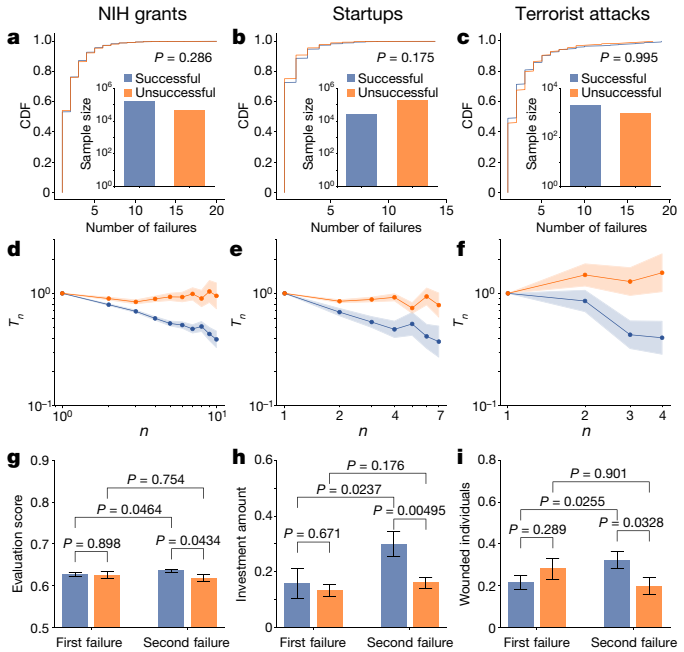


Fig. 3 | Testing model predictions. a–**c**, Cumulative distribution function (CDF) of the number of consecutive failures before the last attempt for successful and unsuccessful groups. To eliminate the possibility that agents were simply in the process of formulating their next attempt, we focus on cases for which it has been at least five years since their last failure. In each of our three datasets, the two distributions are statistically indistinguishable (Kolmogorov–Smirnov test for samples with at least one failures). For clarity, here we show results for less than 21 failures (Supplementary Information 5.2). Inset, the sample size of successful and unsuccessful groups, showing their size is of a similar order of magnitude. **d**–**f**, Early temporal signals separate successful and unsuccessful groups. **d**, $n = 43{,}705$ (successful), 15,132 (unsuccessful). **e**, $n = 2{,}455$ (successful), 16,656 (unsuccessful). **f**, $n = 446$ (successful), 321 (unsuccessful). For each group, we measure the average inter-event time between two failures $T_n \equiv t_n/t_1$ as a function of the number of attempts. Dots and shaded areas are mean ± s.e.m. measured from data (Supplementary Information 5.3). All successful groups manifest power-law scaling $T_n \sim n^{-\gamma}$ (Extended Data Fig. 2). The two groups show distinguishable temporal dynamics for $n = 2$. Two-sided Welch's t-test; $P = 3.02 \times 10^{-4}$, 7.18×10^{-3}, 9.42×10^{-2} for comparisons of successful and unsuccessful groups in **d**, **e**, **f** respectively. This temporal scaling is absent for unsuccessful groups. **g**–**i**, Performance at first attempt appears indistinguishable between successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3, two-sided Welch's t-test), but becomes distinguishable at the second attempt (two-sided Welch's t-test). Whereas performance improves for the successful group (one-sided Welch's t-test), this improvement is absent for the unsuccessful group (one-sided Welch's t-test). Data are mean ± s.e.m. **g**, $n = 628$, 145, 571, 123 (from left to right). **h**, $n = 248$, 1,332, 237, 1,312 (from left to right). **i**, $n = 231$, 173, 229, 174 (from left to right).

suggesting noteworthy early signals that separate those who eventually succeed from those who do not.

Observations uncovered in Fig. 3d–f are notable for two main reasons. First, failures captured by the three datasets differ widely in their scope, scale, definition and temporal resolution, yet despite these differences, they are characterized by remarkably similar dynamical patterns predicted by our simple model. Second, although one might expect that the last attempt was crucial in separating the two groups, as the model predicts, successful and unsuccessful groups each follow their respective, highly predictable patterns, which are distinguishable long before the eventual outcome becomes apparent. Indeed, we use D_1 to set up a prediction task (Extended Data Fig. 3, Methods, Supplementary Information 6.1) to predict ultimate success or failure using only temporal features, which yielded substantial predictive power. To test whether the observed patterns in Fig. 3d–f may simply reflect preexisting population differences, we take agents who experienced a large number of failures, and measure performance from their first attempt. We find that for all three domains, the two populations were statistically indistinguishable in their initial performance (Fig. 3g–i), which leads us to the next prediction.

Diverging patterns of performance improvement

Although the two groups may have begun with similar performances, the model predicts that they may experience different performance gains through failures (Fig. 2k). We compared performance at first and second attempts, finding significant improvement for the successful group (Fig. 3g–i), which is absent for the unsuccessful group. We further repeated our measurements by comparing the first and penultimate or halfway attempt, arriving at the same conclusion (Extended Data Fig. 9j–o, Supplementary Information 7.3). This prediction explains the patterns that were observed in Fig. 1c–e, which leads us to the second puzzle described in Fig. 1h–j: if performance improves, why are failure streaks longer than we expect?

Failure streaks follow a Weibull distribution

One key difference between progression and stagnation regimes is the propensity to reuse past components. From the perspective of exploration versus exploitation[27,28], however, reuse helps one to retain a good version when it appears, but it could also keep one in a suboptimal position for longer, leading to our final prediction: the length of failure streaks follows a Weibull distribution (Supplementary Table 4):

$$P(N \geq n) \sim e^{-(n/\lambda)^{\beta}} \tag{2}$$

Moreover, the shape parameter β is connected with the temporal scaling exponent γ through a scaling identity (Supplementary Information 3.8)

$$\beta + \gamma = 1 \tag{3}$$

This means that if we fit the streak length distribution in Fig. 1h–j to obtain the shape parameter β, it should relate to the temporal scaling exponent γ, which is obtained from Fig. 3d–f. Comparing β and γ measured independently across all three datasets shows consistency between our data and the scaling identity Eq. (3) (Supplementary Table 4).

We test the robustness of our results along several dimensions, arriving at broadly consistent conclusions (Methods, Extended Data Figs. 5–9, Supplementary Information 7). We include further quantitative tests for model assumptions and additional interpretations of the model in the Methods.

Discussion

As a single parameter, k necessarily combines individual, organizational and environmental factors in learning[19,22] (Supplementary Information 3.1). The one-parameter model developed here represents a minimal model (Supplementary Information 3.7), which can be extended into more complex frameworks. For example, agents may have varied incentives to improve or may differ in their confidence and ability to judge their previous work. Such factors trace heterogeneity in the population and can be captured by the α parameter, which quantifies the propensity of individuals to change given feedback. This led us to develop the k–α model (Methods), which predicts a two-dimensional phase diagram with three distinct phases (Extended Data Fig. 10a, b, Methods, Supplementary Information 4.1). The model can be further extended to capture fuzzy inference from past feedback, allowing agents to not always choose the best previous versions (see 'k–α–δ

model' in the Methods, Extended Data Fig. 10c, d, Supplementary Information 4.2).

The model also offers relevant insights for the understanding of learning curves. For example, the second critical point of the model suggests the existence of a minimum number of failures one needs to consider (k^*+1), indicating that it is unnecessary to learn from all past experiences to achieve a maximal learning rate. This finding poses a potential explanation for the widespread nature of Wright's law across a wide variety of domains, particularly given the fact that in many of those domains not all past experiences can be considered (Supplementary Information 2).

Furthermore, our simple model does not explicitly account for many of the complexities that characterize real settings that may affect failure dynamics, such as knowledge depreciation[29], competition, forgetting and transfer[12] or vicarious learning from others[30]. However, the model offers a theoretical basis to incorporate additional factors, including individual and organizational characteristics that may affect learning[12,17] (see Methods for various factors related to learning rate, including organizational learning, previous achievements and gender differences), demonstrating that our modelling framework can serve as a springboard for anchoring future models and analyses.

Concluding remarks

Together, these results support the hypothesis that if future attempts systematically build on past failures, the dynamics of repeated failures may reveal statistical signatures that are discernible at an early stage. Traditionally the main distinction between ultimate success and failure following repeated attempts has been attributed to differences in luck, learning strategies or individual characteristics, but here our model offers an important explanation with crucial implications: Even in the absence of distinguishing initial characteristics, agents may still experience fundamentally different outcomes. Indeed, Thomas Edison once said, 'Many of life's failures are people who did not realize how close they were to success when they gave up.' Our results unveil identifiable early signals that help us to predict the eventual outcome to which failures lead. Together, they not only deepen our understanding of the complex dynamics beneath failure, but also hold lessons for individuals and organizations that experience failure and the institutions that aim to facilitate or hinder their eventual breakthrough.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1725-y.

1. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
2. Harford, T. *Adapt: Why Success Always Starts with Failure* (Farrar, Straus and Giroux, 2011).
3. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
4. Jones, B. F. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
5. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).
6. Liu, L. et al. Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**, 396–399 (2018).
7. Hu, Y., Havlin, S. & Makse, H. A. Conditions for viral influence spreading through multiplex correlated social networks. *Phys. Rev. X* **4**, 021031 (2014).
8. Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
9. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
10. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
11. Malmgren, R. D., Stouffer, D. B., Campanharo, A. S. & Amaral, L. A. N. On universality in human correspondence activity. *Science* **325**, 1696–1700 (2009).
12. Argote, L. *Organizational Learning: Creating, Retaining and Transferring Knowledge* (Springer Science & Business Media, 2012).
13. Sitkin, S. B. Learning through failure: the strategy of small losses. *Res. Organ. Behav.* **14**, 231–266 (1992).
14. Yelle, L. E. The learning curve: historical review and comprehensive survey. *Decis. Sci.* **10**, 302–328 (1979).
15. Dutton, J. M. & Thomas, A. Treating progress functions as a managerial opportunity. *Acad. Manage. Rev.* **9**, 235–247 (1984).
16. Huber, G. P. Organizational learning: the contributing processes and the literatures. *Organ. Sci.* **2**, 88–115 (1991).
17. Cannon, M. D. & Edmondson, A. C. Failing to learn and learning to fail (intelligently): how great organizations put failure to work to innovate and improve. *Long Range Plann.* **38**, 299–319 (2005).
18. Kaplan, S. N. & Lerner, J. in *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges* (Univ. Chicago Press, 2016).
19. Eggers, J. P. & Song, L. Dealing with failure: serial entrepreneurs and the costs of changing industries between ventures. *Acad. Manage. J.* **58**, 1785–1803 (2015).
20. National Consortium for the Study of Terrorism and Responses to Terrorism. *Global Terrorism Database (GTD)* https://www.start.umd.edu/research-projects/global-terrorism-database-gtd (2018).
21. Clauset, A. & Gleditsch, K. S. The developmental dynamics of terrorist organizations. *PLoS ONE* **7**, e48633 (2012).
22. Johnson, N. et al. Pattern in escalations in insurgent and terrorist activity. *Science* **333**, 81–84 (2011).
23. Newell, A. & Rosenbloom, P. S. in *Cognitive Skills and their Acquisition* **1** (ed. Anderson, J. R.) 1–55 (Erlbaum, 1981).
24. Anderson, J. R. Acquisition of cognitive skill. *Psychol. Rev.* **89**, 369–406 (1982).
25. Muth, J. F. Search theory and the manufacturing progress function. *Manage. Sci.* **32**, 948–962 (1986).
26. Wright, T. P. Factors affecting the cost of airplanes. *J. Aeronaut. Sci.* **3**, 122–128 (1936).
27. March, J. G. Exploration and exploitation in organizational learning. *Organ. Sci.* **2**, 71–87 (1991).
28. Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
29. Arbesman, S. *The Half-life of Facts: Why Everything We Know Has an Expiration Date* (Penguin, 2013).
30. Madsen, P. M. & Desai, V. Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. *Acad. Manage. J.* **53**, 451–476 (2010).

Methods

Model assumptions

Parameter k in our model can be viewed as approximating the 'memory' of past versions. The rationale of using k for the model is rooted in the learning literature showing that the general notion of 'forgetting' takes multiple forms, often representing a combination of individual, organizational and environmental factors. Indeed, several relevant factors may be at play, which can generate patterns similar to forgetting. For example, in rapidly shifting innovation domains, not all past failures remain useful over time and some become obsolete. Consider the possibility of knowledge depreciation[31], which could also apply in our settings as environments (of scientific knowledge, capital markets or security situations) evolve over time, such that past experience could become useless even if memorized. For example, an NIH proposal four failures ago may become irrelevant as the ideas proposed have been proven wrong, or published by the principal investigator or another research group[32,33]. Similarly, startup ideas from the dot-com era may be irrelevant in the era of artificial intelligence and Blockchain[34]. Terrorist tactics can also depreciate over time, as past strategies attracted media coverage and gave rise to tighter security measures to defend against them[22]. This line of reasoning supports the intuition that recent attempts are most relevant. It is also consistent with the learning literature, which suggests knowledge forgetting can happen in distinct ways, either voluntarily or involuntarily[35]. Given these factors, here we select a single parameter k to encapsulate a variety of potential contributing factors.

Quantifying component dynamics

To empirically measure the dynamics of components, we collected abstract information for all R01 applications submitted after 2008 (Supplementary Information 5.4). To this data corpus we applied a natural-language-processing technique to extract MeSH (medical subject headings) terms from each abstract, which approximate the methods, physical states and processes involved in the proposed research. This allows us to quantify the dynamics of component reuse from previous proposals for the successful group. We measure the new versions of components by the number of new MeSH terms (terms that did not appear in the previous k submissions, defined as m_n) and plot $M_n \equiv \langle m_n \rangle / \langle m_1 \rangle$ as a function of n. Our model suggests that given k, we can use M_n to mimic the temporal dynamics of T_n. More precisely, for the successful group, we expect to observe that for large k ($k > k^*$), M_n and T_n are characterized by similar dynamics. For small k ($k < k^*$), however, the two quantities could be quite different. As shown in Extended Data Fig. 4, our empirical analysis shows that the two curves indeed follow different dynamics for small k ($k \leq 3$), but the dynamics of M_n and T_n become statistically indistinguishable for $k > 3$ (from 4 to ∞), approximately following a power law with $\gamma \sim 0.35$. We cannot directly examine component dynamics for the unsuccessful group due to the lack of sufficient data—by definition agents in this group submitted no proposal after 2010, and the unsuccessful abstract data only go back to 2008.

Phase transitions

To understand the nature of two transition points in our model, here we consider a canonical ensemble of N particles ($N \to \infty$) and three energy states $E_a(h) = 1$, $E_b(h) = (2h-1)^2$ and $E_c(h) = 1$ where h denotes the external field. We can write down the partition function of the system $Z = e^{-NE_a(h)} + e^{-NE_b(h)} + e^{-NE_c(h)}$, and calculate its free energy density $f = \ln[Z/N]$. In this system, it can be shown that the magnetization density $m = \frac{df}{dh}$ is discontinuous at the boundary of two energy states $E_a(h) = E_b(h)$ and $E_b(h) = E_c(h)$, characterized by two phase transitions at $h = 0$ and $h = 1$, respectively.

We notice that the canonical ensemble considered above has a mapping to our model. Indeed, denoting $\Gamma \equiv k^* \times \gamma/(1-\gamma)$ and $K \equiv k - k^*$, we can rescale the system as $\Gamma = \min\{\max\{\Gamma_a(K), \Gamma_b(K)\}, \Gamma_c(K)\}$ where $\Gamma_a(K) = 0$, $\Gamma_b(K) = K$ and $\Gamma_c(K) = 1$, allowing us to map the two systems through $f \to (2\Gamma - 1)^2$, $N \to \ln[n]$, $h \to K$ and $E_i(h) = (2\Gamma_i(K) - 1)^2$ (Extended Data Fig. 1g–j).

To understand the origin of the two transition points, we can calculate the expected lifespan of a high-quality version, obtaining $\langle u(x) \rangle \sim \langle (1-x)^{-\min\{k/k^*, 1/k^*+1\}} \rangle$ (Supplementary Information 3.4). The first critical point k^* occurs when the first moment $\langle u \rangle$ diverges. Indeed, when k is small ($k < k^*$), $\langle u \rangle$ is finite, indicating that high-quality versions can only be reused for a limited period of time. Once k passes critical point k^*, however, $\langle u \rangle$ diverges, offering the possibility for a high-quality version to be retained for an unlimited period of time. The second critical point arises due to the competition between two dynamical forces: (1) whether the current best version becomes forgotten after k consecutive attempts in creating new versions (dominated by the k/k^* term); or (2) it is substituted by an even better version (dominated by the $1/k^* + 1$ term).

Note that while phase transitions carry exceptional importance in statistical physics, similar phenomena and concepts are also of fundamental relevance in the social and behavioural science literatures. For example, critical thresholds have been observed and modelled in social settings that include shifts in the segregation of neighbourhoods[36], the formation of social networks[37] and changes in collective opinions[38]. In each case, slight shifts in microscale phenomena, such as average preference, group size or interaction intensity, condition a qualitative transition in macroscale outcomes.

Alternative hypothesis, interpretation and robustness checks

To better understand the role of heterogeneity in learning, we separated the successful group into narrow-win and clear-win subgroups based on their eventual performance. We find that, despite their eventual difference, the temporal dynamics of the two groups remain statistically indistinguishable (two-sided Welch's t-test, $P = 0.763$ (D_1), 0.813 (D_2), 0.259 (D_3), Extended Data Fig. 4), suggesting that the distinction between successful and unsuccessful groups appears the most critical, whereas agents within the successful group are characterized by similar dynamics, consistent with the predictions of our model.

An alternative interpretation for the stalled efficiency of the unsuccessful group is an effort to hedge against failures—their efficiency did not improve because they spent more effort elsewhere. The three professions that we studied, NIH investigators, entrepreneurs and terrorists, involve varied levels of risk, exposure and commitment, which renders this explanation less likely.

To test the robustness of our results, we vary the definitions of what constitutes the successful group (Supplementary Information 7.1) by excluding revisions in D_1 (Extended Data Fig. 6), changing the threshold of high-value mergers and acquisitions or controlling for unicorn companies in D_2 (Extended Data Fig. 7), and varying the types of attack or changing the threshold for fatal attacks in D_3 (Extended Data Fig. 8). We also vary the definition of unsuccessful groups (Extended Data Fig. 5, Supplementary Information 7.2) and test other measures to approximate performance (Extended Data Fig. 9j–o, Supplementary Information 7.4, 7.5). We further adjust for temporal variation by controlling for the overall success rate across different years (Extended Data Fig. 9a–i, Supplementary Information 7.3). Across all variations, our conclusions remain the same.

Predicting ultimate success

We use a simple logistic model to predict whether one may achieve success following N previously failed attempts in D_1, using only temporal features t_n ($1 \leq n \leq N-1$) as predictors. To evaluate prediction accuracy, we calculate the area under the receiver operating characteristic (AUC) curve with tenfold cross-validation. We find that, by observing the timing of the first three failures alone, our simple temporal feature yields high accuracy in predicting the eventual outcome with an AUC close to 0.7, which is significantly higher than random guessing (Mann–Whitney U-test, $P < 10^{-180}$; Extended Data Fig. 3a, Supplementary

Information 6.1). We repeated the same prediction task on D_2 and D_3, arriving at similar conclusions (Extended Data Fig. 3b, c, Supplementary Information 6.1). The predictive power from temporal features alone is somewhat unexpected. Indeed, there are a large number of documented factors that affect the outcome of a grant application[39–43], ranging from the previous success rate to publication and citation records to the race, ethnicity and gender of the applicant. Here we ignore these factors, however, using only features that pertain to temporal scaling as prescribed by our model. This suggests that our predictive power represents a lower bound, which could be further improved and leveraged by incorporating additional factors.

k–α model

Agents may differ in the judgment of their own work or incentives to change given feedback, which can be captured by varying the α parameter in the original k model. Of the many influences on p, one key factor is the quality of existing versions, suggesting that p should be a function of x^*. Consider the following two extreme cases. If $x^* \to 0$, existing versions of this component have one of the worst scores and, hence, a high potential for improvement when replaced with a new version. In this case, the likelihood of creating a new version is high, that is, $p \to 1$. On the other hand, $x^* \to 1$ corresponds to a near-perfect version, yielding a decreased incentive to create a new one ($p \to 0$). Indeed, $P(x \geq x^*) = 1 - x^*$ captures the potential to improve on previous versions, prompting us to assume that $p = (1 - x^*)^{\alpha}$ where $\alpha > 0$ characterizes the propensity of an agent to create new versions given the quality of existing ones. Therefore, $\alpha \to 0$ indicates that regardless of one's evaluation, the agent will always create a new version, whereas $\alpha \to \infty$ points to the other extreme where one does not create a new version unless it is extremely bad (Extended Data Fig. 10a). Considering α another tunable parameter, we arrive at a two-parameter model: the k–α model (Supplementary Information 4.1).

To solve this model we can substitute k^* with $1/\alpha$, and the indexes k/k^* and $1/k^* + 1$ now become $k\alpha$ and $\alpha + 1$. The extended model thus predicts the existence of three different phases on a two-dimensional phase diagram, with boundaries $k\alpha = 1$ and $(k-1)\alpha = 1$ that separate the three phases (Extended Data Fig. 10b). The k–α model reduces back to the two critical points in the original k model when we fix α. The two parameters jointly define an effective $K \equiv k - k^* = k - 1/\alpha$. The critical boundaries therefore reduce into two simple equations: $K = 0$ and $K = 1$. Note that the assumed relationship between p and $(1 - x^*)$ is not limited to a power law but can be relaxed into its asymptotic form. Indeed, we show that as long as the function satisfies $\frac{\ln[p]}{\ln(1-x^*)} \to \alpha$ as $x^* \to 1$, the model offers the same predictions[25] (Extended Data Fig. 3, Supplementary Information 3.6).

k–α–δ model

Agents may have fuzzy or unclear inference regarding past feedback, and may therefore not always choose the version with highest quality. We can model the choice between different versions in a probabilistic fashion, by introducing a δ parameter to the k–α model. Here the probability to choose the ith version as a baseline follows

$$P(i) = \frac{1}{Z}(1 - x_i)^{-\delta} 1_{n-k \leq i \leq n-1}$$

where Z is the normalization factor, $Z \equiv \sum_{i=n-k}^{n-1}(1 - x_i)^{-\delta}$ and $k \geq 1$. $\delta = 0$ means one cannot differentiate between the quality of past versions and selects randomly among different versions, whereas $\delta \to \infty$ indicates that one always chooses the previous version with highest quality, converging back to our original k model or the k–α model. Incorporating δ leads to the k–α–δ model (Supplementary Information 4.2).

Analytically solving the model reveals interesting scaling behaviours based on δ (Supplementary Information 4.2). Indeed, we find the scaling behaviour of the system follows

$$\gamma(k, \alpha, \delta) = 1 - \{\max[\min(\alpha + (k-1)\min\{1, \alpha, \delta\}, \alpha + 1), 1]\}^{-1}$$

with rich mathematical properties. When $\delta \to \infty$, the new solutions converge back to the original solution for the k–α model. With δ, the three-parameter model is characterized by four different phases. Three of the regimes are generalizations of those found in the k–α model, where the scaling exponent γ does not depend on δ in the limit of $\delta \to \infty$, that is, $\gamma(k, \alpha, \delta) = \gamma(k, \alpha, \infty)$. The fourth one, however, is a new phase and only exists for small δ. The intuition is that in this regime the inability to select a high-quality version (small δ) dominates the scaling behaviour, with exponent $\gamma(k, \alpha, \delta) = 1 - [(k-1)\delta + \alpha]^{-1}$. Together, these extensions offer further support for the predictions of our original model, while demonstrating the theoretical potential of the model by enriching its mathematical properties for more realistic interpretations. They also point to promising future research that explores the interplay between different perspectives on learning.

Note that although all three variations of the model predict the existence of different phases, the primary focus of this paper concerns the fundamental differences in the nature of these regimes (that is, stagnation versus progression), rather than the behaviour of the system as it approaches the critical threshold. As such, the conclusions of the paper hold the same regardless of any specific critical behaviour around the threshold.

Factors related to learning rate

Our model offers a framework to anchor potential factors relevant to learning[44,45]. As an example, here we test three different factors. First, the literature has identified several factors for the emergence of learning at the level of organizations[12,21], suggesting that individual learning is just one factor in how and why organizations learn. This suggests that settings closer to organizational learning (such as terrorist groups) should correspondingly experience higher learning rates than those closer to individual learning (such as NIH principal investigators) (Supplementary Information 5.5). We test this hypothesis by calculating the average scaling exponent γ measured from our data (Supplementary Table 4), and find that our estimations support this hypothesis; learning rates are lowest for individual researchers, higher for entrepreneurs and their founding teams and highest for terrorist organizations. Note that although these results show consistency with theories from the organizational learning literature, these differences could also be due to inherent domain-specific differences.

Second, higher previous achievements often bring recognition and resources, a phenomena referred to as the Matthew effect[46], which might translate into higher learning rates. To test this we link NIH grant application data to the Web of Science citation database through a systematic effort to disambiguate authors, and match the citations of previous research papers with submitted proposals[5,47] (Supplementary Information 5.6). We take principal investigators who failed more than three times before their eventual success and calculate the total number of citations from all his/her papers including only papers published before the first failure. We find that prior acclaim is positively and significantly correlated with learning rate γ ($P < 0.001$).

Third, persistent gender inequalities in science and entrepreneurship[48–50] suggest the possibility that failure dynamics may be mediated by gender[51,52]. Our regression analysis reveals a significant correlation between gender and learning rate (Supplementary Information 5.7). All else being equal, the learning rate γ of a male principal investigator in the NIH system exceeds that of a female principal investigator by 0.14 ($P = 0.001$), suggesting that male principal investigators fail faster than their female colleagues. This difference appears substantial, considering that the average learning rate is centred around 0.35. We further test this relationship in the startup dataset, finding a similar gap of 0.10 between male and female innovators, but this result is not significant, possibly owing to a smaller sample size. Note that these gender differences probably flow from institutional as well as individual causes, such

as a culture that discourages women from persistence and encourages oversensitivity to feedback. Indeed, one irony suggested by our model is that agents in the stagnation region did not work less. Rather they made more, albeit unnecessary modifications to what were otherwise advantageous experiences.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This paper makes use of restricted access data from the National Institutes of Health (NIH), protected by the Privacy Act of 1974 as amended (5 U.S.C. 552a). Deidentified data necessary to reproduce all plots and statistical analyses are freely available at https://yian-yin.github.io/quantifyFailure. Those wishing to access the raw data can apply for access following the procedures outlined in the NIH Data Access Policy document (http://report.nih.gov/pdf/DataAccessPolicy.pdf). The VentureXpert database is available from Thomson Reuters. The Global Terrorism Database is publicly available at https://www.start.umd.edu/gtd/.

Code availability

Code is available at https://yian-yin.github.io/quantifyFailure.

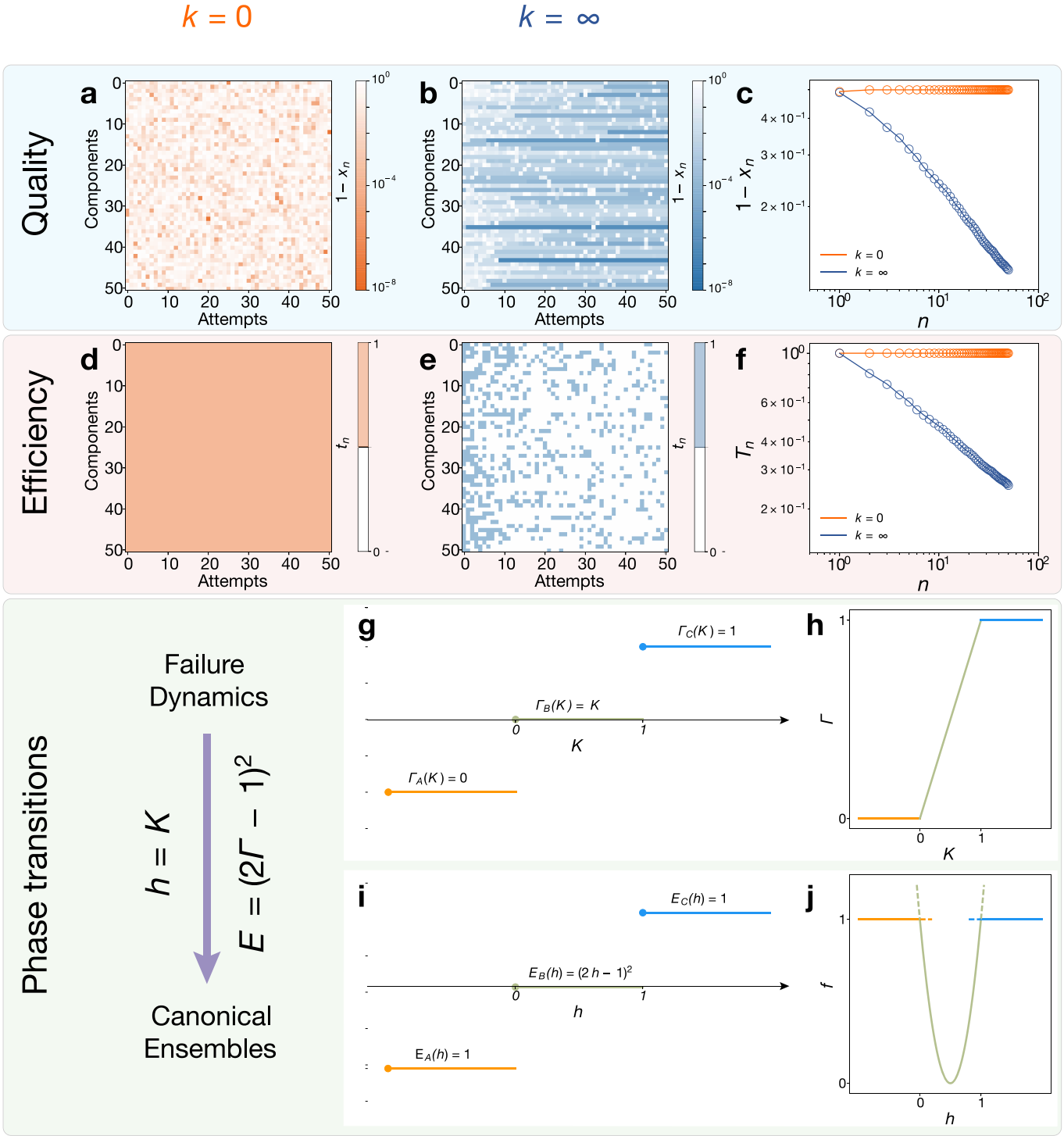
31. Argote, L., Beckman, S. L. & Epple, D. The persistence and transfer of learning in industrial settings. *Manage. Sci.* **36**, 140–154 (1990).
32. Kuhn, T. S. *The Structure of Scientific Revolutions* (Chicago Univ. Press, 2012).
33. Merton, R. K. Singletons and multiples in scientific discovery: a chapter in the sociology of science. *Proc. Am. Phil. Soc.* **105**, 470–486 (1961).
34. Gompers, P., Kovner, A., Lerner, J. & Scharfstein, D. Performance persistence in entrepreneurship. *J. Financ. Econ.* **96**, 18–32 (2010).
35. de Holan, P. M. & Phillips, N. Remembrance of things past? the dynamics of organizational forgetting. *Manage. Sci.* **50**, 1603–1613 (2004).
36. Schelling, T. C. *Micromotives and Macrobehavior* (WW Norton & Company, 2006).
37. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
38. Holme, P. & Newman, M. E. Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E* **74**, 056108 (2006).
39. Ginther, D. K. et al. Race, ethnicity, and NIH research awards. *Science* **333**, 1015–1019 (2011).
40. Boudreau, K. J., Guinan, E. C., Lakhani, K. R. & Riedl, C. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. *Manage. Sci.* **62**, 2765–2783 (2016).
41. Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).
42. Banal-Estanol, A., Macho-Stadler, I. & Pérez Castrillo, D. *Key Success Drivers in Public Research Grants: Funding the Seeds of Radical Innovation in Academia?* CESifo Working Paper Series 5852 (CESifo, 2016).
43. Ma, A., Mondragón, R. J. & Latora, V. Anatomy of funded research in science. *Proc. Natl Acad. Sci. USA* **112**, 14760–14765 (2015).
44. Levitt, B. & March, J. G. Organizational learning. *Annu. Rev. Sociol.* **14**, 319–338 (1988).
45. Argote, L. & Epple, D. Learning curves in manufacturing. *Science* **247**, 920–924 (1990).
46. Merton, R. K. et al. The Matthew effect in science. *Science* **159**, 56–63 (1968).
47. Huang, J., Ertekin, S. & Giles, C. L. Efficient name disambiguation for large-scale databases. In *European Conference on Principles of Data Mining and Knowledge Discovery* 536–544 (Springer, 2006).
48. Shen, H. Inequality quantified: Mind the gender gap. *Nature* **495**, 22–24 (2013).
49. Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: global gender disparities in science. *Nature* **504**, 211–213 (2013).
50. Yang, T. & Aldrich, H. E. Who's the boss? Explaining gender inequality in entrepreneurial teams. *Am. Sociol. Rev.* **79**, 303–327 (2014).
51. Argote, L., Insko, C. A., Yovetich, N. & Romero, A. A. Group learning curves: the effects of turnover and task complexity on group performance. *J. Appl. Soc. Psychol.* **25**, 512–529 (1995).
52. Bailey, C. D. Forgetting and the learning curve: a laboratory study. *Manage. Sci.* **35**, 340–352 (1989).

Acknowledgements We thank C. Song, A. Clauset, B. Uzzi, B. Jones, E. Finkel, J. Van Mieghem, A. Bassamboo and Y. Xie for helpful discussions, and H. Sauermann and S. Havlin for suggesting extensions of the model, leading us to discover the k–α and k–α –δ models. This work is supported by the Air Force Office of Scientific Research under award number FA9550-15-1-0162, FA9550-17-1-0089 and FA9550-19-1-0354, National Science Foundation grant SBE 1829344, the Alfred P. Sloan Foundation G-2019-12485, and Northwestern University Data Science Initiative. This work does not reflect the position of NIH.

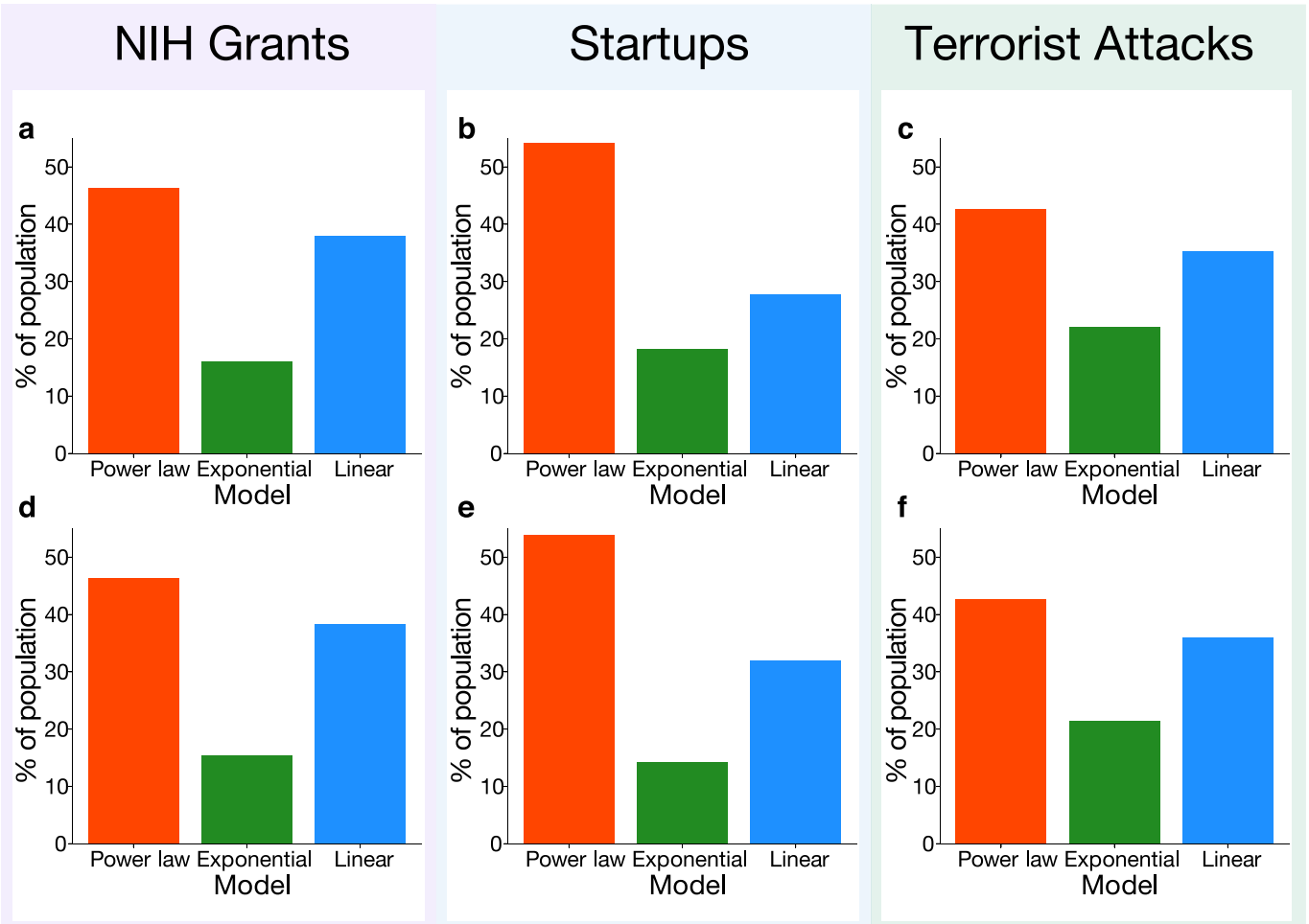
Author contributions D.W. conceived the project and designed the experiments; Y.Y. and Y.W. collected data and performed empirical analyses with help from D.W. and J.A.E.; Y.Y. and D.W. carried out theoretical calculations; all authors collaboratively designed the model and interpreted results; D.W. and Y.Y. wrote the manuscript; all authors edited the manuscript.

Competing interests Y.W. and D.W. serve as special volunteers (unpaid) to the NIH. The remaining authors declare no competing interests.

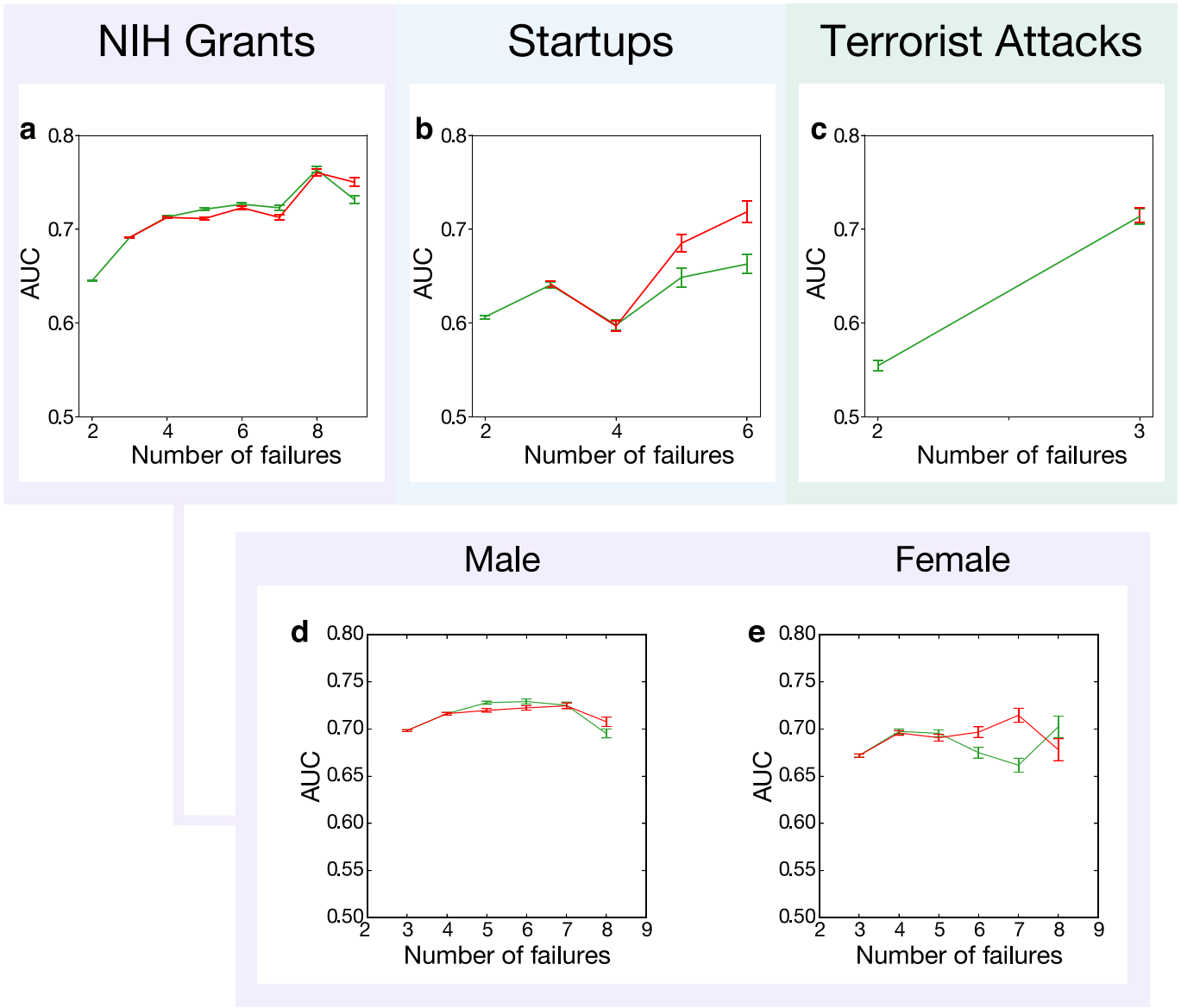
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1725-y.
Correspondence and requests for materials should be addressed to D.W.
Peer review information *Nature* thanks Shlomo Havlin and Henry Sauermann for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



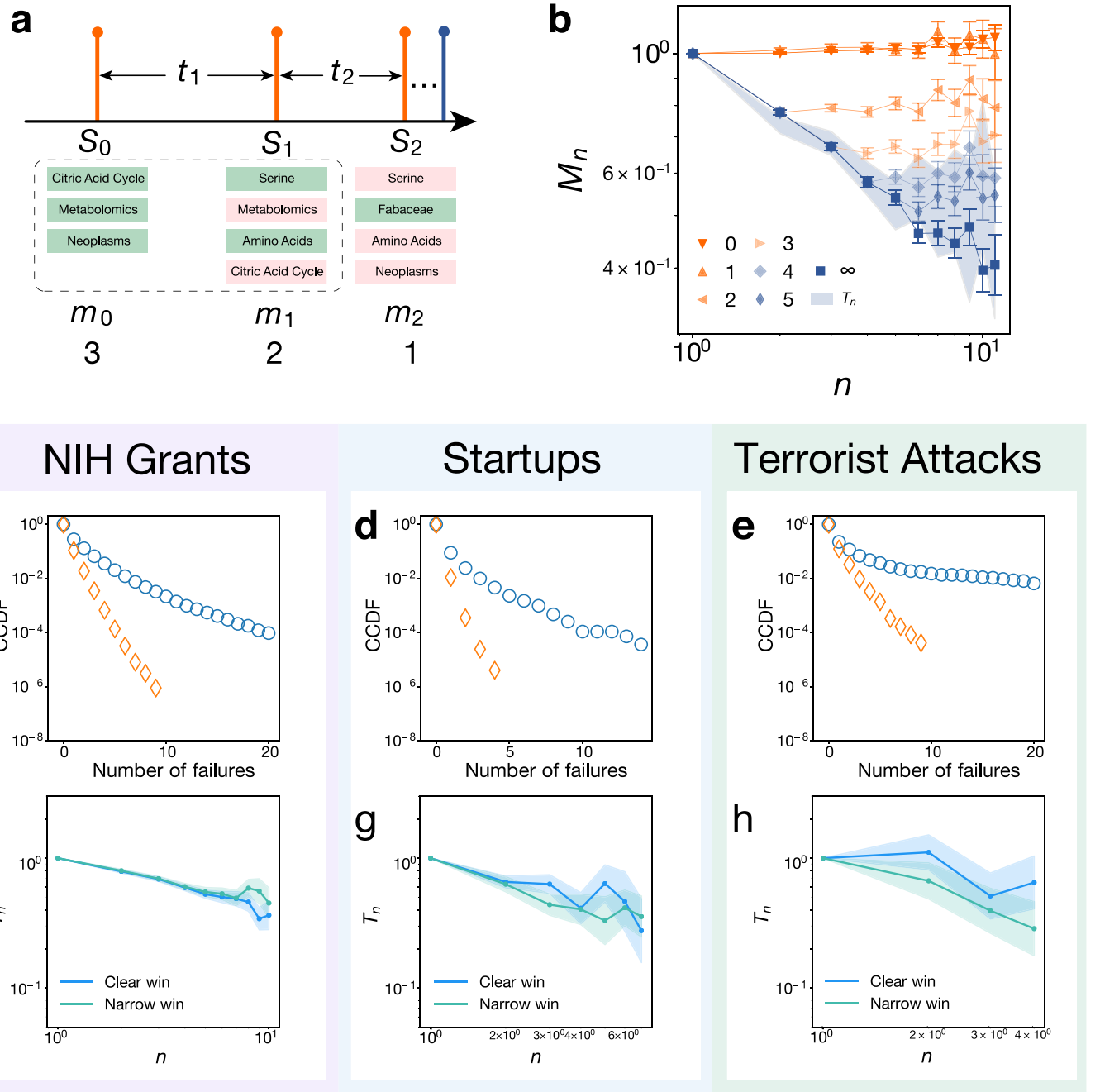
Extended Data Fig. 1 | The *k* model. a–f, Simulation results from the model ($\alpha = 0.6$) for the cases of $k = 0$ (**a**, **d**) and $k \to \infty$ (**b**, **e**) in terms of the average quality (**a**–**c**) and efficiency (**d**–**f**) of each attempt. $k = 0$ recovers the chance model, predicting a constant quality (**c**) and efficiency (**f**). $k \to \infty$ predicts temporal scaling that characterizes the dynamics of failure (**e**) with improved quality (**b**), recovering predictions from learning curves and Wright's law. **g**–**j**, Illustration of mapping between failure dynamics (**g**, **h**) and canonical ensembles (**i**, **j**). The canonical system is characterized by three different states a, b, c with corresponding energy densities $E_a(h), E_b(h), E_c(h)$. Here we assume $E_a(h) = (2\varepsilon h - 1)^2$, $E_b(h) = (2h - 1)^2$ and $E_c(h) = [2\varepsilon(1 - h) - 1]^2$ where $\varepsilon \to 0^+$. The introduction of ε is to distinguish state a from state c, both of which can be approximated in the limiting condition $E_a(h) = E_c(h) = 0$. We map $f \to (2\Gamma - 1)^2$, $N \to \ln[n]$, $h \to K$ and $E_i(h) = [2\Gamma_i(K) - 1]^2$. In this case, the two transition points k^* and $k^* + 1$ correspond to $h = 0$ and 1 in the canonical ensemble systems.



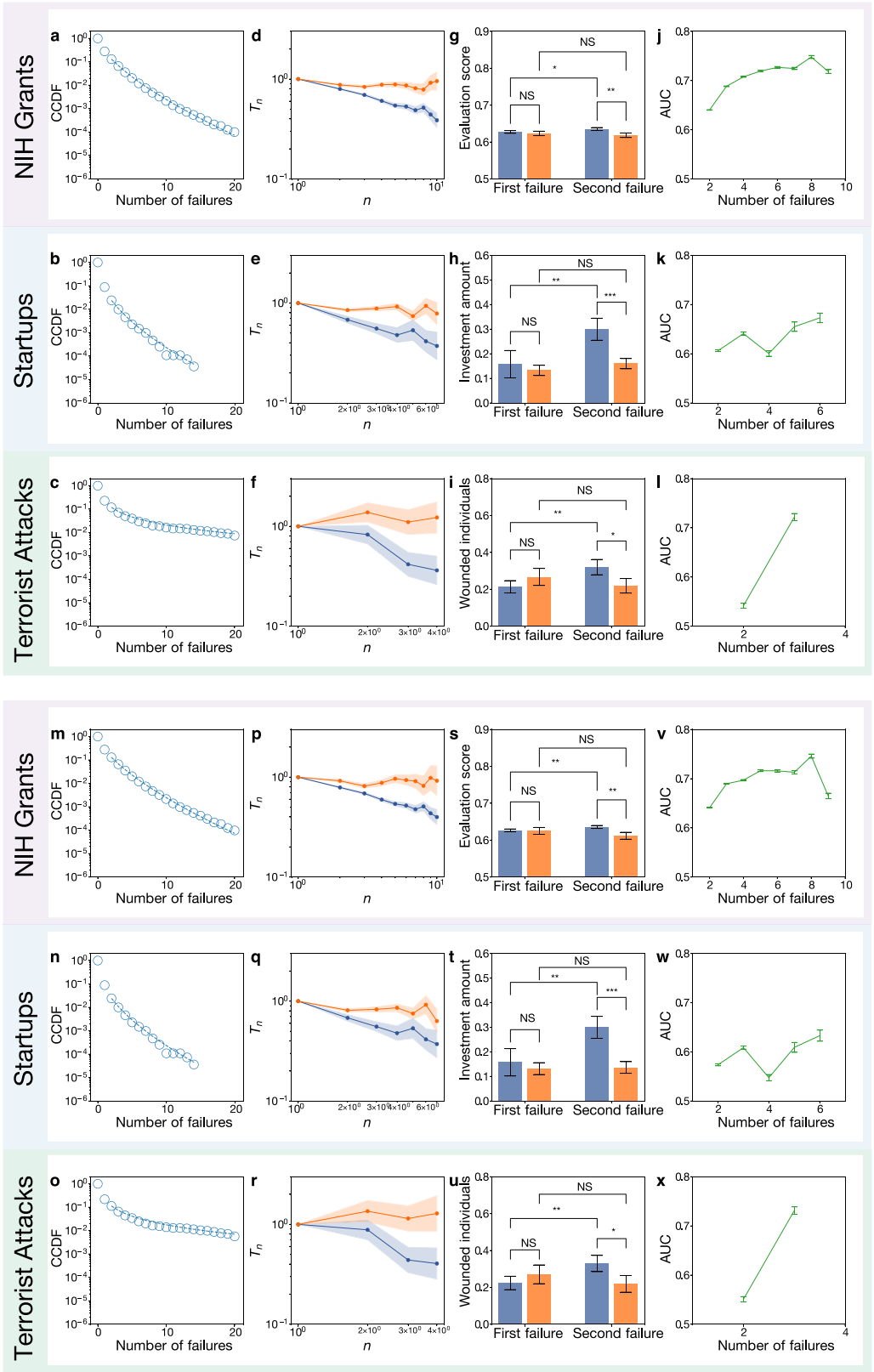
Extended Data Fig. 2 | Predicting temporal dynamics in science, entrepreneurship and security. a–**c**, We compare the goodness of fit for three different models in temporal dynamics in NIH grants (**a**, $n = 10345$), startups (**b**, $n = 275$) and terrorist attacks (**c**, $n = 136$). For each individual sample, we take all but the last inter-event time for model fitting ($n = 1, \ldots, N-1$), comparing model predictions for the last inter-event time. The tested functional forms are power law, $t_n = an^b$; exponential, $t_n = ab^{-n}$; and linear, $t_n = a + bn$. We then calculate the frequency that each model reaches minimum error, defined as $|\log(t_N) - \log(\hat{t}_N)|$, among all three forms. The power-law model offers consistently better predictions. **d**–**f**, As in **a**–**c**, but using $|t_N - \hat{t}_N|$ as the loss function.



Extended Data Fig. 3 | Predicting ultimate success in science, entrepreneurship and security. a–**c**, Area under the receiver operating characteristic curve (AUC) of the prediction task. We apply two logistic regression models (Supplementary Information 6.1) to predict ultimate success in NIH grants (**a**), startups (**b**) and terrorist attacks (**c**). The centres and error bars of AUC scores denote the mean ± s.e.m. calculated from tenfold cross-validation over 50 randomized iterations (green, model 1; red, model 2). **d**, **e**, As in **a** but predicting ultimate success in NIH grants for male (**d**) and female (**e**) investigators.

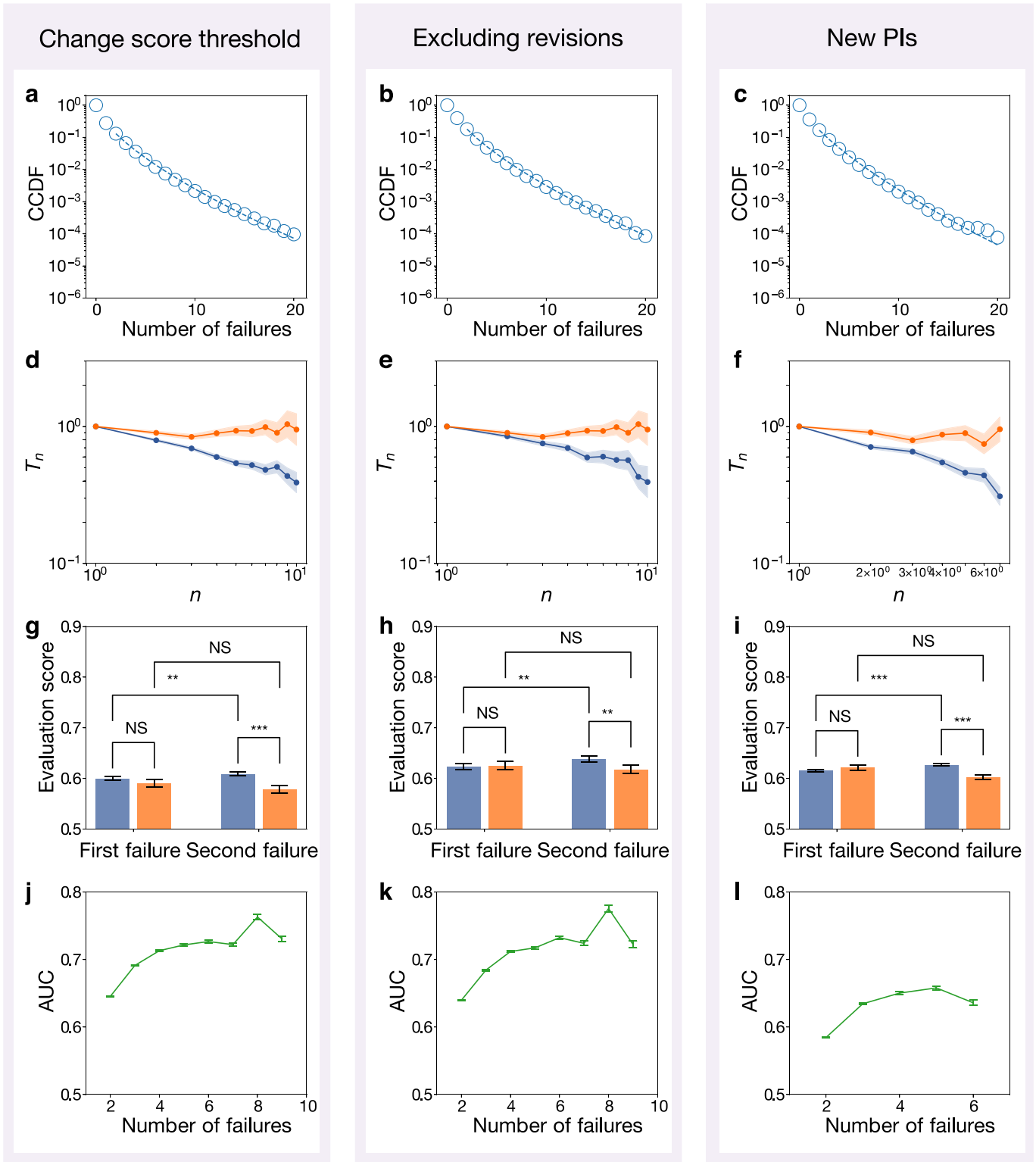


Extended Data Fig. 4 | Model validations. a, **b**, An illustration of the component dynamics. We extract all MeSH terms associated with the nth attempt, S_n, and calculate the number of new terms m_n, defined as $|S_n - (S_{n-1} \cup \cdots \cup S_{n-k})|$. **b**, Testing component dynamics in NIH grant applications. We calculate the dynamics of $M_n = \langle m_n \rangle / \langle m_1 \rangle$ using different k and compare it with T_n. The centres and error bars of M_n show the mean ± s.e.m. (n = 5,899) for different k. The shaded area shows mean ± s.e.m. of T_n (log scale) measured on the same subset. All $k > 3$ lead to similar trends between M_n and T_n. **c**–**e**, Length of failure streak after randomization in science (**c**), entrepreneurship (**d**) and security (**e**). We take the samples used in Fig. 1 and shuffle the success/failure label from each attempt. This operation keeps both the overall success rate and the total number of attempts for each individual constant. **f**–**h**, Temporal scaling patterns within the successful group in science (**f**), entrepreneurship (**g**) and security (**h**). We separated the successful group into two subgroups (narrow winners and clear winners) based on eventual performance (0.9 in evaluation score for D_1, 0.5 in investment amount for D_2 and 1 in wounded individuals for D_3). The shaded area shows mean ± s.e.m. of T_n (log scale).



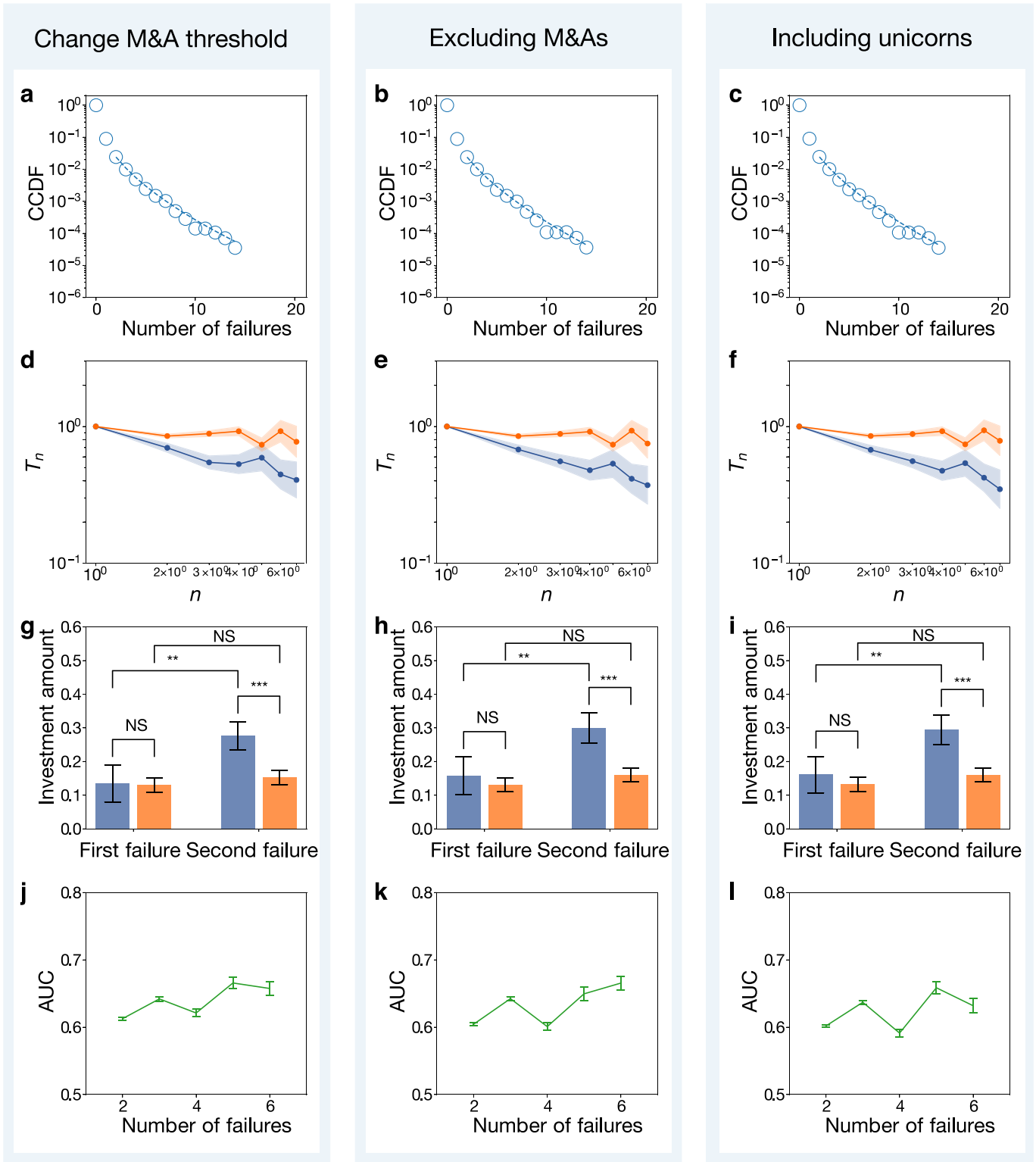
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Robustness check on definition of unsuccessful group. a–**l**, Robustness check as we change the threshold of inactivity to 3 years. **a**–**c**, Failure streak in science (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data from the successful group and dashed lines represent fitted Weibull distributions. **d**–**f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean ± s.e.m. of T_n (log scale). **g**–**i**, Performance dynamics in science (**g**, $n = 641, 231, 578, 190$, from left to right), entrepreneurship (**h**, $n = 248, 1{,}332, 237, 1{,}312$ from left to right) and security (**i**, $n = 238, 198, 236, 199$, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable for first failures (two-sided Welch's *t*-test; $P = 0.566, 0.671$ and 0.349), but quickly diverge for second failures (two-sided Welch's *t*-test; $P = 2.09 \times 10^{-2}$, 4.95×10^{-3} and 7.77×10^{-2}). The successful group also shows significant improvement in performance (one-sided Welch's *t*-test; $P = 7.03 \times 10^{-2}$, 2.37×10^{-2} and 2.32×10^{-2}), which is absent for the unsuccessful group (one-sided Welch's *t*-test; $P = 0.717, 0.176$ and 0.786). Data are mean ± s.e.m. **j**–**l**, AUC score of predicting ultimate success in science (**j**), entrepreneurship (**k**) and security (**l**). The centres and error bars of AUC scores denote the mean ± s.e.m calculated from tenfold cross-validation over 50 randomized iterations. **m**–**x**, As in **a**–**l** but using 7 years as the threshold of inactivity. Sample sizes are **s**: $n = 620, 101, 559, 76$; **t**: $n = 248, 977, 237, 989$; **u**: $n = 216, 152, 214, 153$. *P* values in **s**–**u** (from bottom to top) are $P = 0.883$ (**s**), 0.671 (**t**), 0.456 (**u**); $P = 2.25 \times 10^{-2}$ (**s**), 1.38×10^{-3} (**t**), 8.34×10^{-2} (**u**); $P = 4.59 \times 10^{-2}$ (**s**), 2.37×10^{-2} (**t**), 3.33×10^{-2} (**u**); $P = 0.838$ (**s**), 0.446 (**t**), 0.775 (**u**). $^{*}P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$, NS, not significant ($P \geq 0.1$).



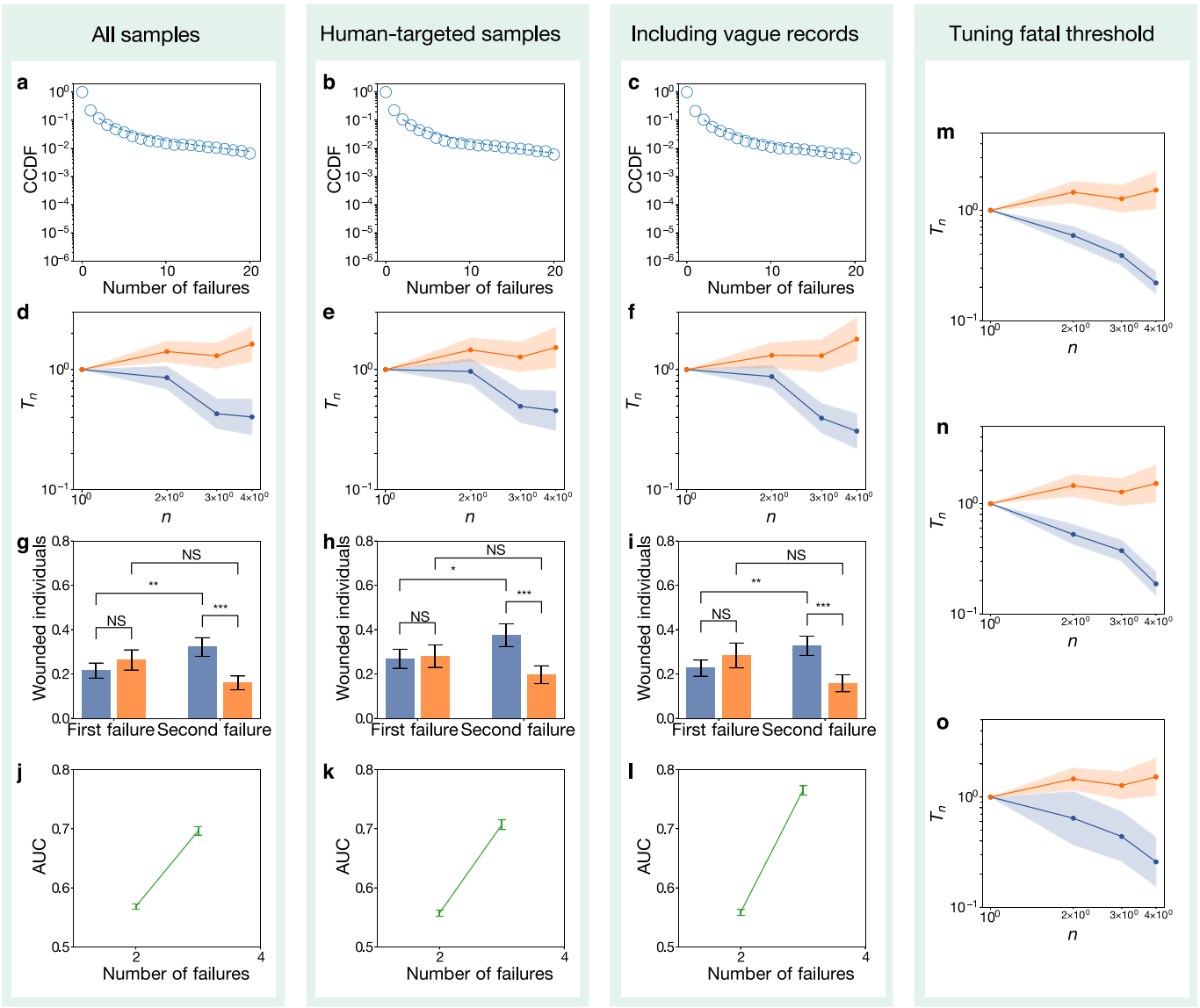
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Robustness check on D_1. a–**c**, Failure streak as we change the score threshold to 55 (**a**), exclude revisions as successes (**b**) and only focus on new principal investigators without previous R01 grants (**c**). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. **d**–**f**, Temporal scaling patterns as we change the score threshold to 55 (**d**), exclude revisions as successes (**e**) and only focus on new principal investigators without previous R01 grants (**f**). The shaded area shows mean ± s.e.m. of T_n (log scale). **g**–**i**, Performance dynamics as we change the score threshold to 55 (**g**, n = 768, 189, 686, 170, from left to right), exclude revisions as successes (**h**, n = 252, 145, 216, 123, from left to right) and only focus on new principal investigators without previous R01 grants (**i**, n = 1,164, 308, 1,530, 334, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 5 for **g** and **h**, and 3 for **i**) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.242, 0.819, 0.289) but quickly diverge for second failures (two-sided Welch's t-test; $P = 3.40 \times 10^{-4}, 3.40 \times 10^{-2}, 9.70 \times 10^{-7}$). The successful group also shows a significant improvement in performance (one-sided Welch's t-test; $P = 4.23 \times 10^{-2}$, $3.04 \times 10^{-2}, 1.92 \times 10^{-4}$), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.863, 0.754, 0.997). Data are mean ± s.e.m. **j**–**l**, AUC score of predicting ultimate success as we change the score threshold to 55 (**j**), exclude revisions as successes (**k**) and only focus on new principal investigators without previous R01 grants (**l**). The centres and error bars of AUC scores denote the mean ± s.e.m calculated from tenfold cross-validation over 50 randomized iterations. $^*P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$, NS, $P \geq 0.1$.

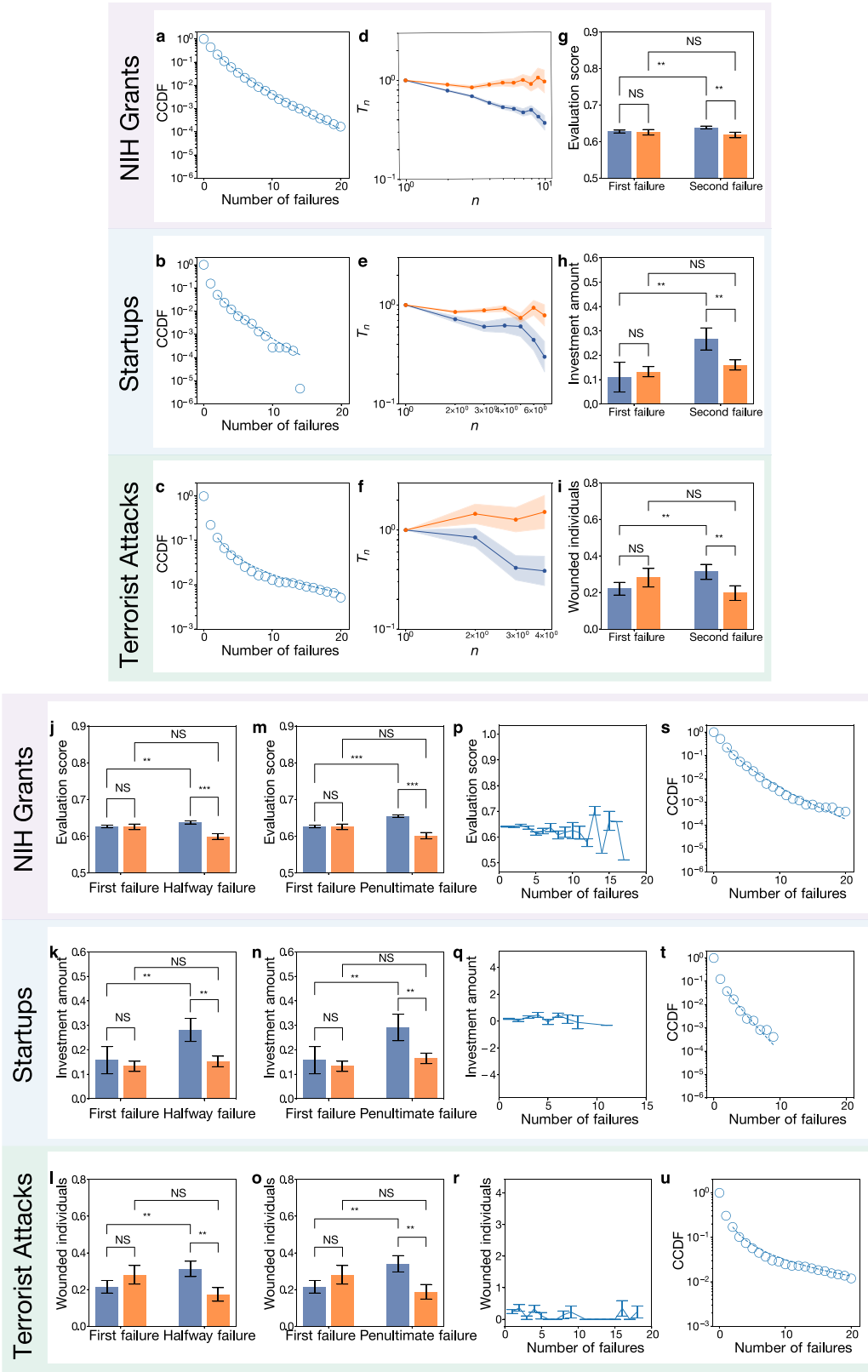


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Robustness check on D_2. a–**c**, Failure streak as we change the threshold of high-value mergers and acquisitions (M&A) to 5% (**a**), exclude M&As as successes (**b**) and classify unicorns as successes (**c**). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. **d**–**f**, Temporal scaling patterns as we change the threshold of high-value M&A to 5% (**d**), exclude M&As as successes (**e**) and include unicorns as successes (**f**). The shaded area shows mean ± s.e.m. of T_n (log scale). **g**–**i**, Performance dynamics as we change the threshold of high-value M&A to 5% (**g**, n = 251, 1,304, 243, 1,284, from left to right), exclude M&As as successes (**h**, n = 248, 1,335, 237, 1,315, from left to right) and include unicorns as successes (**i**, n = 257, 1,330, 244, 1,311, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 3) appear indistinguishable for first failures (two-sided Welch's t-test; P = 0.937, 0.647, 0.620) but quickly diverge for second failures (two-sided Welch's t-test; $P = 9.92 \times 10^{-3}, 4.94 \times 10^{-3}, 6.33 \times 10^{-3}$). The successful group also shows a significant improvement in performance (one-sided Welch's t-test; $P = 2.16 \times 10^{-2}, 2.37 \times 10^{-2}, 2.77 \times 10^{-2}$), which is absent for the unsuccessful group (one-sided Welch's t-test; P = 0.224, 0.158, 0.167). Data are mean ± s.e.m. **j**–**l**, AUC score for predicting ultimate success as we change threshold of high-value M&A to 5% (**j**), exclude M&As as successes (**k**) and include unicorns as successes (**l**). The centres and error bars of AUC scores denote the mean ± s.e.m calculated from tenfold cross-validation over 50 randomized iterations. $^{*}P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$, NS, $P \geq 0.1$.

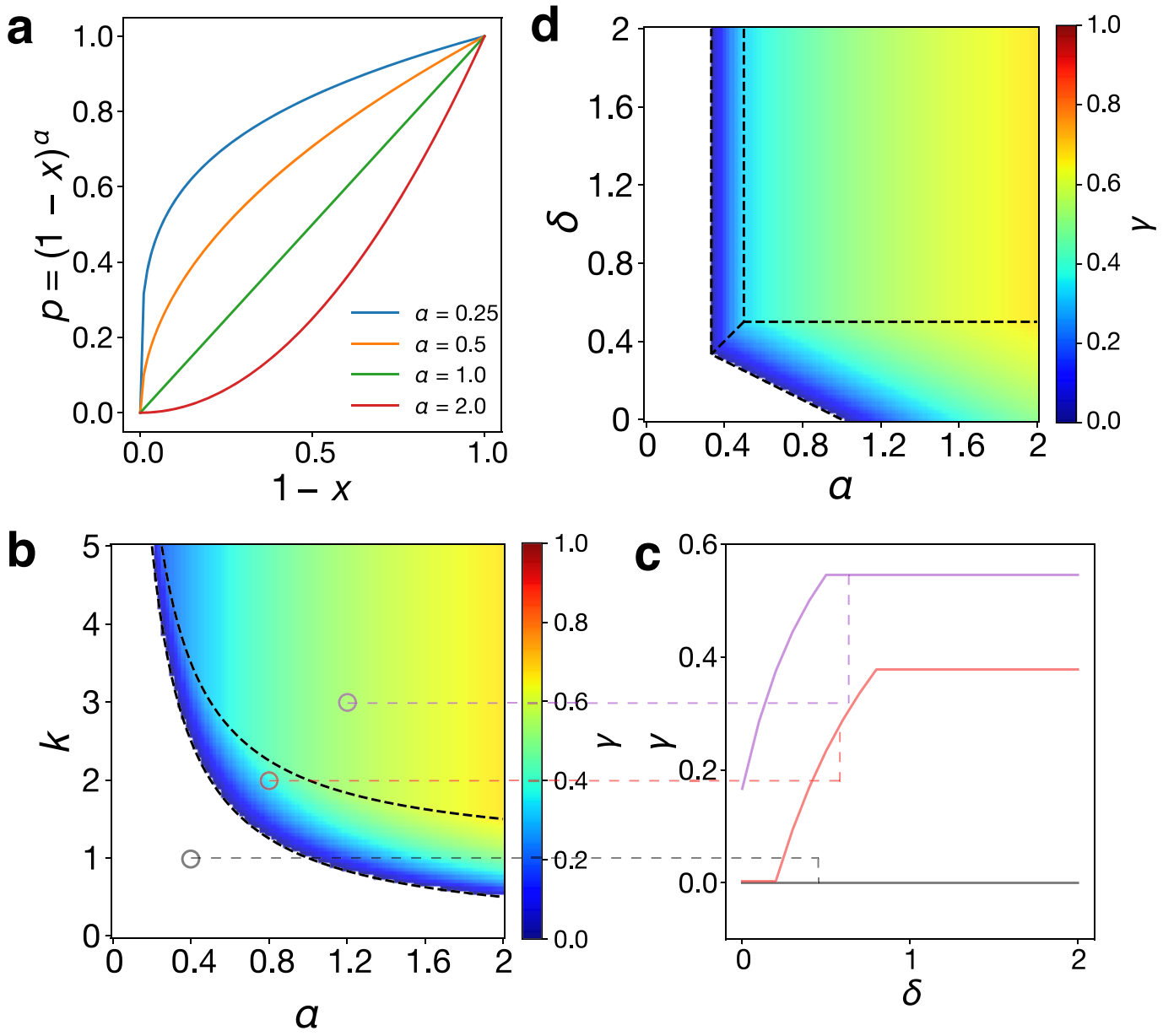


Extended Data Fig. 8 | Robustness check on D_3. **a**–**c**, Failure streak as we focus on all samples (**a**), samples of human-targeted attacks (**b**) and include vague data on fatalities (**c**). Blue circles represent real data from successful groups and dashed lines represent fitted Weibull distributions. **d**–**f**, Temporal scaling patterns as we focus on all samples (**d**), samples of human-targeted attacks (**e**) and include vague data on fatalities (**f**). The shaded area shows mean ± s.e.m. of T_n (log scale). **g**–**i**, Performance dynamics as we focus on all samples (**g**, $n = 231, 231, 229, 232$, from left to right), samples of human-targeted attacks (**h**, $n = 176, 173, 173, 174$, from left to right) and include vague data on fatalities (**i**, $n = 227, 147, 225, 148$, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 2) appear indistinguishable for first failures (two-sided Welch's t-test; $P = 0.400, 0.859, 0.395$), but quickly diverge for second failures (two-sided Welch's t-test; $P = 2.08 \times 10^{-3}, 6.70 \times 10^{-3}, 3.76 \times 10^{-3}$). The successful group also shows a significant improvement in performance (one-sided Welch's t-test; $P = 2.55 \times 10^{-2}, 5.65 \times 10^{-2}, 3.77 \times 10^{-2}$), which is absent for the unsuccessful group (one-sided Welch's t-test; $P = 0.970, 0.901, 0.967$). Data are mean ± s.e.m. **j**–**l**, AUC score of predicting ultimate success as we focus on all samples (**j**), samples of human-targeted attacks (**k**) and include vague data on fatalities (**l**). The centres and error bars of AUC scores denote the mean ± s.e.m calculated from tenfold cross-validation over 50 randomized iterations. **m**–**o**, Temporal scaling patterns as we change the threshold for the successful group to fatal attacks that killed at least 5 (**m**), 10 (**n**) and 100 (**o**) people. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$, NS, $P \geq 0.1$.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Additional robustness checks. a–i, Robustness check as we control for temporal variation. **a–c**, Failure streak in science (**a**), entrepreneurship (**b**) and security (**c**). Blue circles represent real data of successful groups and dashed lines represent fitted Weibull distributions. **d–f**, Temporal scaling patterns in science (**d**), entrepreneurship (**e**) and security (**f**). The shaded area shows mean ± s.e.m. of T_n (log scale). **g–i**, Performance dynamics in science (**g**, $n = 628, 145, 571, 123$, from left to right), entrepreneurship (**h**, $n = 248, 1{,}332, 237, 1{,}312$, from left to right) and security (**i**, $n = 231, 173, 229, 174$, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable for first failures (two-sided weighted Welch's *t*-test; $P = 0.814, 0.728, 0.330$) but quickly diverge for second failures (two-sided weighted Welch's *t*-test; $P = 1.80 \times 10^{-2}$, 3.10×10^{-2}, 4.56×10^{-2}). The successful group also shows significant improvement in performance (one-sided weighted Welch's *t*-test; $P = 2.10 \times 10^{-2}$, 1.92×10^{-2}, 4.53×10^{-2}), which is absent for the unsuccessful group (one-sided weighted Welch's *t*-test; $P = 0.755, 0.175, 0.903$). Data are mean ± s.e.m. **j–l**, Performance dynamics as we compare first and halfway attempts in science (**j**, $n = 628, 145, 582, 111$, from left to right), entrepreneurship (**k**, $n = 248, 1{,}332, 240, 1{,}294$, from left to right) and security (**l**, $n = 231, 173, 228, 175$, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before their last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable for first failures (two-sided Welch's *t*-test; $P = 0.898, 0.671, 0.289$) but diverge for halfway failures (two-sided Welch's *t*-test; $P = 2.18 \times 10^{-5}, 1.34 \times 10^{-2}, 1.34 \times 10^{-2}$). The successful group also shows significant improvement in performance (one-sided Welch's *t*-test; $P = 2.35 \times 10^{-2}$, 4.54×10^{-2}, 3.69×10^{-2}), which is absent for the unsuccessful group (one-sided Welch's *t*-test; $P = 0.992, 0.252, 0.955$). Data are mean ± s.e.m. **m–o**, Performance dynamics as we compare the first and penultimate attempts in science (**m**, $n = 628, 145, 896, 87$, from left to right), entrepreneurship (**n**, $n = 248, 1{,}332, 227, 1{,}199$, from left to right) and security (**o**, $n = 231, 173, 230, 173$, from left to right). The successful and unsuccessful groups that experienced a large number of consecutive failures before the last attempt (at least 5 for D_1, 3 for D_2 and 2 for D_3) appear indistinguishable for first failures (two-sided Welch's *t*-test, $P = 0.898, 0.671, 0.289$) but diverge for penultimate failures (two-sided Welch's *t*-test; $P = 8.50 \times 10^{-8}, 3.12 \times 10^{-2}, 1.13 \times 10^{-2}$). The successful group also shows a significant improvement in performance (one-sided Welch's *t*-test; $P = 5.79 \times 10^{-9}, 4.30 \times 10^{-2}, 1.33 \times 10^{-2}$), which is absent for the unsuccessful group (one-sided Welch's *t*-test; $P = 0.980, 0.138, 0.923$). Data are mean ± s.e.m. **p–r**, The correlation between length of failure streak and initial performance (samples with repeated failures) in science (**p**, $n = 12{,}171$), entrepreneurship (**q**, $n = 2{,}086$) and security (**r**, $n = 441$). Correlation is weak across all three datasets (Pearson correlation; $r = -0.051, -0.011, -0.107$ for **p**, **q**, **r**, respectively). **s–u**, Length of failure streak still follow fat-tailed distributions conditional on bottom 10% initial performance samples in science (**s**, $n = 6{,}339$), entrepreneurship (**t**, $n = 2{,}438$) and security (**u**, $n = 1{,}092$). Two-sided Kolmogorov–Smirnov test between sample and exponential distributions rejects the hypothesis that the two distributions are identical with $P < 0.01$. $^*P < 0.1$, $^{**}P < 0.05$, $^{***}P < 0.01$, NS, $P \geq 0.1$.



Extended Data Fig. 10 | Generalization of the *k* model. a, The α parameter connects the potential to improve $(1-x)$ with the likelihood of creating new versions p through $p=(1-x)^{\alpha}$. **b**, Phase diagram of the k–α model. The two-dimensional parameter space is separated into three regimes, with boundaries at $k\alpha=1$ and $(k-1)\alpha=1$. **c**, The impact of δ parameter on scaling exponent γ for given $k=1, 2, 3$ and $\alpha=0.4, 0.8, 1.2$. We find that δ may affect the temporal scaling parameter when it is small, but has no further effect beyond a certain point $\delta^{*}=\min(\alpha, 1/(k-1))$. **d**, Phase diagram of the k–α–δ model for $k=3$, with boundaries at $\alpha=\delta$, $(k-1)\delta=1$, $(k-1)\delta+\alpha=1$, $k\alpha=1$ and $(k-1)\alpha=1$, respectively.

nature research

Corresponding author(s): Dashun Wang

Last updated by author(s): Sep 2, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	This paper makes use of restricted access data from the National Institutes of Health, protected by the Privacy Act of 1974 as amended (5 U.S.C. 552a). De-identified data necessary to reproduce all plots and statistical analyses will be made freely available. Those wishing to access the raw data may apply for access following the procedures outlined in the NIH Data Access Policy document available at http://report.nih.gov/pdf/DataAccessPolicy.pdf. The VentureXpert database is available via Thomson Reuters. The Global Terrorism Database is publicly available at https://www.start.umd.edu/gtd/.
Data analysis	Data analyses were conducted using Python 3.4. Regression analysis were conducted using Stata 14.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

De-identified data necessary to replicate results of this study will be made freely available.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative study of dynamics of failure based on pre-existing datasets.
Research sample	We collected three large-scale datasets from three domains: (1) R01 grant applications ever submitted to the National Institutes of Health (NIH), (776,721 applications by 139,091 investigators from 1985 to 2015); (2) Start-up investment records from VentureXpert database (58,111 startup companies involving 253,579 innovators) ; and (3) Terrorist attack data from Global Terrorism Database (70,350 terrorist attacks by 3,178 terrorist organizations from 1970 to 2017).
Sampling strategy	No statistical methods were used to predetermine sample size.
Data collection	This study is based on pre-existing datasets.
Timing	The NIH dataset was collected in 2016. The VentureXpert and GTD datasets were collected in 2017.
Data exclusions	The analysis has no data exclusions. Selection criteria within a dataset are described in the supplementary information.
Non-participation	There are no participants in this study.
Randomization	This is a data driven study, not a randomized experiment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging