Innovation under Uncertainty: Perspectives from Reinforcement Learning *

Haoran Wang †

LATEST VERSION HERE

November 13, 2026

Abstract

I incorporate reinforcement learning into a quality-ladder model to study how firms innovate when the distribution of innovation returns across technological fields is unknown. Firms learn value functions from noisy market feedback following successful innovations and use these estimates to determine innovation expenditure and directions of external innovation. Learning makes innovation decisions path dependence conditional on firms' experience: favorable feedback reinforces innovation in existing fields and encourages expansion into nearby technologies, while unfavorable feedback redirects search toward more distant fields. Patent-level evidence from U.S. public firms is consistent with these predictions. Simulations of the calibrated model show that temporary market optimism produces persistent changes in the level and technological composition of patenting. Field-specific optimism redirects innovation toward favored technologies, while broad optimism raises patenting primarily through internal innovation, reducing the share of external innovation.

*I am most grateful to Luminita Stevens for her continuous advising and invaluable support to this project. I thank Boragan Aruoba, John Haltiwanger, Ham Huang, Tyler Pike and William Witheridge for helpful comments.

†Department of Economics, University of Maryland, College Park. Email: hw2688@umd.edu

1 Introduction

The modern growth literature allows innovation outcomes to be stochastic but generally assumes that firms know the distribution of returns to innovative activity. However, in reality, firms may be uncertain about this distribution and therefore unable to calculate the value of future innovation opportunities. Market feedback following successful innovations provides information that helps firms revise these valuations and adjust subsequent research decisions. A striking recent example is the reception of ChatGPT. Before its public release in late 2022, the product was not widely viewed within OpenAI as a transformative innovation, and its developers reportedly expected only modest adoption.[1] The extraordinary response from users following its release was far beyond the expectations of the developers. The company subsequently redirected its R&D resources towards GPT-based products from 2023 onwards.[2]

The motivating example brings up an important question: How do uncertainty and learning shape firms' allocation of R&D resources and direction of innovation? To answer this question, I introduce reinforcement learning into an endogenous-growth framework. Firms do not know the distribution of innovation returns across technological fields and therefore cannot calculate field values as in a full-information benchmark. Instead, they learn these values directly from market feedback and use their estimates to determine research expenditure and the direction of innovation. Their decisions in turn determine where and how frequently they receive subsequent feedback to update their future estimates. The reinforcement-learning model predicts an *history-dependent* path of endogenous innovation, which is consistent with the empirical evidence I document from the patenting data of U.S. public firms. The reinforcement-learning model also makes it possible for non-fundamental factors, such as sentiments and narratives, to affect the direction of technological growth. I explore this possibility using a calibrated model.

The paper starts by proposing a reinforcement-learning framework for endogenous innovation. When firms do not know the distribution of innovation returns across technological space, they cannot calculate how current research expenditure translates into expected future profits and hence cannot determine the value of innovating in each field. This problem has

[1] For instance, in a podcast interview, Jiayi Weng, a key developer in the ChatGPT team, said "We developed the chatbot to elicit some real-world user data... we expect it to have 10k or 20k active users and close it if it drops to 0... but the market response was totally beyond expectation: the user base grew exponentially...". For details, see https://music.youtube.com/podcast/I0DrcsDf3Os.

[2] There was a massive media coverage on this. For example, https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/.

the structure of a dynamic programming problem in which the decision maker has imperfect knowledge of the underlying Markov decision process. Reinforcement learning (RL) provides a framework for learning values and policies through experience. Under RL, decision makers maintain estimates of the value functions. In each period, they update their estimates of value functions and decision rules in two steps. In the "policy evaluation" step, decision-makers update their estimates of value functions based on the delayed rewards or feedback from the actions they took in previous periods. Then, in the "policy improvement" step, they choose actions based on the updated estimates, which would generate rewards in the future. Through this repeated interaction between value updating and policy updating, decision makers progressively refine their assessments of future payoffs and their decision rules. This process can improve their choices even though their value estimates remain imperfect and their policies fall short of the optimum.[3]

I incorporate this mechanism into a quality-ladder model in which firms improve products in their existing technological fields through internal innovation and expand into other fields through external innovation. Firms hold estimates of field values, which summarize the expected profitability of innovating in each field. Following a successful innovation, the firm uses the market response as a noisy assessment of field value to update its value estimate in the policy-evaluation step. Gaussian-process beliefs allow firms to generalize this feedback across technologically related fields, including those they have not yet entered. In the policy-improvement step, firms use their updated valuations to determine internal research expenditure through a decision rule motivated by the full-information benchmark. They direct external innovation toward fields with the highest estimated values and choose external expenditure accordingly. These innovation choices would in turn determine where and how frequently firms generate experience and receive subsequent feedback.

The model generates rich predictions on the innovation patterns of the firm. First, reinforcement learning makes firms' choice of innovation history-dependent. As a result, firms with same technological portfolio and product qualities can choose different research expenditures and directions because they have experienced different histories of market feedback. Favorable feedback raises the perceived value of further innovation in a field and encourages additional research, whereas disappointing feedback lowers its perceived value and discourages further investment. In the canonical full-information quality-ladder benchmark, innovation decisions depend on firms' current technological states and known innovation returns;

[3]Successful applications of the RL algorithm, such as AlphaGo, follow the same broad logic. AlphaGo uses feedback from previous games to update its evaluation of different positions and improve its subsequent decisions. Through repeated experience, it learns useful approximations to the value and policy functions that allow it to play effectively (Silver et al., 2016).

past market assessments have no independent role once these determinants are held fixed. Reinforcement learning therefore introduces an additional source of heterogeneity in innovation decisions, allowing firms with the same technological capabilities to allocate research differently depending on what they have learned from past innovations.

Additionally, the reinforcement learning model generates a distinctive geometric pattern of external innovation in the technological space. Through Gaussian-process generalization, favorable feedback in an existing field can raise the firm's perceived values of nearby external fields, encouraging expansion into the surrounding technological neighborhood. As a result, firms with more favorable signal histories tend to have a more concentrated path of innovation. Disappointing feedback can instead lower these valuations and redirect search toward more distant fields, resulting in a more dispersed path of innovation. Even when the true value is identical across fields, differences in learned valuations can direct external innovation toward particular technological regions. The mechanism also differs from knowledge-based path dependence in directed-innovation models (e.g. Acemoglu (2002)), where accumulated technological knowledge raises the productivity of subsequent research regardless of the rewards from past innovation results.

To examine these predictions, I test the predictions using the patenting history of U.S. public firms. I combine the universe of patents filed or granted by U.S. firms from USPTO with the patent-firm linkage in Kogan, Papanikolaou, Seru, and Stoffman (2017), which also contains a measure of patent value constructed from abnormal stock price movement in a three-day window around the grant date of each patent by U.S. public firms. I use the component of this patent value that is not explained by its citations, the patenting firm, or contemporaneous field conditions as a proxy for the market signal. I define the technological space as CPC subclasses from USPTO, and use the backward citation similarity between CPC subclasses to construct a proxy of distance between fields, following the classical approach by Jaffe (1986) and Bloom, Schankerman, and Van Reenen (2013).

I document the following facts from the combined dataset. First, within a firm and year, higher past signals predict greater subsequent innovation in the corresponding existing fields, after controlling for the firm size and knowledge stock in the field. A one-unit increase in average realized signals increases the probability of subsequent innovation in the field by 12.5%. Second, new-field entry is systematically directed toward technological regions receiving stronger signals: firms are more likely to enter a candidate external field that is closer to the fields that have accumulated high realizations of signals and are technologically similar to the candidate field. Within a firm-year, a candidate field at the 90th percentile

of the similarity-weighted signal is 7.7% more likely to be entered than one at the 10th percentile. Firms move away from fields that are technologically similar to those associated with unfavorable feedback. Lastly, the distance between a firm's new field and its existing field in the technological space is negatively correlated with the signal realizations of the existing field. A one-within-firm-year-standard-deviation increase in the signal implies a reduction in technological distance from a new field by approximately 2 percentile points closer. These empirical patterns are consistent with the predictions of the reinforcement learning model.

The past four decades of U.S. stock market witnessed ebbs and flows of sentiments in various technological fields, from the revolution of Information and Communication Technologies (ICT) in 1990s to the recent AI boom. The reinforcement learning model makes it possible for market sentiments of this kind to shape the path of innovation. In the last part of the paper, I calibrate the reinforcement-learning growth model and use it to study the response of firms' innovation behaviors to a wave of optimism by 2 standard deviations of signal noise for a 10-year episode. An optimistic signal shift applied for ten years to a connected block of technologies raises the block's patent share from 7.76 percent in the paired benchmark to 31.29 percent at the end of the episode. The reallocation persists after sentiment disappears: the difference remains 6.77 percentage points ten years later and 1.08 percentage points after twenty years. The strong and persistent effects of sentiment shock come from a novel feedback loop between patents and learning: as firms in the shocked fields file patents in the sentiment episode, the optimism wave shifts the signal realization upwards, which makes firms increase their evaluations of patenting values in these fields. They allocate more resources to this field, which results in a higher probability of success and generates more favorable signals in the field, further reinforcing their beliefs.

When the same optimistic shift applies to all fields, cumulative patenting rises by 73.1 percent during the episode relative to the no-sentiment benchmark, and the effect remains persistent. Interestingly, the boom is driven primarily by internal innovation, which rises by 164.5 percent cumulatively during the episode, compared with 26.0 percent for external innovation. The share of external patents keeps declining during the sentiment episode. This is because the feedback loop between patent and learning is weaker for external innovation: firms update their beliefs about the external fields by extrapolating from the higher beliefs in their internal fields, which is not repeatedly reinforced by the signals generated from that external field. The learning model offers one possible interpretation of three salient patterns in the U.S. economy in the past four decades: persistently booming U.S. stock market return, persistently increasing number of patents in aggregate, and persistently declining share of

external patents in the economy.

Related Literature First, this paper contributes to the literature that uses reinforcement learning to model economic decisions based on value and policy functions learned from experience. The RL method has been applied to model learning and choices in psychology (e.g. Gershman and Daw (2017), Wu et al. (2018),Plate et al. (2023)) and recently in behavioral finance (Barberis and Jin, 2023). Recent macroeconomic applications include Woodford (2019), which combines finite-horizon planning with learned continuation values to study monetary policy under limited foresight. Ilut and Valchev (2024) and Kaplowitz (2026) apply reinforcement learning to study the role of experience in household consumption behavior. Yang et al. (2026) treat reinforcement learning as a computational method to solve heterogeneous-agent macro models. The generalization of experience across related alternatives also connects my paper to Hodgson and Lewis (2025), who use Gaussian-process beliefs to model how experience with previously searched products informs valuations of nearby alternatives, generating path dependence in consumer search. My paper adopts reinforcement learning to study how firms choose the expenditure and direction of innovation when they are uncertain about the distribution of innovation returns across different technologies.

Second, this paper builds on quality-ladder models of innovation and creative destruction and their extensions to firm dynamics. The quality-ladder literature (Grossman and Helpman (1991), Aghion and Howitt (1992), Klette and Kortum (2004), Lentz and Mortensen (2008), Akcigit and Kerr (2018), Berlingieri et al. (2026)) shows that firms of different sizes choose different strategies of innovation. With uncertainty and learning integrated to the quality-ladder framework, my model allows innovation to depend on firms' history of patenting experiences, which proposes a new source of heterogeneity and allows innovation intensity to differ for firms that are of the same size but experienced different histories of signal realizations.

Third, this paper contributes to the study of direction and path of innovation. The quality ladder models typically assume that creative destruction is undirected. The directed technical change literature (Acemoglu (2002), Acemoglu et al. (2012)) discusses the conditions under which innovation is directed towards binary features of technologies (skill-based vs unskill-biased, clean vs dirty, ...). In this class of model, path dependence of innovation is generated by assuming the accumulated knowledge base increases the productivity of future innovation in a field. Similarly, Redding (2002) proposes that technology-specific secondary knowledge transfers imperfectly to the next fundamental technology, so accumulating knowl-

edge around the incumbent technology reduces the return to searching for a new technological trajectory. My model proposes an alternative channel of path dependence through learning. In my model, path dependence is conditional upon the history of accumulated signals: unfavorable market feedback can discourage further innovation even in fields where firms have accumulated substantial technological knowledge.

Lastly, this paper relates to research on how uncertainty and learning shape innovation and entrepreneurial decisions. Akcigit and Liu (2016) study how private information about failed research affects firms' experimentation and creates inefficiencies in research effort. Lei and Shi (2024) show that Knightian uncertainty discourages entrepreneurial entry, particularly among less wealthy households. Jo and Kim (2025) study delays in learning other firms' frontier technologies, which induce incumbents to use own-innovation strategically to deter creative destruction. My paper studies a complementary mechanism through which firms learn field values from market feedback on their own innovations and generalize this experience across technologically related fields.

Layout The rest of the paper proceeds as follows. Section 2 presents a simplified, illustrative model of innovation with reinforcement learning and sketches the key predictions of the model. Section 3 presents the empirical analysis that tests the model predictions. Section 4 specifies the details of the full quantitative model and calibration, and studies the effect of sentiment shock to the growth path. Section 5 concludes.

2 A Reinforcement-learning Framework of Innovation

2.1 The Model

We begin with a simplified endogenous innovation problem of a multi-product firm to illustrate the learning mechanism. Let $x \in X$ be a vector that characterizes a technological field dominated by a firm. Let $q(x)$ be the quality of goods that the firm produces in field x. Following Akcigit and Kerr (2018), assume that the monopolistic profit of the firm in field x is equal to $\bar{\pi} \cdot q(x)$ where $\bar{\pi}$ is a constant, and the expenditure of innovation in field x is equal to $c(x) \cdot q(x)$, where $c(x)$ is the innovation intensity of the firm in field x. The innovation

intensity c translates into a Poisson arrival rate of successful innovations z by

$$z(c) = \sqrt{2c}.$$

A successful innovation (e.g. the firm files a new patent in the field) improves the quality of goods from q to q' in field x. For now, I abstract from the business-stealing features in Schumpeterian growth models: the firm can generate profits from the field forever once it dominates it.

Given interest rate r and suppose time is continuous, the firm chooses the innovation intensity c on field x that solves the Bellman equation

$$rV(q,x) = \max_c \bar{\pi}q - cq + z(c) \cdot \mathbb{E}[V(q',x) - V(q,x)] \tag{1}$$

where $\mathbb{E}[V(q',x) - V(q,x)]$ is the expected return of a successful innovation in field x.

Let $\lambda(x) \equiv \frac{q'(x)-q(x)}{q(x)}$ be the realized step size of quality improvement in a successful innovation, which I will refer to as the "return" or "prize" of the innovation. In the endogenous growth literature, it is typically assumed that the distribution of $\lambda(x)$ is known to the firm (e.g. Akcigit and Kerr (2018); Berlingieri et al. (2026)). In that case, this problem is easy to solve: $V(q,x)$ takes the form $V(q,x) = v(x) \cdot q$, where $v(x)$ solves

$$rv(x) = \max_c \bar{\pi} - c + \sqrt{2c}\mathbb{E}[\lambda(x)]v(x) \tag{2}$$

Since the probability distribution of $\lambda(x)$ is known, the firm is able to compute $\mathbb{E}[\lambda(x)]$ and solve the value function coefficient $v(x)$ exactly.

Apart from spending on improving the quality of field x that the firm already knows about, the firm can also innovate externally on a field in which it has not yet filed a patent. Suppose that the firm does not know the quality of such external fields *ex ante*. Following Akcigit and Kerr (2018), assume that the external innovation expenditure scales with average quality $\bar{q} \equiv \int_0^1 q_j \, dj$. The firm chooses external innovation field x' and intensity e by solving

$$\max_{e,x'} -e\bar{q} + z(e)\mathbb{E}[V(q,x')] \tag{3}$$

Under full information, the firm knows that $V(q,x')$ takes the functional form $V(q,x') =$

$v(x')q$, and hence external choices (3) degenerate to

$$v^e(x') \equiv \max_{e,x'} -e + z(e)v(x') \tag{4}$$

If the probability distribution of $\lambda(x)$ is different across fields, the firm would choose to spend on the external field with the highest $v(x')$. If the probability distribution of $\lambda(x)$ is identical across field x, then $v(x)$ should also be identical, and the search of external fields becomes undirected.

Let $\mathbf{X} \equiv \{x_1, \cdots, x_n\}$ denotes the fields in which the firm has filed patents, and let $\mathbf{q} \equiv \{q_1, \cdots, q_n\}$ denotes the corresponding quality of each field goods. Aggregating the field-level value functions (1) and the value of external innovation (3) gives us the firm's value function

$$\begin{aligned} rV(\mathbf{q}, \mathbf{X}) = \max_{c(x),e,x'} & \sum_{x \in \mathbf{X}} \bar{\pi} q(x) - c(x)q(x) - e\bar{q} \\ & + \sum_{x \in \mathbf{X}} z(c(x)) \mathbb{E}[V(\mathbf{q} \backslash \{q(x)\} \cup \{\lambda(x)q(x)\}, \mathbf{X}) - V(\mathbf{q}, \mathbf{X})] \\ & + z(e) \mathbb{E}[V(\mathbf{q} \cup \{q(x')\}, \mathbf{X} \cup \{x'\}) - V(\mathbf{q}, \mathbf{X})] \end{aligned} \tag{5}$$

which has a closed-form solution $V(\mathbf{q}, \mathbf{X}) = \sum_{x \in \mathbf{X}} v(x)q(x) + v^e(x')\bar{q}$.

An Algorithm that Computes $v(x)$ Under (2), the optimal innovation intensity in the field x is given by $c^*(x) = \frac{1}{2}\mathbb{E}[\lambda(x)]^2 v(x)^2$, and the value function coefficient is $v^*(x) = \frac{r - \sqrt{r^2 - 2\bar{\pi}\mathbb{E}[\lambda(x)]^2}}{\mathbb{E}[\lambda(x)]^2}$. Plugging them back into (2) gives us a clean mapping from the true value $v(x)$ to optimal innovation intensity $c(x)$:

$$c^*(x) = rv^*(x) - \bar{\pi} \tag{6}$$

This relationship has a simple economic interpretation. Without further innovation, a unit of product quality generates a discounted profit stream worth $\bar{\pi}/r$. The difference $v^*(x) - \bar{\pi}/r$ therefore measures the value from future innovation efforts, net of research costs. Optimal innovation intensity therefore equals this additional value expressed as an annual flow.

Even though it is straightforward to compute the value and policy functions that solve (2), in order to motivate our discussions on reinforcement learning, I introduce an algorithm based on equation (6) that helps the firm figure out $v^*(x)$ and $c^*(x)$ from some initial guess $v^{(0)}(x)$:

1. Start from some initial guess of $v^*(x)$, denoted as $v^{(0)}(x)$;

2. Given $v^{(k)}(x)$, choose innovation intensity according to

$$c^{(k+1)}(x) = rv^{(k)}(x) - \bar{\pi} \tag{7}$$

3. Update the value function by

$$rv^{(k+1)}(x) = \bar{\pi} - c^{(k+1)}(x) + z(c^{(k+1)}(x))E[\lambda(x)]v^{(k)}(x) \tag{8}$$

4. Iterate until $c^{(k)}(x)$ or $v^{(k)}(x)$ converges.

The following proposition shows the convergence of the value-learning algorithm to the true value function coefficient $v^*(x)$ and policy function $c^*(x)$ in (2).

Proposition 1. *Assume* $r^2 \geq 2\mathbb{E}[\lambda(x)]^2\bar{\pi}$*, and suppose the initial guess* $v^{(0)}(x) \in (\frac{\bar{\pi}}{r}, \frac{r-\sqrt{r^2-2\bar{\pi}\mathbb{E}[\lambda(x)]^2}}{\mathbb{E}[\lambda(x)]^2}]$. *The algorithm (7) and (8) satisfies* $v^{(k)}(x) \leq v^{(k+1)}(x)$ *and*

$$\lim_{k\to\infty} v^{(k)}(x) = v(x), \quad \lim_{k\to\infty} c^{(k)}(x) = c(x).$$

Proof. See Appendix A.1. □

The proposition suggests that under full information, the algorithm (7) and (8) can lead the firm to gradually learn the value and policy functions. It is analogous to policy-iteration-style algorithms that macroeconomists are familiar with. Equation (7) can be regarded as a "policy improvement" step: it defines how the existing value function estimate is mapped to the choice of innovation expenditure in the learning process, and the proposition shows that following this decision rule necessarily leads to policy improvement, in the sense that $c^{(k+1)}$ would give the firm a weakly higher payoff than $c^{(k)}$. In other words, the firm keeps improving its decisions over time. Equation (8) is akin to a "policy evaluation" step: based on the updated decision of innovation expenditure, the firm updates its evaluation of field x from its previous estimate. The right hand side of (8) gives the firm a correct evaluation of field x under innovation expenditure $c^{(k+1)}$. These successive adjustments bring both the valuation and innovation expenditure toward their full-information optima.

Uncertainty and Reinforcement Learning The standard endogenous growth literature assumes that firms know the distribution of innovation returns, even though they do not

know the realized outcome *ex ante*. In practice, firms may also be uncertain about the distribution itself. This information friction can be captured in my model by assuming that a firm does not know the distribution of quality improvements $\lambda(x)$ across technological fields, including its mean $\mathbb{E}[\lambda(x)]$ as well as how $\lambda(x)$ correlates with returns in other fields $\lambda(x')$. The firm observes its current quality and knows its profit flow, research costs, and the expenditure–arrival mapping $z(c)$, but this knowledge is not sufficient for the firm to determine the expected gains from successful innovation.

Reinforcement learning provides a framework for studying how decision makers learn to make choices when the true value function cannot be correctly computed and has to be learned from experiences (Sutton and Barto, 2018). A reinforcement-learning agent cannot and need not begin with a complete description of the environment from which it can calculate all relevant values. Instead, he maintains estimates of the values based on the states he has experienced and the actions he has taken, uses these estimates to guide its current decisions, and updates the value estimates using the rewards of these decisions in the future.

As Moll (2024) points out, policy iteration algorithms in dynamic programming, such as the one defined by (7) and (8), can capture this "learning-by-doing" spirit of reinforcement well: the "policy improvement" step (7) helps the firm figure out a decision that yields a better payoff, and the "policy evaluation" step (8) describes how the firm update its evaluation from what he chose. A reinforcement-learning firm would also follow these two steps, with two modifications. First, since the expectation $\mathbb{E}[\lambda(x)]$ is not known to the firm, the value update (8) can no longer use $\mathbb{E}[\lambda(x)]$ on the right hand side, and the firm can no longer correctly evaluate the value of the field at a given expenditure level. It would instead maintain an imperfect estimate $\hat{v}_t(x)$ and update it with an alternative policy evaluation equation to (8). Second, because the correct evaluation of previous period's decision $v^{(k)}$ is no longer available, when doing the "policy-improvement" step (7), the firm has to use the imperfect estimate $\hat{v}_t(x)$ to guide its expenditure decision c_t.

Market Feedback as a Value Assessment Now I introduce how to replace the "policy-evaluation" step (8) with an alternative learning rule that does not require the formation of expectation on $\lambda(x)$. Let $\hat{V}_{t-1}(q,x)$ denote the firm's evaluation of the field-level value $V(q,x)$ before learning event t. Here, t indexes the number of patenting events the firm has experienced. Suppose now the firm successfully files the t-th patent in field x from its past innovation efforts in this field, and the new patent generates a new reward $R_t(q,x)$ about

the value of the firm innovating in field x. The firm revises its estimate according to

$$\hat{V}_t(q,x) = \hat{V}_{t-1}(q,x) + \alpha_t \left[R_t(q,x) - \hat{V}_{t-1}(q,x) \right] \tag{9}$$

where the learning rate $\alpha_t \in [0,1]$ determines the weight placed on the new reward. The term in brackets is a reward prediction error, which captures the surprises that the new reward R_t brings to the current estimate $\hat{V}_{t-1}$ about the field x. In standard dynamic programming, this R_t is simply the right-hand side of the Bellman equation. In reinforcement learning, we have more flexibility in choosing this reward object R_t. For example, Monte Carlo methods use realized discounted returns as R_t, while temporal-difference methods combine an observed payoff with an estimated continuation value (Sutton and Barto, 2018). Two recent applications in macroeconomics also takes a similar form to equation (9): Woodford (2019) uses an estimated value from finite-horizon planning as R_t, and Yang et al. (2026) uses an estimated value from Monte-Carlo simulation of price trajectories. In my application to innovation, I use market feedback following successful innovations as R_t, which I will specify in detail shortly.

Since the firm knows that the profit and innovation expenditure scales with quality q, in its estimate $\hat{V}_t(q,x)$ it preserves the linear structure on q in the true $V(q,x)$. Formally, the firm's estimate $\hat{V}(q,x)$ takes the form $\hat{V}_t(q,x) = q\hat{v}_t(x)$, where $\hat{v}_t(x)$ is a subjective estimate of the value function coefficient $v(x)$ defined in (2). Dividing on both sides of (9) by q gives us an equivalent value updating equation on the value function coefficient $v(x)$:

$$\hat{v}_t(x) = \hat{v}_{t-1}(x) + \alpha_t \left[r_t(q,x) - \hat{v}_{t-1}(x) \right]. \tag{10}$$

where $r_t(q,x) \equiv R_t(q,x)/q$ denote the per-quality reward in period t.

Now I specify the reward R_t in my model. Suppose the new patent in field x increases the quality of goods in field x by an observed amount $\Delta q_t(x) \equiv q_t - q_{t-1}$. Under full information, this patenting event will lead the firm value to change by $\Delta V_t(x) \equiv V(q_t,x) - V(q_{t-1},x) = \Delta q_t \cdot v(x)$. In an efficient financial market, the market value of the firm is correctly priced at $V(\mathbf{q},\mathbf{X})$, and the stock market response to this patenting event should be exactly equal to $\Delta q_t \cdot v(x)$. However, since the distribution of innovation returns across technological fields is not known to the investors either, the variations in stock market values are not equal to $\Delta q_t \cdot v(x)$ exactly. Instead, I assume that the observed stock market response ξ_t is a noisy signal of this value gain:

$$\xi_t = \Delta q_t \cdot v^*(x) + \varepsilon_t \tag{11}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is independent of the unknown value of the firm. The firm can then use the observed quality increment to back out a reward r_t by:

$$r_t(x) \equiv \frac{\xi_t}{\Delta q_t} = v^*(x) + e_t, \qquad \text{where } \sigma_{r,t}^2 \equiv \text{Var}(e_t \mid \Delta q_t) = \frac{\sigma_\varepsilon^2}{(\Delta q_t)^2}. \tag{12}$$

The firm then uses the r_t in the updating rule of the value coefficient (10). This market-value-based reward provides the firm with a delayed feedback of its previous efforts in the innovation of field x, and the firm's estimate of value $\hat{v}_t(x)$ is therefore dependent upon the history of this market response.

From Value Estimates to Innovation Decisions I now specify how the firm use its estimated value $\hat{v}_t(x)$ to guide its innovation decisions. The full-information policy-improvement step (7) provides a natural benchmark. Its policy-update step, combined with the full-information value update, generates improving innovation policies and converges to the optimal expenditure. I assume that it retains the mapping given by this policy-improvement step (7), substituting its estimated value $\hat{v}_t(x)$ for the value iterate in the benchmark algorithm: in each patenting event t,

$$c_t(x) = \max\{r\hat{v}_t(x) - \bar{\pi}, 0\} \tag{13}$$

The economic interpretation of this decision rule is similar to (7). Maintaining the current product without further innovation yields a value of $\bar{\pi}/r$ per unit of quality. The difference $\hat{v}_t(x) - \bar{\pi}/r$ therefore measures the firm's estimated additional value from future innovation, net of research costs. The firm translates this perceived innovation premium into expenditure, subject to the nonnegativity constraint.

This decision rule makes innovation expenditure experience-based. A market reward exceeding the firm's previous valuation raises its estimate and increases subsequent expenditure. In turn, higher innovation efforts in a field increases the frequency of successful innovations, allowing the firm to generate more market feedback to the innovations in this field. If the feedback is again positive, then the firm will further reinforce its high estimate of the field's value and keep spending on the field. A negative market reward has an opposite effect: if a firm receive a negative market feedback in the field, the evaluation $\hat{v}_t(x)$ is revised downwards, and the innovation expenditure will be lowered. This would in turn lower the probability of drawing further market signals from successful patents in this field, making the evaluation $\hat{v}_t(x)$ and expenditure $c_t(x)$ stuck at the low level.

Using estimated values $\hat{v}_t(x)$ in place of $v^{(k)}$ in the policy-improvement step is a standard practice in reinforcement learning.[4] With exact current-policy values, this adjustment weakly improves performance relative to policies satisfying the same exploration constraint. If the policy improvement step uses estimated values instead, the gap between estimated current-policy values and the true current-policy values can lead the firm to make temporarily worse decisions. If the firm can accumulate long enough experience, so that the estimate $\hat{v}_t(x)$ converges to the true value function coefficient $v^*(x)$ in the long run, then the innovation effort would also converge to the full-information optimum $c^*(x)$ eventually. Note that this is feasible only in our current setup where the firm will never lose the field to another firm due to creative destruction. In a more general setup in Section 4 where creative destruction is allowed, the firm can only accumulate a limited episode of experiences, and $\hat{v}_t(x)$ and $c_t(x)$ do not converge to v^* and c^* during the life cycle of each firm-field pair.

Gaussian-Process Generalization across Fields So far, I have only considered the cases where the firm learns the value of a field x from the market signals in that field only. In reality, market signals in another field x' may also be informative to the value of innovating in field x if the two fields are technologically similar. Additionally, since the firm also needs to decide the intensity and direction of innovation in the fields that it has not yet filed patents, it might find it appealing to generalize the knowledge accumulated from the fields he already owns patents in and inform the external innovation decisions.

I capture this generalization by assuming that the firm's subjective evaluation $\hat{v}_t(x)$ operates on a Gaussian Process Regression. Specifically, the firm holds a Gaussian Process prior about the true value function coefficient $v^*(x)$ over the technological field X, i.e. for any finite collections of fields $x_1, x_2, \cdots, x_N$, the distribution of values in these fields $v^*(x_1), v^*(x_2), \cdots, v^*(x_N)$ follows a joint Gaussian distribution

$$\begin{bmatrix} v^*(x_1) \\ \vdots \\ v^*(x_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{v}_0(x_1) \\ \vdots \\ \hat{v}_0(x_N) \end{bmatrix}, \begin{bmatrix} k_0(x_1,x_1) & \cdots & k_0(x_1,x_N) \\ \vdots & \ddots & \vdots \\ k_0(x_N,x_1) & \cdots & k_0(x_N,x_N) \end{bmatrix} \right), \tag{14}$$

where the mean function $\hat{v}_0(x)$ is the initial guess (or the prior mean) of the firm about the value of innovation in field x, and the kernel function $k_0(x, x')$ gives us the covariance between the values in field x and x'. Here I assume that $k_0(x, x')$ is the standard radial-basis

[4] Examples include ϵ-greedy control method (Sutton and Barto (2018), Chapter 5), which assigns most probability to actions with the highest estimated values while retaining some exploration, and softmax Barberis and Jin (2023), which is multinomial logit decision rule based on estimated values.

functions, as in Rasmussen and Williams (2006):

$$k_0(x, x') = \sigma_0^2 \exp\left(-\frac{\|x - x'\|^2}{2\varphi}\right). \tag{15}$$

This covariance function declines as the distance between field x and x' becomes larger. In other words, the evaluations of technological similar fields are more strongly correlated in the prior. The degree of generalization is captured by the parameter $\varphi > 0$: a larger φ means that it is easier for the firm to generalize its evaluations from experienced fields to external fields. σ_0^2 denote the prior uncertainty of the firm in any given field.

This Gaussian Process prior has important implications for the value updating equation (10). First, the Gaussian Process implies that the signal ξ_t in (11) is Gaussian, and hence the updating equation (10) becomes a Bayesian updating equation, with $\hat{v}_t(x)$ being the posterior mean and α_t being the Kalman gain. Moreover, the Gaussian Process provides a convenient way for the firm to update its estimates on an external field x':

$$\hat{v}_t(x') = \hat{v}_{t-1}(x') + \alpha_t(x', x)\left[r_t(x) - \hat{v}_{t-1}(x)\right] \tag{16}$$

where $\alpha_t(x', x) \equiv \frac{k_{t-1}(x,x')}{k_{t-1}(x,x)+\sigma_{r,t}^2}$ is a learning rate that captures how much information the signals in field x carries about the value of field x', with kernel $k_{t-1}(x, x')$ defined recursively as

$$k_t(x'', x') = k_{t-1}(x'', x') - \frac{k_{t-1}(x'', x)k_{t-1}(x, x')}{k_{t-1}(x, x) + \sigma_{r,t}^2}. \tag{17}$$

If field x and x' are distant in technological space, then the learning rate $\alpha_t(x', x)$ is small and the value estimate $\hat{v}_t(x')$ updates less from the previous estimates.

External Innovation Choice under RL Lastly, I specify how the firm chooses external innovation intensity and direction using $\hat{v}_t(x)$. Replacing the true value $v(x')$ on the right hand side of equation (3) with the estimated value $\hat{v}_t(x')$, the external choices are given by

$$\max_{e,x'} -e + z(e)\hat{v}_t(x') \tag{18}$$

and therefore the external innovation direction x_t and intensity e_t is given by

$$x_t \in \arg\max_{x' \notin \mathbf{X}} \hat{v}_t(x'), \qquad e_t = \frac{1}{2}\left[\hat{v}_t(x_t)\right]_+^2.$$

2.2 Implications for Innovation Paths

To illustrate how experience-based valuation and generalization shape innovation paths, I simulate a firm over a time horizon of $T = 80$ on a 15×15 grid of technological fields. The true value coefficient is $v^*(x) = 1$ in every field, but the learning firm does not know that this mapping is flat. Its Gaussian-process prior has constant mean $\hat{v}_0(x) = 1$, standard deviation $\sigma_0 = 0.5$, and covariance parameter $\varphi = 0.25$, unless otherwise specified. The standard deviation of the raw market-signal noise is $\sigma_\varepsilon = 0.15$. For illustrative purposes, the figures display the first 25 successful external entries from each simulation.

Figure 1 compares a full-information firm with a reinforcement-learning firm, both receiving favorable market assessments following successful innovations. We can see that the reinforcement-learning firm has a much more concentrated path in the technological space than the full-information benchmark. Under full information, the firm knows that payoffs are identical across fields. Therefore, the external search path is dispersed and undirected, spreading over the whole space. In contrast, the reinforcement-learning firm does not know the distribution of innovation returns and instead learns field valuations directly from market feedback following successful innovations. Gaussian-process generalization allows this experience to inform its assessments of nearby technologies. Favorable early experience raises valuations around the initial field, and generalization makes neighboring external fields relatively attractive. Subsequent entry and positive feedback further reinforces the high valuations in this technological region, resulting in a concentrated innovation path. This concentration arises despite homogeneous underlying returns: it reflects the firm's history of market feedback rather than an intrinsic advantage of nearby technologies.

Figure 2 isolates the role of signal histories by comparing two learning firms with the same parameters and initial location. In the simulation, every successful patent, including the initial patent, generates the normalized assessment $r_t = 1.3$ in the left panel and $r_t = 0.7$ in the right panel. We can see that the favorable-history external path is more concentrated than the unfavorable-history path. The reason is that the unfavorable signal history depresses local valuations, making more distant fields whose assessments remain relatively unaffected more attractive. The resulting path involves larger technological steps and more dispersed entry.

Figure 3 illustrates the role of generalization by comparing $\varphi = 0.01$ with $\varphi = 4$, holding the normalized assessment fixed at $r_t = 1.3$ following every successful patent. In this comparison, we can see that the path with $\varphi = 4$ is more concentrated than the path with $\varphi = 0.01$.

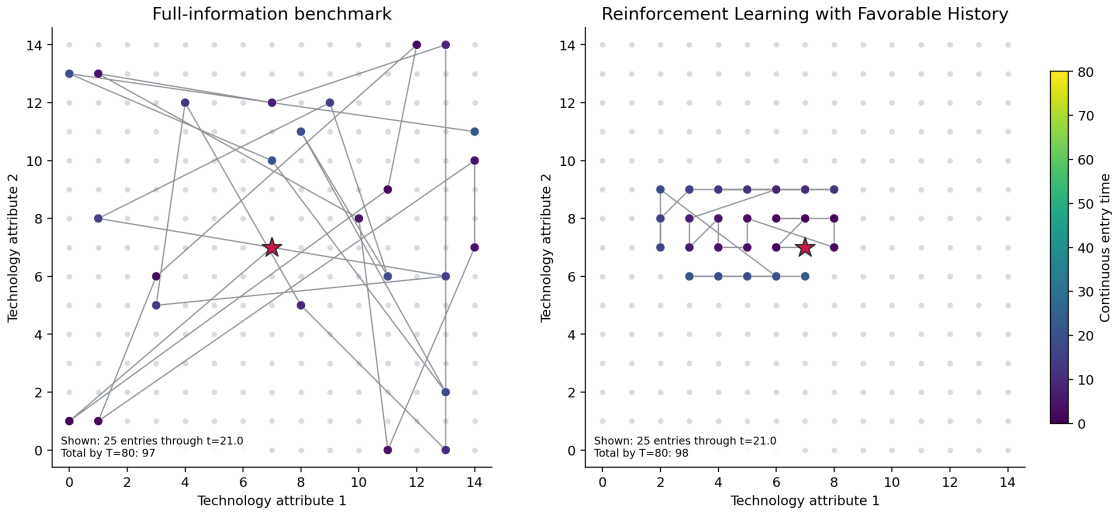


Figure 1: External innovation paths under full information (left) and reinforcement learning with Gaussian-process generalization (right, $\varphi = 0.25$). The true value coefficient is $v^*(x) = 1$ in every field. The full-information firm breaks ties among unowned fields randomly; the learning firm receives the normalized assessment $r_t = \xi_t/\Delta q_t = 1.3$ after every successful patent, including the initial patent. Each panel shows the first 25 successful external entries on a 15×15 grid from a simulation with horizon $T = 80$ and the same seed. The red star marks the initial field, lines connect successive entries, and colors indicate continuous entry time; panel annotations report the displayed time span and total entries by $T = 80$. Parameters are $r = 0.05$, $\bar{\pi} = 0.04$, $\mathbb{E}[\lambda] = \sqrt{0.02}$, $\hat{v}_0 = 1$, $\sigma_0 = 0.5$, and $\sigma_\varepsilon = 0.15$, with initial and externally acquired quality normalized to one.

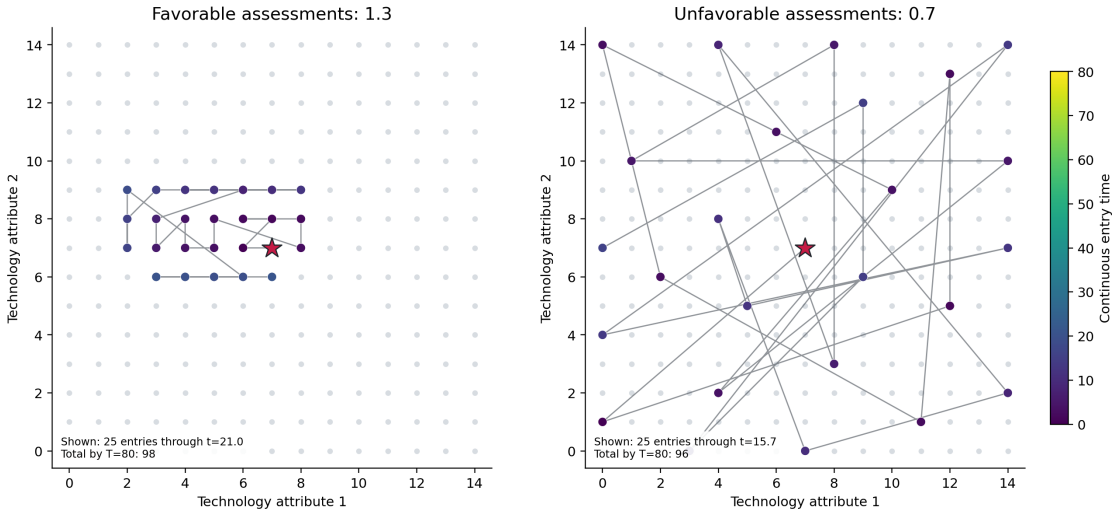


Figure 2: External innovation paths conditional on favorable (left) and unfavorable (right) market assessments. Every successful patent, including the initial patent, generates the normalized assessment $r_t = \xi_t/\Delta q_t = 1.3$ in the left panel and $r_t = 0.7$ in the right panel. These are conditional signal histories, not representative Gaussian draws. Both firms have $\varphi = 0.25$ and the same initial field. Each panel displays the first 25 successful external entries from a simulation with horizon $T = 80$ and the same seed. The red star marks the initial field, lines connect successive entries, and colors indicate continuous entry time. Other parameters are as in Figure 1.

When φ is small, covariance decays rapidly with distance, so favorable market signals update beliefs locally only at or very close to the sampled field. Therefore, the firm's unexplored fields remain nearly symmetric in values, and external entry is undirected and dispersed across the technological space. When φ is larger, the same signals raise posterior means throughout a wider neighborhood. Subsequent external innovation is therefore directed toward the region surrounding the initial field, producing a more concentrated path.

In sum, the decision rules and these illustrations highlight three implications. First, internal innovation expenditure responds to learned valuations, increasing with the perceived value of further innovation. Second, generalization can direct external innovation toward neighborhoods associated with favorable experience and away from those associated with disappointing experience. Third, innovation paths are conditionally history-dependent: firms with the same initial technological portfolio and underlying opportunities can follow different paths after receiving different market feedback. The source of this path dependence is not solely the accumulation of technological knowledge (as in Acemoglu (2002)), but also the history of value assessments and their transmission across fields. These predictions motivate the empirical test in Section 3.

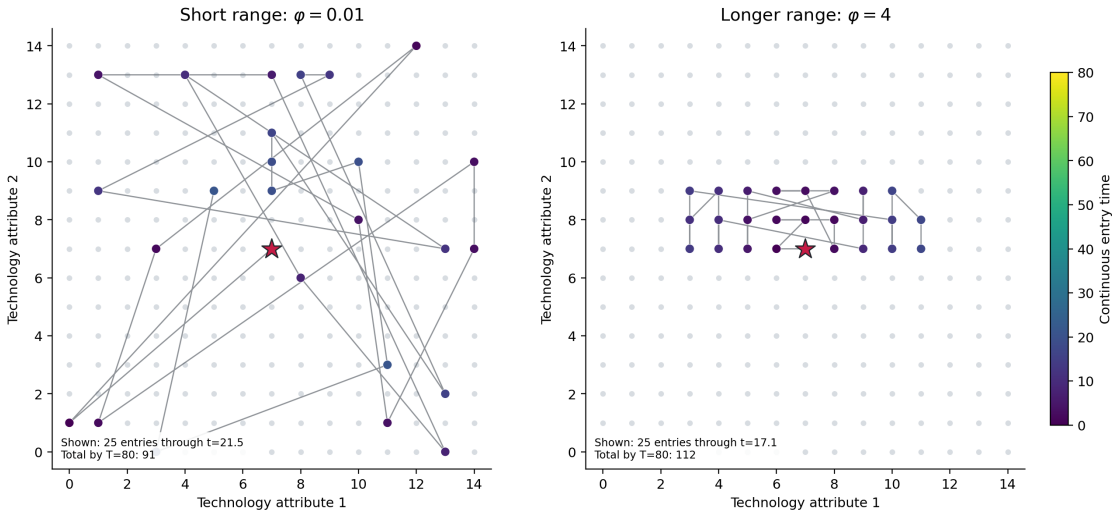


Figure 3: External innovation paths with effectively independent-field learning ($\varphi = 0.01$, left) and spatial generalization ($\varphi = 4$, right), under the kernel $k_0(x, x') = \sigma_0^2 \exp[-\|x - x'\|^2/(2\varphi)]$. Both firms receive the normalized assessment $r_t = \xi_t/\Delta q_t = 1.3$ after every successful patent, including the initial patent. Each panel shows the first 25 successful external entries from a simulation with horizon $T = 80$ and the same seed. The red star marks the initial field, lines connect successive entries, and colors indicate continuous entry time. Other parameters are as in Figure 1. This comparison illustrates spatial generalization, not a globally monotone relationship between φ and concentration.

3 Empirics

In this section, I test the three predictions of reinforcement learning from the illustrative model.

3.1 Data and Measurement

In the empirical analysis, I focus on the patenting behaviors and market responses to successful patents of U.S. public firms from 1981 to 2020. The analysis draws on two data sources. Patent data come from PatentsView, which contains the universe of patents granted by the USPTO since 1976. For each patent, the dataset provides the patent metadata, citation information and CPC classifications. The firm information, such as the public-firm identifiers, comes from the CRSP/Compustat dataset. I merge the patent- and firm-level data using the patent-firm linkage provided by Kogan et al. (2017).

From the illustrative model, we can see that the innovation path of firms over the techno-

logical space is crucial in distinguishing the reinforcement-learning innovation model from other endogenous growth models in the literature. Throughout the empirical analysis, I use CPC subclasses as the empirical counterpart of "fields" X in the illustrative model. The 640 CPC subclasses $X = \{x_1, \cdots, x_{640}\}$ are classified by the USPTO based on the technological features of patents and provide an appropriate level of granularity for our analysis. The key object that links theory to data is a distance measure that captures the technological similarity between two different CPC subclasses. I use patent-level citation information and construct a similarity measure based on the location of CPC subclasses in the backward citation network. Unlike the forward citation network in Acemoglu et al. (2016), which captures how the knowledge from a field spills over to *future* patents in other fields, the backward citation network captures how the development of patents in a field depends upon *past* knowledge from other fields, which speaks more directly to the mechanism in the illustrative model.

Specifically, for each decade $d \in \{1980\text{–}1989, 1990\text{–}1999, 2000\text{–}2009, 2010\text{–}2019\}$ and each CPC subclass x, I first construct the share of backward citations from another field x':

$$s_{x\to x',d} = \frac{\#\text{ of citations from field-}x\text{ patents to field-}x'\text{ patents in decade }d}{\text{Total citations from field-}x\text{ patents in decade }d} \tag{19}$$

It is the share of citations made by field-x patents that point to field-x' patents in decade d, which captures the extent to which innovation in field x depends on prior knowledge in x'. For each field x, we have a vector of backward-citation shares $\mathbf{x} \equiv (s_{x\to x_1}, s_{x\to x_2}, \cdots, s_{x\to x_{640}})$, summarizing the technological dependence of field x on other fields. I then define the technological similarity between fields x and $\tilde{x}$ as the cosine similarity between vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$, as in Jaffe (1986) and Bloom et al. (2013):

$$s(x, \tilde{x}) \equiv \frac{\mathbf{x} \cdot \tilde{\mathbf{x}}}{\|\mathbf{x}\| \cdot \|\tilde{\mathbf{x}}\|} \tag{20}$$

This similarity measure is higher if two fields draw on similar knowledge. Appendix B.2 shows the four decade matrices. The off-diagonal Pearson correlations between successive decades range from 0.748 to 0.787.

The reinforcement-learning algorithm relies on market signals generated by patenting events. The patent value in the KPSS dataset is useful for constructing a proxy for these signals. It is inferred from the firm's abnormal stock return in a three-day window around the patent grant and therefore captures the market's valuation of the patent when the grant becomes public.[5]

[5]For details on the construction of the KPSS measure, see Kogan et al. (2017).

To isolate the patent-specific component of this valuation, I run the following regression:

$$\log \xi_i = \alpha \log(1 + C_i) + \gamma_{f(i)} + \gamma_{t(i)\times x(i)} + \varepsilon_i \tag{21}$$

where i indexes patents, ξ_i is the KPSS patent value, C_i is the number of forward citations to patent i, $\gamma_{f(i)}$ is a firm fixed effect, and $\gamma_{t(i)\times x(i)}$ is an issue-year-by-field fixed effect. The estimated residual $\hat{\varepsilon}_i$ is the component of log patent value unexplained by citations, persistent firm-level differences, and common field-year variation. $\hat{\varepsilon}_i > 0$ means that the market response to patent i was unusually favorable relative to other patents in the same field-year and other patents filed by the firm, while $\hat{\varepsilon}_i < 0$ indicates an unusually unfavorable response.

I interpret $\hat{\varepsilon}_i$ as a proxy for the patent-specific market signal $r^f_{t_i}(x_{t_i})$ in the illustrative model. Because the KPSS measure is inferred from the firm's abnormal stock return in a narrow window around the patent grant, it provides a direct measure of investors' contemporaneous response to the patent, whereas citations primarily capture technological importance. As the ChatGPT example illustrates, innovators may understand the scientific and technical merits of an innovation better than its commercial prospects. The market response can therefore provide information that is valuable to the firm. Residualizing patent values helps isolate variation in market assessments beyond citations, persistent firm differences, and common field-year conditions.

I then aggregate these $\hat{\varepsilon}$ signals to the firm-field-year level. Let $N_{f,x,t}$ denote the cumulative number of patents that firm f has filed in field x by year t; this is the firm's knowledge stock used below. Because a patent's market signal becomes observable at issuance, let t^I_i denote patent i's issue year and define the number of observed signals by

$$M^I_{f,x,t} \equiv \sum_i \mathbf{1}\{f(i) = f,\ x(i) = x,\ t^I_i \le t\}. \tag{22}$$

For firm-field-years with $M^I_{f,x,t} > 0$, the accumulated market signal is

$$\hat{\varepsilon}_{f,x,t} \equiv \frac{1}{M^I_{f,x,t}} \sum_{i:\, f(i)=f,\, x(i)=x,\, t^I_i \le t} \hat{\varepsilon}_i. \tag{23}$$

Thus, $\hat{\varepsilon}_{f,x,t}$ is the average market signal that firm f has observed by year t about field x and serves as a proxy for $\hat{v}^f_t(x)$ in the illustrative model. This distinction keeps the filing-date knowledge stock $N_{f,x,t}$ separate from the issue-date information set used to construct signals.

3.2 Testing the Learning Mechanism

I next use these measures to test the three predictions derived from the illustrative model.

Test 1: Directional Internal Innovation The first test asks whether a firm reallocates internal innovation toward the fields in which it has received more favorable market signals. For each field x already in firm f's portfolio at the beginning of year t, let $P_{f,x,t}$ be the share of the firm's year-t patent filings assigned to that field. The main specification is the Poisson regression

$$\mathbb{E}\left[P_{f,x,t} \mid \mathcal{I}_{f,t-1}\right] = \exp\left\{\alpha \log(1 + N_{f,x,t-1}) + \beta \hat{\varepsilon}_{f,x,t-1} + \delta_{f\times t} + \delta_x\right\} \tag{24}$$

where $N_{f,x,t-1}$ is the stock of patents that the firm had filed in field x by $t-1$, and $\hat{\varepsilon}_{f,x,t-1}$ is the average residualized market signal from its field-x patents issued by $t-1$. I estimate equation (24) by PPML and cluster standard errors at the firm level. The coefficient of interest is β, which measures whether the firm directs a larger share of its current filings toward an existing field with a better prior signal. Firm-year fixed effects control the firm's total patenting in year t, along with any common change in its R&D budget or innovation opportunities. Field fixed effects absorb persistent differences in patenting intensity across CPC subclasses. The lagged patent stock is included to separate signal-driven reinforcement from the familiar knowledge-stock mechanism. Both the quality-ladder model and the Acemoglu-style directed-innovation model predict that larger existing fields attract more innovation, i.e., $\alpha > 0$, but do not predict an additional response to the realization of the market signal, i.e., $\beta = 0$. The reinforcement learning mechanism in our illustrative model instead predicts $\beta > 0$: firms reallocate their resources internally towards fields that receive more favorable market responses.

Table 1 supports both knowledge-based and signal-dependent path dependence. The coefficient on $\log(1 + N_{f,x,t-1})$ of 1.156 confirms the knowledge accumulation channel: firms continue to innovate in fields where they have already accumulated a larger knowledge stock. More importantly, conditional on this stock, the coefficient on $\hat{\varepsilon}_{f,x,t-1}$ is 0.1178 and statistically significant at the 1% level. This means that a one-unit increase in the cumulative average signal is associated with a 12.5% increase in the expected filing share relative to the firm's other existing fields in that year. This pattern is consistent with experience-dependent allocation of innovation at the firm level under the reinforcement-learning framework.

Dependent Variable:	Filing share
Variables	
$\hat{\varepsilon}_{f,x,t-1}$	0.1178***
	(0.0160)
$\log(1 + N_{f,x,t-1})$	1.156***
	(0.0069)
Fixed-effects	
Firm × year	Yes
Field	Yes
Fit statistics	
Observations	1,239,635
Pseudo R^2	0.41157
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*	

Table 1: Internal innovation is directed towards high-signal fields. Each observation is at firm-field-year level. The sample covers 1981–2020. SEs are clustered by firms.

Test 2: Directional External Innovation The second test studies the direction of external innovation: when a firm expands beyond its existing portfolio, is it more likely to enter a candidate field that is technologically close to fields in which its prior patents received high market signals?

Unlike the internal innovation test (24), external fields have no firm-specific signal history of their own. I therefore cannot directly observe how a firm evaluates fields that it has not previously explored. Each firm has a different patent and field portfolio and history of signals, and it would be infeasible to include separate explanatory variables for each prior field because the number of explanatory variables would differ across firms with heterogeneous innovation histories. To handle this challenge, I use a flexible similarity-weighted average to summarize how firms use their past patenting experiences in evaluating an external field. Specifically, for every field x' outside firm f's portfolio at $t-1$, I define firm f's similarity-weighted evaluation of field x' as

$$S_{f,t-1}(x';\phi) \equiv \frac{\sum_{x\in\mathcal{P}_{f,t-1}} s(x',x)^{\phi}\hat{\varepsilon}_{f,x,t-1}}{\sum_{x\in\mathcal{P}_{f,t-1}} s(x',x)^{\phi}}, \tag{25}$$

where $\mathcal{P}_{f,t-1}$ is the set of fields in the firm's portfolio before the entry decision. The logic of this construction is similar to the experience-weighting function in Malmendier and Nagel (2011) and Malmendier and Nagel (2016). Those studies weight an individual's macro-

financial experience by its recency; here, the weighting dimension is technological proximity between the firm's prior experience and a candidate external field. This similarity-weighted value can be regarded as the data counterpart of $\hat{v}_t^f(x')$ for the external field x': the signals accumulated from an internal field x receive a higher weight in the evaluation of external field x' if the two fields are technologically similar to each other. The curvature parameter ϕ, which will be jointly estimated from the data, controls the strength of this extrapolation and plays a similar role as the parameter φ in the illustrative model. When $\phi = 0$, all technologically connected portfolio fields receive equal weight, so that the evaluations of external fields are not biased towards technologically similar internal fields. When $\phi > 0$, experience in fields more similar to x' receives greater weight. Thus, a positive estimate of ϕ indicates that firms extrapolate prior experience locally in technological space as in the illustrative reinforcement-learning model. Taken together, $S_{f,t-1}(x';\phi)$ is high when the portfolio fields most similar to x' have favorable accumulated signals. Favorable experiences in nearby fields therefore contribute most strongly to the firm's evaluation of the candidate field.

Similarly, I construct a similarity-based knowledge base for the candidate external field x' as the similarity-weighted average of patent counts in the firm's existing fields:

$$N_{f,t-1}(x';\phi) \equiv \frac{\sum_{x\in\mathcal{P}_{f,t-1}} s(x',x)^{\phi}\log(1+N_{f,x,t-1})}{\sum_{x\in\mathcal{P}_{f,t-1}} s(x',x)^{\phi}} \tag{26}$$

which captures how much the firm knows about x' based on the knowledge accumulated in the firm's internal fields.

I next specify the test for directed external innovation. For each eligible firm-year with at least one first entry, I compare candidate CPC subclasses after excluding fields with both positive prior filing stock and an observed market signal. Appendix B.3 gives the sample and estimation details. Let $\text{Enter}_{f,x',t}$ indicate whether firm f first enters field x' in year t. I consider a Poisson regression

$$\mathbb{E}\left[\text{Enter}_{f,x',t} \mid \mathcal{I}_{f,t-1}\right] = \exp\left\{\beta S_{f,t-1}(x';\phi) + \gamma N_{f,t-1}(x';\phi) + \delta_{f\times t} + \delta_{x'}\right\} \tag{27}$$

and simultaneously estimate coefficients β, γ and the weighting parameter ϕ by maximizing the likelihood. The coefficients β and γ capture the sensitivity of entry decision into a new field x' to experienced market signals and accumulated knowledge. The canonical quality ladder model predicts that β and γ should both be zero since the external search is undirected. Acemoglu's directed innovation models would predict $\gamma > 0$, i.e. firms continue to innovate

in fields associated with larger knowledge stocks because of increasing returns to innovation. Neither model speaks about the weighting parameter ϕ. Under reinforcement learning, we should expect that $\beta > 0$ and $\phi > 0$: the external innovation is directed towards fields with high similarity-weighted signals after controlling for the firm's existing knowledge, and greater weight is placed on the fields that are more similar to the candidate field x'.

Table 2 supports the predictions of the reinforcement learning model. In Column (1), the coefficient on the similarity-based signals is positive and significant at the 1% level, after controlling for the knowledge base of the firm. To interpret the magnitude of the coefficients, for each firm in each year, I compute the difference between the 90th and 10th percentiles for the similarity-weighted signals, take the median of the gap across firm-year observations, and multiply by the raw coefficients in Table 2. Within a firm in a given year, a candidate field at the 90th percentile of the similarity-weighted signal is about 7.7% more likely to be entered than one at the 10th percentile. The comparison indicates that U.S. public firms direct their external R&D towards fields that receive higher market signals. The estimated weighting parameter $\phi = 0.482$ implies that market responses and knowledge from technological fields that are more similar to the candidate external field receive greater weight in the evaluation of the candidate field.

Column (2) of Table 2 investigates whether the dependence of entry probabilities on experienced signals comes from entering fields with favorable similarity-weighted signals or from avoiding fields with unfavorable similarity-weighted signals. I estimate the following slope-shift version of regression (27):

$$\begin{aligned}\mathbb{E}\left[\text{Enter}_{f,x',t} \mid \mathcal{I}_{f,t-1}\right] = \exp\Big\{ & \\ & \beta^{-} S_{f,t-1}(x';\phi) + \Delta\beta\, S_{f,t-1}(x';\phi) D^{+}_{f,t-1}(x';\phi) \\ & + \gamma N_{f,t-1}(x';\phi) + \delta_{f\times t} + \delta_{x'}\Big\}.\end{aligned} \tag{28}$$

where $D^{+}_{f,t-1}(x';\phi) \equiv \mathbf{1}\{S_{f,t-1}(x';\phi) > 0\}$, β^{-} is the slope for unfavorable signals, and $\Delta\beta \equiv \beta^{+} - \beta^{-}$ is the change in the slope when the similarity-weighted signal is positive. Thus, the slope for favorable signals is $\beta^{-} + \Delta\beta = \beta^{+}$, and the coefficient on the interaction directly tests the null hypothesis $H_0 : \beta^{+} = \beta^{-}$. Column (2) reports an unfavorable-signal slope of $\widehat{\beta}^{-} = 0.6432$ (SE $= 0.1352$), which is positive and statistically significant at the 1% level. The estimated slope difference is $\widehat{\Delta\beta} = -0.5807$ (SE $= 0.2212$), and the null of equal slopes is rejected ($p = 0.0086$). The implied favorable-signal slope is $\widehat{\beta}^{+} = 0.0625$ and is not statistically different from zero. Hence, the positive pooled association between experienced

signals and entry is concentrated on the unfavorable-signal side: firms disproportionately avoid candidate fields associated with more unfavorable prior market signals.

Dependent Variable:	Entry	
	Baseline	Slope Difference
Model:	(1)	(2)
Experienced Signals β	0.4268***	
	(0.1053)	
Knowledge base γ	2.882***	2.864***
	(0.0878)	(0.0879)
Unfavorable-Signal Slope β^-		0.6432***
		(0.1352)
Slope Difference $\Delta\beta = \beta^+ - \beta^-$		-0.5807***
		(0.2212)
Weighting Parameter ϕ	0.482***	0.482 [fixed]
	(0.034)	
Fixed-effects		
Firm-year (choice occasion)	Yes	Yes
Candidate field	Yes	Yes
Fit statistics		
Observations	9,763,345	9,763,345
Squared Correlation	0.01727	0.01727
Pseudo R^2	0.17280	0.17285

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2: External innovation is directed towards high-signal fields, and higher weights are placed on more similar fields. The results in Column (1) are obtained by maximum likelihood estimation. The sample period is 1990–2020. Each observation is a candidate external field for a firm in a year. The sample contains 16,742 firm-years with at least one field entry and 38,248 entry events. Column (2) keeps the weighting parameter ϕ fixed at the estimate in Column (1) and reports the slope-difference specification in regression (28); $\Delta\beta$ directly tests whether the favorable- and unfavorable-signal slopes are equal. SEs are clustered by firm for parameters other than ϕ. The SE for ϕ is computed using a firm-block bootstrap.

Test 3: Unfavorable Signals Push External Innovation Farther Away Test 2 shows that external entry is positively associated with similarity-weighted feedback, with the strongest evidence coming from avoidance of fields associated with unfavorable experience. I now examine whether this pattern also appears in the distance of realized external entry. As I have established in the illustrative model, under reinforcement learning, a favorable signal in

an existing field raises the perceived value of nearby fields and hence reinforces the external search towards these nearby fields, whereas an unfavorable signal reduces their perceived value and may redirect external search toward more distant parts of the technological space. Thus, conditional on the firm's accumulated knowledge, the distance between an external entry and an existing field should be negatively related to the signals previously experienced in that field.

For each external field x' that firm f enters in year t, I pair the entry with every field x in the firm's pre-entry portfolio $\mathcal{P}_{f,t-1}$. The outcome is the angular technological distance

$$d(x', x) = \arccos\left(s(x', x)\right),$$

where similarity is measured using the predetermined citation network: the network constructed in decade d is used for entry decisions in decade $d + 1$. I retain all entry-event-by-existing-field pairs rather than selecting fields according to their realized distance from the entry. I estimate

$$d(x', x) = \beta\hat{\varepsilon}_{f,x,t-1} + \gamma \log(1 + N_{f,x,t-1}) + \delta_{f\times t} + \delta_{x'} + \delta_x + u_{f,x,x',t}, \tag{29}$$

where $\delta_{f\times t}$ denotes firm-year fixed effects, $\delta_{x'}$ entered-field fixed effects, and δ_x existing-field fixed effects. Firm-year fixed effects absorb factors common to all of a firm's external entries in a year, including its overall resources and propensity to explore. Entered- and existing-field fixed effects absorb persistent differences in the centrality and average distance of fields in the technological network. The coefficient β is therefore identified from differences in signals across the fields in a firm's existing portfolio, conditional on their accumulated patent stocks. Reinforcement learning predicts $\beta < 0$: an external field should land closer to an existing field for which the firm has experienced a more favorable signal.

Column (1) of Table 3 supports this prediction. The coefficient on the accumulated signal is -0.0005 and statistically significant at the 1% level. In terms of magnitude, a one-standard-deviation increase in the signal is associated with shorter technological distances by approximately 2 percentile points on average in the firm's own distribution of technological distances.

The pooled coefficient in Column (1) can result from two sources: unfavorable signals can induce firms to leave a poorly evaluated technological neighborhood, and favorable signals may generate movements toward an already attractive neighborhood. I decompose the pooled

coefficient using the slope-shift specification

$$d(x', x) = b^{-}\hat{\varepsilon}_{f,x,t-1} + \Delta b\,\hat{\varepsilon}_{f,x,t-1}D^{+}_{f,x,t-1} + \gamma \log(1 + N_{f,x,t-1}) + \delta_{f\times t} + \delta_{x'} + \delta_{x} + u_{f,x,x',t}, \tag{30}$$

where

$$D^{+}_{f,x,t-1} \equiv \mathbf{1}\{\hat{\varepsilon}_{f,x,t-1} > 0\}, \qquad \Delta b \equiv b^{+} - b^{-}.$$

The coefficient b^{-} is the distance response to unfavorable signals, while $b^{-} + \Delta b = b^{+}$ is the response to favorable signals. The interaction coefficient Δb therefore directly tests whether the two slopes are equal.

Column (2) shows that the negative relationship is concentrated on the unfavorable-signal side. The estimated unfavorable-signal slope is $\widehat{b^{-}} = -0.0011$ and is statistically significant at the 1% level. Within this region, a one-standard-deviation deterioration in the signal increases technological distance by approximately 4.6 percentile points. This supports the interpretation that firms shift their external search paths away from negative-signal fields. Interestingly, the data show limited support for a response on the favorable-signal side. The estimated slope difference is $\widehat{\Delta b} = 0.0012$ and is statistically significant at the 1% level. The implied favorable-signal slope is only $\widehat{b^{+}} = 0.00004$ and is not statistically different from zero. Thus, favorable signals do not appear to pull external innovation toward existing fields.

3.3 Discussion

In sum, the empirical evidence from the patenting histories of U.S. public firms is consistent with the predictions of the illustrative model. First, firms reallocate R&D resources toward high-signal internal fields. Second, they are more likely to enter new fields that are technologically associated with high-signal internal fields. Finally, the technological distance of external entry is negatively associated with signals from existing fields: firms direct external search away from the neighborhoods of low-signal internal fields.

The first limitation concerns the interpretation of the residualized market response. Although $\hat{\varepsilon}_i$ measures patent-specific public market feedback, it is neither a direct observation of the firm's belief nor an exogenous shock to that belief. The residual may contain fundamental patent quality that was already known to the firm, investors' expectations about the firm's future innovation plans, contemporaneous firm news, and measurement error. The design therefore cannot fully distinguish firms learning from market feedback from firms acting on

Dependent Variable:	Angular distance $d(x', x)$	
	Baseline	Slope shift
Model:	(1)	(2)
Variables		
Accumulated signal (β in (1), b^- in (2))	-0.0005***	-0.0011***
	(0.0001)	(0.0002)
Signal slope difference $\Delta b = b^+ - b^-$		0.0012***
		(0.0002)
$\log(1 + N_{f,x,t-1})$	0.0001	0.0001
	(0.0002)	(0.0002)
Fixed effects		
Firm × year	Yes	Yes
Entered field	Yes	Yes
Existing field	Yes	Yes
Fit statistics		
Observations	2,166,979	2,166,979
R^2	0.18385	0.18387

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 3: Path dependence of external search. Each observation is an entry-event-by-existing-field pair. The sample spans 1990–2020 and contains 38,251 entry events for 3,006 firms. Distance is measured on the predetermined decade citation network.

private information that is also reflected in market valuations. The timing restrictions, fixed effects, and controls for accumulated patent stocks reduce some of these concerns but do not eliminate them. The results should consequently be interpreted as conditional behavioral patterns consistent with reinforcement learning, rather than as causal estimates of the effect of stock-market feedback on R&D allocation. Future work could improve the measurement of market feedback by constructing an independent measure of investors' patent-specific evaluations. For instance, coverage of patenting news in financial newspapers, analyst reports, and other market commentary could be linked to individual patents using firm identifiers, event dates, and textual similarity to patent abstracts. The tone of this commentary would provide a direct measure of whether investors viewed a patent's prospects favorably or unfavorably. Residualizing this measure with respect to firm-wide sentiment, technology-wide sentiment, and observable patent fundamentals could further isolate its patent-specific and potentially nonfundamental component. Comparing such a measure with the residualized stock-market response would help validate the current proxy and clarify whether subsequent innovation responds to information about fundamentals, investor sentiment, or both.

Another limitation of the current analysis is that technological proximity is measured using a generated backward-citation network aggregated to CPC subclasses. Because citation vectors are sparse, most pairwise cosine similarities are close to zero. This concentration does not mechanically eliminate the identifying variation, but it implies that economically meaningful comparisons may depend on small—and potentially noisy—differences in the estimated network. The field-level construction may also suppress heterogeneity within CPC subclasses and reflect citation practices, including examiner-added citations, in addition to technological relatedness. The estimates should therefore be interpreted as conditional on this particular representation of technology space. Several alternatives are worth considering. One possible alternative is to aggregate CPC subclasses into broader CPC classes to reduce the number of small cosine similarities. In practice, this aggregation does not resolve the problem. The majority of resulting cosine similarities under that construction are either 0 or 1, indicating that the granularity is insufficient for this definition. A promising next step is to construct an independent text-based geometry using natural-language-processing techniques, following the methods in Argente et al. (2020), Bergeaud et al. (2025) or Carvalho and Freund (2026). Future work could create embeddings for each patent using patent titles and abstracts, cluster patents in the resulting semantic space, and use these representations to trace firms' innovation paths. For the present design, continuous patent embeddings have two merits. First, they may help with clustering the patents at a more appropriate level of granularity, potentially more granular than CPC classes and coarser than CPC subclasses,

so that the distribution of cosine similarities is better-behaved. Additionally, they could be used to construct predetermined field centroids and distances using only patents observed before each entry decision, after which the external-entry tests could be repeated using the text-based measure.

Lastly, apart from the measurement issues, the current tests are not able to rule out alternative mechanisms that links market feedback to subsequent innovation. A related mechanism is analyzed by Angeletos et al. (2023), in which investors' beliefs affect IPO valuations and entrepreneurs' opportunities to expand. If innovation projects rely heavily on external funding, then investor's sentiment might determine the direction and intensity of innovation. Firms could consequently redirect innovation toward technologies favored by investors without revising their own assessments of fundamental returns. If financing opportunities favor particular technological areas, this channel could generate both continued innovation within those fields and expansion into related technologies. The evidence therefore does not separately identify the contributions of external financing and firms' value learning. Distinguishing these channels would require additional evidence on firms' assessments of project returns and the financing of their innovation activities, which I am working on.

4 Quantitative Model

This section embeds the reinforcement-learning mechanism in a canonical quality-ladder environment for quantitative analysis. The quantitative exercise has three purposes. First, it disciplines the prior uncertainty, signal noise, and spatial reach of learning using U.S. patent data. Second, it assesses whether the calibrated model reproduces the conditional path dependence documented in Section 3, including its disappearance when innovation is undirected. Third, it allows us to study how market sentiment can shape the path of innovation.

4.1 Model Setup

In the model, time is continuous and the horizon is infinite.

Household A representative consumer supplies a unit of labor inelastically to the labor market, owns the firms in the economy and consumes $C(t)$ units of final goods in each period.

His lifetime utility is $\int_0^\infty \exp(-\rho t)\log(C(t))\,dt$.

Final Good Producer Final goods are produced by labor and a continuum of intermediate goods $x \in [0,1]$, with technology

$$Y(t) = \frac{L(t)^\theta}{1-\theta}\int_0^1 q(x,t)^\theta y(x,t)^{1-\theta}\,dx$$

where $q(x,t)$ is the quality of intermediate good x. Here, the intermediate good index x should be regarded as "fields", and each field is associated with a vector defining its technological characteristics, as in Section 2. The price of the final good is normalized to 1, and the input and output markets are competitive. The final good producer chooses $\{y(x,t)\}$ and $L(t)$ that maximizes the profit, given prices $\{p(x,t)\}$ and wage $w(t)$.

Intermediate Good Producers Intermediate good producers sell differentiated goods in a monopolistically competitive market. Each field is dominated by one firm which takes the demand function into account. The production technology for intermediate good x is

$$y(x,t) = \bar{q}(t)\cdot l(x,t)$$

where $\bar{q}(t) \equiv \int_0^1 q(x,t)\,dx$ is the average quality of intermediate goods. Solving the profit maximization problem of intermediate producers and combining it with the labor market clearing condition gives us the monopolistic profit of each intermediate good x:

$$\pi(x) = \underbrace{\frac{\theta^{\theta+1}(1-\theta)^{2-2\theta}}{\theta+(1-\theta)^2}}_{\equiv\bar{\pi}} q(x)$$

Appendix A.2 gives the derivations.

Innovation under Full Information As in Akcigit and Kerr (2018), intermediate good producers are engaged in two types of innovations: internal innovation improves products in fields the firm already owns, while external innovation enables entry into fields it does not currently own.

Successful internal innovation in field x improves the quality of the field by a proportion λ:

$$q(x, t + \Delta t) = (1 + \lambda)q(x, t)$$

The internal innovation arrival rate $z(x)$ is chosen at a cost

$$c(z(x)) = \chi z(x)^{\psi} \tag{31}$$

where c is the innovation intensity on x. Total innovation expenditure on field x is equal to $c(x) \cdot q(x)$, as in Section 2.

Similarly, successful external innovation in field x' improves the quality of the field by a proportion η, with $\eta > \lambda$:[6]

$$q(x', t + \Delta t) = (1 + \eta)q(x', t)$$

The external innovation arrival rate $z_e(x')$ is subject to a cost function

$$e(z_e(x')) = \tilde{\chi} z_e(x')^{\tilde{\psi}} \tag{32}$$

where e is the innovation intensity on x'. Total external expenditure is equal to $e(x') \cdot \bar{q}$, as in Section 2.

Under full information, the step sizes η and λ are known to the firms in the economy. For an incumbent firm that currently owns fields $\{x_1, \cdots, x_k\}$ with quality portfolio $\mathbf{q} \equiv \{q_1, \cdots, q_k\}$, the value function is

$$\begin{aligned} rV(\mathbf{q}) - \dot{V}(\mathbf{q}) = \max_{\{c\},e,x'} & \sum_{i=1}^{k} \bar{\pi} q_i - \sum_{i=1}^{k} c_i q_i - e(x')\bar{q} \\ & + \sum_{i=1}^{k} z_i \left[V((1+\lambda)q_i) - V(q_i)\right] - \tau \sum_{i=1}^{k} V(q_i) \\ & + z_e(x') \cdot \mathbb{E}_j \left[V((1+\eta)q_j)\right] \end{aligned}$$

where τ is the creative destruction rate, a Poisson rate at which a product is taken over by another firm. As in Section 2, I assume that the firms only know the quality of products they own, but not the quality of products owned by other firms. Since V is linear in q (which will be shown below), under full information, the expected payoff of successfully innovating

[6] External innovation or creative destruction is assumed to be of higher quality than internal innovation. This assumption is standard in the literature and supported empirically by Jo and Kim (2025).

in field x' is given by $\mathbb{E}_j[V((1+\eta)q_j)] = A(1+\eta)\mathbb{E}[q_j] = A(1+\eta)\bar{q}$, and hence the search of external field is undirected.

Appendix A.2 shows that the optimal internal and external innovation arrival rates are given by

$$z = \left(\frac{A\lambda}{\chi\psi}\right)^{1/(\psi-1)}, z_e = \left(\frac{(1+\eta)A}{\tilde{\chi}\tilde{\psi}}\right)^{1/(\tilde{\psi}-1)}$$

and the value function $V(\mathbf{q})$ takes the form

$$V(\mathbf{q}) = \sum_{i=1}^{k} Aq_i + B_k\bar{q}$$

where A and B_k are constants that are pinned down in general equilibrium. This value function consists of two components: the first part, $\sum_{i=1}^{k} Aq_i$, is the value of innovating on existing fields, and the second part, $B_k\bar{q}$, is the option value from external innovation. The value of each existing field x_i can therefore be written as

$$V(q_i, x_i) = Aq_i \tag{33}$$

We can see that neither the innovation arrival rate z_i nor the value of field $V(q_i, x_i)$ depends on field x_i. This $V(q_i, x_i)$ has the same structure as the value function in Section 2: under this full-information benchmark, the value function coefficient $v^*(x) = A$ for every field x. In the current setup, A is defined by equation

$$(r+\tau)A = \max_z\{\bar{\pi} - \chi z^{\psi} + \lambda z A\} \tag{34}$$

and the optimal innovation intensity c^* is defined as

$$c^* = \frac{(r+\tau)A - \bar{\pi}}{\psi - 1} \tag{35}$$

which is also identical across fields.

Besides the incumbent firms, there is an exogenously given mass M_0 of potential entrants. The entrant chooses an innovation arrival rate z_0 subject to a cost $\nu z_0\bar{q}$, where ν is an entry cost. The value function of entrants is therefore

$$rV_0 - \dot{V}_0 = \max_{z_0} z_0(\mathbb{E}_j[V((1+\eta)q_j)] - V_0) - \nu z_0\bar{q}$$

Free entry requires $V_0 = 0$. The balanced growth path of the full-information economy is characterized in Appendix A.2.

Innovation under Reinforcement Learning I now incorporate the reinforcement-learning framework of Section 2 into the quantitative model. Firms observe their current product qualities, profit flows, and research costs, and know how research expenditure determines innovation arrival rates. However, firms do not know the distribution of quality improvements generated by internal or external innovation, or whether these distributions differ across technological fields. Following Section 2, firms therefore maintain estimates of field values and revise these estimates using market feedback from successful innovations.

If firms could obtain unbiased market feedback infinitely often in every technological field, they could learn the true innovation values in each field $v^*(x)$ and their innovation decisions would converge to the full-information benchmark $c^*(x)$ in (35). However, this is not feasible because of creative destruction: the fact that other firms may take over a field from a firm limits both the duration of the firm owning a particular field and the number of fields in which the firm can accumulate experience. As a result, the firm can only use some estimate of field value instead of the true value $v^*(x)$ to guide its choice of innovation intensity and direction.

As in Section 2, firm f maintains an estimate of value $\hat{V}_t^f(q, x)$ for innovating in field x with quality q. With the primitives that the firm knows, it can back out the functional form of $\hat{V}_t^f(q, x)$ as a linear function of q:

$$\hat{V}_t^f(q, x) = q\hat{v}_t^f(x) \tag{36}$$

where $\hat{v}_t^f(x)$ is the firm's estimate of the value function coefficient $v^*(x)$. Although the true value function coefficient $v^*(x)$ is identical across fields, the firms' estimates $\hat{v}_t^f(x)$ can differ across fields and firms because they reflect different innovation histories. These estimates guide research expenditure and the direction of external innovation. The resulting choices determine where and how frequently the firm receives new market feedback, which subsequently revises its valuations.

Firms generalize their experience across technological fields using Gaussian-process beliefs about $v^*(x)$, as in Section 2. Specifically, each firm begins with the subjective prior

$$v^*(\cdot) \sim \mathcal{GP}(\hat{v}_0, k_0), \qquad \text{with } k_0(x, x') = \sigma_0^2 \exp\left[-\frac{d(x, x')^2}{2\varphi}\right], \tag{37}$$

where $d(x, x')$ measures distance between the technological characteristics of the two fields. The parameter σ_0^2 captures prior uncertainty, while φ determines the degree of generalization. The estimate $\hat{v}_t^f(x)$ is the posterior mean conditional on firm f's observed history. I set the common prior mean as $\hat{v}_0(x) = A$, so that initial valuations coincide with the full-information benchmark.

Suppose the firm's j-th innovation occurs at time t_j in field x_j. Following the innovation, the market generates a signal $\xi_{t_j}^f(x_j)$ that takes the form

$$\xi_{t_j}^f(x_j) = \Delta q_{t_j}^f(x_j) \cdot v^*(x_j) + \varepsilon_{t_j}^f(x_j) \tag{38}$$

where $\epsilon_t^f(x)$ follows a Gaussian distribution with mean 0 and variance σ_ε^2, and the quality gain $\Delta q_{t_j}^f(x_j)$ is equal to $q_{t_j}^f(x_j) - q_{t_j-\Delta t}^f(x_j)$ if field x_j is owned by the firm and $q_{t_j}^f(x_j)$ if x_j is an external field to the firm. Since $\Delta q_{t_j}^f$ is observable by the firm, it can use the reward signal $r_{t_j}^f(x_j) \equiv \frac{\xi_{t_j}^f}{\Delta q_{t_j}^f}$ to update its estimate $\hat{v}_t^f(x)$ by

$$\hat{v}_{t_j}^f(x) = \hat{v}_{t_{j-1}}^f(x) + \alpha_{t_j}(x, x_j) \cdot [r_{t_j}^f(x_j) - \hat{v}_{t_{j-1}}^f(x_j)] \tag{39}$$

where the learning rate $\alpha_{t_j}(x, x_j) = \frac{k_{t_{j-1}}^f(x,x_j)}{k_{t_{j-1}}^f(x_j,x_j)+(\sigma_{r,j}^f)^2}$ with $k_{t_{j-1}}^f(x, x_j)$ being the posterior covariance between field x and x_j following the $(j-1)$-th innovation, and $(\sigma_{r,j}^f)^2 \equiv \frac{\sigma_\varepsilon^2}{\Delta q_{t_j}^f(x_j)^2}$ being the variance of the reward signal noise.[7] It takes the same form as the Gaussian-process updating rule in (16) in Section 2.

The firm then uses the updated valuations to determine subsequent innovation expenditure and the direction of external research. With $\hat{v}_t^f(x)$, the firm chooses the innovation intensity $c_t^f(x)$ according to the decision rule (35), replacing the full-information value coefficient $v^*(x)$ with the estimated value coefficient $\hat{v}_t^f(x)$:

$$c_t^f(x) = \max\left\{\frac{(r+\tau)\hat{v}_t^f(x) - \bar{\pi}}{\psi - 1}, 0\right\} \tag{40}$$

Without further internal innovation, a field generates a value of $\bar{\pi}/(r+\tau)$ per unit of quality

[7] The posterior covariance is updated according to

$$k_{t_j}^f(x', x'') = k_{t_{j-1}}^f(x', x'') - \frac{k_{t_{j-1}}^f(x', x_j) k_{t_{j-1}}^f(x_j, x'')}{k_{t_{j-1}}^f(x_j, x_j) + (\sigma_{r,j}^f)^2}.$$

before it is lost through creative destruction. The difference $\hat{v}_t^f(x) - \bar{\pi}/(r+\tau)$ therefore measures the firm's estimated additional value from further internal innovation. The decision rule translates this perceived innovation premium into research expenditure.

Under reinforcement learning, even firms with the same current fields and product qualities can have different value estimates and research expenditures because they have experienced different market feedback. Relative to the full-information benchmark, firms' signal histories therefore provide an additional source of heterogeneity in innovation decisions. If the firm has received many favorable signals in field x, its value estimate $\hat{v}_t^f(x)$ is high, which increases the firm's innovation effort $c_t(x)$ in this field. Higher innovation intensity, in turn, raises the arrival rate of successful innovations in the field, allowing the firm to receive market feedback and accumulate experience more frequently.

The firm's history of innovation and market signals also changes the incentives for external innovation. Under full information, the external search is undirected. Under reinforcement learning, this is generally not the case because the evaluation $\hat{v}_t^f(x')$ of an external field x' depends on the patent and signal history of the firm. Let $\mathbf{X}_t^f$ denote the fields currently owned by firm f. I retain the assumption that the firm is uncertain about the quality of goods in external fields, so that the expected qualities of external fields are $\bar{q}$. Its perceived value of stealing field $x' \notin \mathbf{X}_t^f$ from another firm is therefore $\bar{q}(t)\hat{v}_t^f(x')$. Given the expenditure–arrival mapping $z_e(e) = \left(\frac{e}{\tilde{\chi}}\right)^{1/\tilde{\psi}}$, the firm chooses external research intensity and direction to maximize its perceived innovation payoff:

$$\max_{e\geq 0, x'\notin \mathbf{X}_t^f} \left\{ z_e(e)\hat{v}_t^f(x')\bar{q} - e\bar{q} \right\}.$$

The external innovation arrival rate and expenditure intensity are given by

$$z_{e,t}^f = \left(\frac{\max\left\{\hat{v}_t^f(x_t^f), 0\right\}}{\tilde{\chi}\tilde{\psi}} \right)^{1/(\tilde{\psi}-1)}, \qquad e_t^f = \tilde{\chi}\left(z_{e,t}^f\right)^{\tilde{\psi}},$$

and the resulting field choice is

$$x_t^f \in \arg\max_{x'\notin \mathbf{X}_t^f} \hat{v}_t^f(x')$$

with ties broken at random. Since the choice of external field depends on $\hat{v}_t^f(x')$ which is history-dependent and not flat across fields x' in general, the external search under reinforcement learning is directed. With the kernel function (37) and the learning rate in the value updating rule (39), the estimated values of external fields are more closely correlated with

the experienced fields nearby. Hence, if a firm has accumulated favorable market feedback from field x, the estimated values of fields close to x can be revised upwards, making the firm more likely to explore these fields in the external search.

4.2 Calibration

I calibrate the model using the simulated method of moments. In the simulation, each period is 1/50 year, and the technological space is represented by 100 fields arranged on a 10×10 grid. I simulate a panel of 1,000 firms for 2,500 periods and compute the simulated moments after discarding the first 250 periods as burn-in. Note that the simulated panel should be regarded as 1,000 independent simulations of measure-zero reinforcement-learning firms interacting with a universe of full-information firms, and therefore all the aggregate equilibrium objects, such as growth rate and creative destruction rate, come from the full-information balanced growth path. Each reinforcement-learning firm in the panel starts with one field with quality $q = 1$ and Gaussian process prior $\hat{v}_0(x)$ and $k_0(x, x')$ as before. Each reinforcement-learning firm can take over a field from a full-information firm at its chosen external innovation arrival rate, or lose a field to a full-information firm at the equilibrium creative-destruction rate τ. There is no business stealing between two reinforcement-learning firms in the panel. If a firm loses its last field, it exits the simulation panel with its identity and posterior beliefs discarded, and the simulation slot is replaced by a new one-field entrant.

There are in total 12 parameters in the model. I separate them into two categories. The three parameters in Gaussian-process learning, namely, the prior uncertainty σ_0, signal noise σ_ε and the degree of generalization φ, are estimated internally and will be discussed in detail shortly. The rest of the parameters are externally calibrated. Discounted rate $\rho = 0.02$ matches an annual discount factor of 0.98. The curvature parameters of the innovation cost functions ψ and $\tilde{\psi}$ are set to 2, a standard value in the literature. The labor share in final good production function θ is set to match the average profitability of the economy. The remaining five parameters, namely the step sizes λ ane η, innovation cost scales χ and $\tilde{\chi}$, and entry cost ν, are taken from Akcigit and Kerr (2018). The parameter values used in the quantitative exercise are reported in the upper panel of Table 4. Solving the resulting full-information balanced growth path gives us a creative-destruction rate $\tau = 0.1071$, an aggregate growth rate $g = 0.0141$, and an interest rate $r = 0.0341$. These equilibrium objects remain fixed when I vary the learning parameters.

The three parameters governing the GP learning problem are selected to match three mo-

Parameter	Interpretation	Source	Value
θ	Labor share in final-good production	External	0.1059
ρ	Discount rate	External	0.0200
λ	Internal innovation step	External	0.0509
η	External innovation step	External	0.1124
χ	Internal R&D cost	External	0.3462
ψ	Internal R&D cost curvature	External	2.0000
$\widetilde{\chi}$	External R&D cost	External	4.0665
$\widetilde{\psi}$	External R&D cost curvature	External	2.0000
ν	Entry cost	External	0.8303
σ_0	Prior uncertainty	Internal	1.8873
σ_ε	Signal noise	Internal	0.7990
φ	Generalization parameter	Internal	0.4409

Table 4: Parameter Values in Calibration

ments from the KPSS dataset by simulated method of moments. Let $\boldsymbol{\vartheta} = (\sigma_0, \sigma_\varepsilon, \varphi)'$ denote the vector of learning parameters, let $\mathbf{m}^D$ denote the three empirical moments, and let $\mathbf{m}^S(\boldsymbol{\vartheta})$ denote their simulated counterparts. I choose the parameters to minimize

$$Q(\boldsymbol{\vartheta}) = \left[\mathbf{m}^S(\boldsymbol{\vartheta}) - \mathbf{m}^D\right]' \widehat{\Omega}^{-1} \left[\mathbf{m}^S(\boldsymbol{\vartheta}) - \mathbf{m}^D\right],$$

where $\widehat{\Omega}$ is the estimated covariance matrix of the empirical moments.[8]

Now I explain the targeted moments and how they help pin down the learning parameters. The three parameters affect all three moments jointly, so this interpretation describes the primary source of identifying variation rather than a one-to-one assignment of moments to parameters. The first moment is the standard deviation of the patent-level residual $\hat{\varepsilon}_p$ constructed in Section 3. Since we used the residual as a proxy for the market signal to a new patent, the signal-dispersion moment therefore provides the most direct information about σ_ε. The second moment is the average within-size variance of annual patent growth. Under the full-information benchmark, firms with the same number of fields have the same innovation rates. Learning adds variation in these rates through differences in signal histories and estimated field values. The resulting dispersion in innovation intensity generates larger within-size dispersion in patent growth than the full-information case. Conditional on the

[8]I estimate $\widehat{\Omega}$ using 300 firm-block bootstrap draws of the KPSS sample. Each draw resamples firms with replacement, preserves the complete patent history of each selected firm, and re-estimates both the patent-signal first stage and all three downstream moments. I set $\widehat{\Omega}$ equal to the sample covariance matrix of the resulting bootstrap moment vectors.

signal-noise scale disciplined by the first moment, this moment is particularly informative about σ_0, which controls the magnitude of prior uncertainty and hence the dispersion in research expenditure induced by differences in estimated field values.

The third moment resembles the correlation between external entry distance and the signal realizations in the internal field in Test 3 of Section 3. Before each first entry into a new field, I define the firm's "core field" as the field with the largest historical patent stock and calculate that field's share within the firm's patent portfolio as the "core share". I then express the distance between the entered field and the core as its percentile among all external fields available to the firm before the entry. The targeted moment is the correlation between the core share and this entry-distance percentile. This moment captures the difference between reinforcement-learning and canonical quality-ladder model in their spatial properties of innovation over the technological space. Under reinforcement learning, a higher core share results from the core field accumulating more favorable signal realizations and higher posterior mean, which directs the external search towards fields close to the core field and makes the correlation negative. Under the canonical quality-ladder model, since the search is undirected, this correlation should be zero. In the data, the correlation is indeed negative. This moment is most directly informative about φ, because φ determines the extent to which signals from a firm's accumulated experience are informative about other fields in the technological space.

Table 5 shows that the model closely reproduces the targeted moments, and the lower panel of Table 4 tabulates the estimated parameters. We can see that the estimated market signals are informative to firms, with the own-field Kalman gain being approximately 0.84. Moreover, the estimated $\varphi = 0.4409$ indicates that the reinforcement-learning firms make inference about nearby fields from signals and that the inference decays as the distance between fields gets larger in the technological space. It lies between the limiting cases of fully local ($\varphi = 0$) and uniform learning ($\varphi = \infty$).

Moment	Data	Model
Signal SD	0.794	0.796
Average within-size growth variance	0.591	0.591
Correlation between core share and entry distance	-0.230	-0.233

Table 5: Fit of the three-parameter SMM calibration

Untargeted validation I assess the mechanism outside the targeted moments by applying versions of the empirical internal- and external-innovation tests in Section 3 to the simulated panel. These exercises are most informative about signs and the distinction between reinforcement learning and canonical quality-ladder models with undirected innovation. Raw coefficient magnitudes involving technological distance should not be compared literally: the model contains 100 regularly spaced fields with Euclidean distance, whereas the empirical counterpart contains 640 CPC subclasses, with distances based on backward-citation similarity.

The simulated model reproduces the positive internal-allocation response to accumulated signals. The simulated coefficient for Test 1 is 0.803, compared with 0.118 in the data. Thus the direction agrees with the empirical evidence, but the parsimonious model substantially overstates the magnitude of internal reinforcement. This is not surprising: in the quantitative model, learning is the only source of heterogeneity in innovation intensity within firm sizes, and the discrepancy between the model-implied coefficient and data indicates that there are other sources of firm heterogeneity accounting for this within-size variation.

The simulated model is also consistent with patterns of external innovation we highlighted in Section 3. In the simulated analogue of the external-distance regression in Test 3 of Section 3, the signal coefficient is -0.469 under reinforcement learning and approximately zero under an undirected no-learning policy. The reinforcement-learning model therefore directs external entry toward fields associated with more favorable experience. In the test of direction of external innovation in Test 2 of Section 3, the simulated coefficient on the similarity-weighted signals is 2.581 in response to a one-standard-deviation increase in signals, compared with 0.313 in the data. Under no learning, the signal coefficient is close to zero and the localization profile is essentially uninformative. The estimated weighting parameter ϕ from the simulated panel is approximately 0.126, compared with the 0.483 estimated from the data. Although these estimates are not directly comparable since the simulation and data are based on different specifications of technological space, they verify that the mechanism of the reinforcement-learning model is consistent with the empirical patterns I documented for external search.

Taken together, the untargeted results support the qualitative mechanism: reinforcement learning generates internal reallocation of R&D resources towards high-signal fields and directs external innovation towards the neighbors of high-signal internal fields. They are less successful at matching the magnitudes of these responses.

4.3 Counterfactual: Sentiment as a Driver of Innovation Bursts

I now use the quantitative model to run a counterfactual exercise. John Maynard Keynes used the term "animal spirits" to describe the willingness of entrepreneurs to undertake long-horizon, risky investments without knowing the returns precisely. The same force applies naturally to R&D investments: the economic value of a new technology is highly uncertain, and, as I have shown, firms need external signals, such as market responses to their patents, to inform them of which technological directions appear promising. If those responses contain an episodic sentiment component, firms may mistake favorable market valuations for information about fundamental innovation opportunities.

The reinforcement-learning model provides a convenient framework to study the implications of sentiment for innovation. During a sentiment episode, I modify the reward signal specification (38) and let it take the form

$$r^f_{t_j}(x_j) = v^*(x_j) + \Delta_s \mathbf{1}\{x_j \in \mathcal{T},\ t_j \in [t_0, t_1)\} + \frac{\varepsilon^f_{t_j}(x_j)}{\Delta q^f_{t_j}} \tag{41}$$

where $\mathcal{T}$ is the set of fields affected by the sentiment and $\Delta_s > 0$ is a common optimistic shift. Firms observe the combined signal but not its decomposition. They therefore update their GP posterior as if the sentiment component were information about $\tilde{v}(x)$. Note that here I am assuming that the sentiment does not change the fundamentals of the firms and the economy; it affects the firms and the economy only by shifting the belief of reinforcement-learning firms towards a particular field for a period of time.

In the first exercise, I introduce a wave of optimism to a connected 3×3 block of nine fields for ten years, on top of the 1,000-firm, 50-year simulated panel.[9] The size of sentiment shift Δ_s is $2\sigma_\varepsilon$. For comparison, I simulate another panel of firms with identical innovation and signal histories, without the sentiment episode. The beliefs of reinforcement-learning firms in the sentiment and no-sentiment simulations are therefore identical until the sentiment shock hits the economy.

Figure 4 plots the difference in the share of patents from the shocked fields between the sentiment economy and the no-sentiment benchmark. We can see that the sentiment shock can indeed shift the direction of innovation under reinforcement learning. During the sentiment episode, the patent share of the shocked block increases over time. At the end of

[9] Here, among all possible 3×3 blocks, I choose the median block in the pre-sentiment activity distribution. The shocked block accounts for 8.92 percent of learning-sector patenting before the experiment.

the sentiment episode, the treated block accounts for 31.29 percent of annual patents, compared with 7.76 percent in the paired no-sentiment economy, a difference of 23.53 percentage points. The shares of intermediate and distant fields drop accordingly. Moreover, the effect of the sentiment shock seems persistent. The patent-share difference between sentiment and no-sentiment economy in the shocked fields remains 23.24 percentage points one year after sentiment ends, 9.90 points after five years, and 6.77 points after ten years. Twenty years after the episode, the remaining difference in the shocked fields is 1.08 percentage points, and the patent share of other fields catches up gradually.

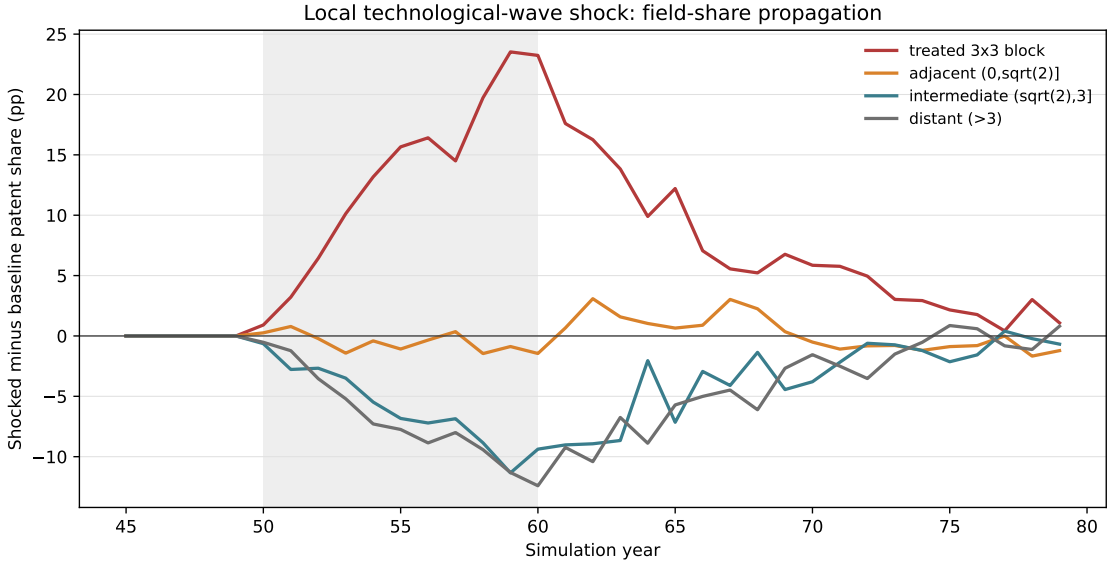


Figure 4: The Effect of a Local Sentiment Wave. The graph plots the difference in the patent share of the sentiment-shocked field between the sentiment economy and the no-sentiment benchmark. The shaded area is the sentiment episode. Signal histories are identical before the shocked episode.

The wave of optimism provides the firms that have innovated in the shocked fields with persistently high realizations of signals during the sentiment episode. These firms will reallocate more R&D resources towards the sentiment fields, which increases the probability of successful patents in these fields. These new patents in turn generate more signals with persistent optimism from these fields, which further reinforces these firms' beliefs about the prospects of innovating in these fields. This feedback loop between belief and innovation is the key to generating the persistent reallocation of innovative resources in response to the sentiment shock.

Besides field-specific sentiment, the same methodology can be applied to study the implications of aggregate sentiment for innovation. After simulating a panel of 1,000 firms for

50 years (2,500 periods), I introduce the same two-standard-deviation shift to all 100 fields for 10 years, and continue the simulation for 20 more years. The upper panel of Figure 5 plots the time series of annual patents for the sentiment economy and an otherwise identical no-sentiment economy. It shows that the aggregate sentiment introduces a large and persistent patenting boom. During the sentiment episode, the learning sector produces 11,935 patents, compared with 6,894 in the paired baseline, an increase of 73.1 percent. At the end of the episode, annual patenting remains 120.5 percent above baseline. Moreover, the difference persists after the sentiment fades away. After five years, the annual patents of the sentiment economy remains 44.8 percent higher than the benchmark, 24.4 percent higher after ten years, and 8.6 percent higher after twenty years.

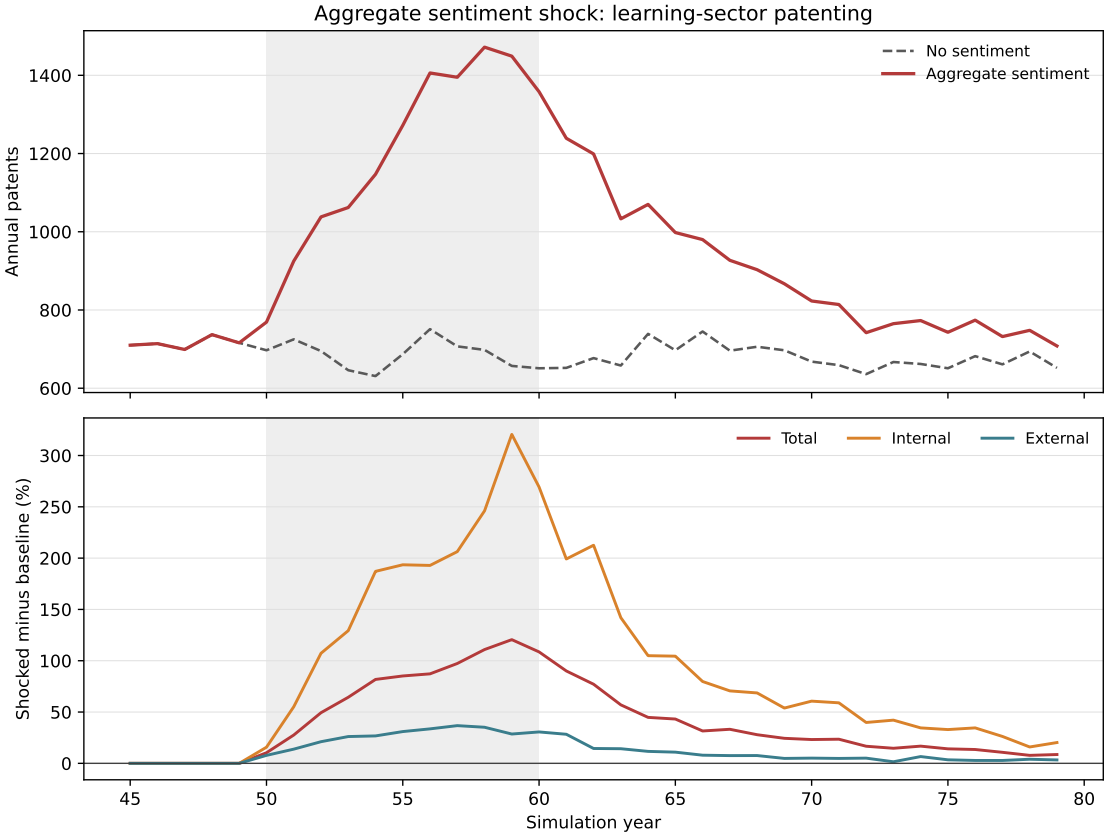


Figure 5: Aggregate Sentiment and Learning-Sector Patenting

The lower panel of Figure 5 shows that internal innovation is the main driver of the patenting boom in response to aggregate sentiment shock. Cumulative internal patenting during the sentiment episode rises by 164.5 percent, whereas external patenting rises by 26.0 percent. Even ten years after sentiment ends, internal patenting remains 53.9 percent above baseline; after twenty years it remains 20.3 percent higher. In comparison, external innovation responds more slowly and less strongly than internal innovation. The reason for this pattern is

that, in our reinforcement-learning model, the feedback loop between innovation expenditure and beliefs is much stronger for internal innovation. Since the firms do not have signals in the external fields, it's difficult for the aggregate sentiment to shift the firms' beliefs on these fields. The beliefs on the internal fields, however, keep moving upwards due to the sentiment, and the reallocation of R&D resources towards these fields generate more signals with the drift from aggregate optimism. Hence, aggregate sentiment creates the innovation boom by reinforcing the own-product innovation instead of encouraging explorations in new fields.

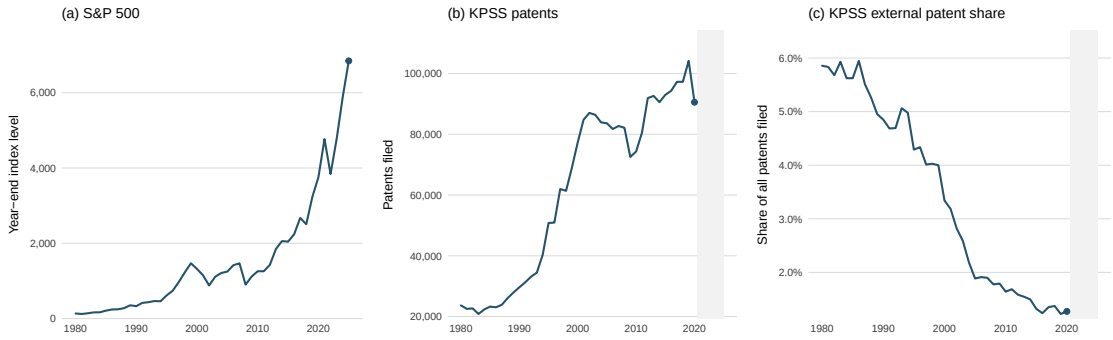


Figure 6: Recent trends of S&P500, total patent filings and share of external patents for U.S. public firms.

This pattern in reinforcement learning is highly relevant to two salient features of stock market returns and innovation of U.S. firms in recent decades. It is well-documented that there is a downward trend of external innovation and creative destruction in the past several decades (e.g. Garcia-Macia et al. (2019), Bloom (2020), Akcigit and Ates (2021), Akcigit and Ates (2023)), while the aggregate number of patents filed by U.S. public firms has been rising persistently. In the meantime, the U.S. stock market has been persistently booming, and the firms have more resources for R&D than before. These empirical patterns are illustrated in Figure 6. The reinforcement-learning mechanism provides a possible interpretation of these facts. An aggregate boom in the stock market leads the firms to update their beliefs about the values of their existing fields more and allocate more resources to internal innovation. This in turn generates more favorable signals in the internal fields. In contrast, the boom propagates to evaluations of external fields much slower because the firms need to enter the field before generating the signals. Hence the share of external innovation drops even though the aggregate patent filing is rising.

In sum, the two experiments illustrate that the reinforcement-learning model provides a convenient and promising framework to analyze the consequences of sentiment on the path of innovation. Field-specific optimism can persistently reallocate innovation toward a popular technological region. Aggregate optimism raises the level of aggregate patenting while

lowering the share of external patents.

5 Conclusion

This paper shows that when firms are uncertain about the value of patenting over the technological space, learning from experience plays a key role in shaping the path and direction of innovation. I develop a novel quality-ladder model that incorporates reinforcement learning and allows firms to use noisy market signals to update their estimates of field values across technologically related fields. Consistent with the model, U.S. public firms subsequently direct more internal innovation toward fields receiving favorable feedback, enter new fields associated with high-signal technologies, and move away from the technological neighborhoods of fields receiving unfavorable feedback. More favorable feedback in owned fields is associated with entry into technologically closer fields. A calibrated version of the model further shows that temporary market sentiment can produce large and persistent changes in both the level and technological composition of innovation through a feedback loop between beliefs, R&D allocation, and subsequent signals. These findings identify a source of conditional path dependence distinct from the conventional knowledge-stock mechanism: even firms with similar technological capabilities can follow different innovation paths because they experience different histories of market feedback.

References

Acemoglu, D. (2002): "Directed Technical Change," *The Review of Economic Studies*, 69, 781–809.

Acemoglu, D., P. Aghion, L. Bursztyn, and D. Hemous (2012): "The Environment and Directed Technical Change," *American Economic Review*, 102, 131–166.

Acemoglu, D., U. Akcigit, and W. R. Kerr (2016): "Innovation network," *Proceedings of the National Academy of Sciences*, 113, 11483–11488.

Aghion, P. and P. Howitt (1992): "A Model of Growth Through Creative Destruction," *Econometrica*, 60, 323–351.

Akcigit, U. and S. T. Ates (2021): "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory," *American Economic Journal: Macroeconomics*, 13, 257–298.

——— (2023): "What Happened to US Business Dynamism?" *Journal of Political Economy*, 131, 2059–2124.

Akcigit, U. and W. R. Kerr (2018): "Growth through Heterogeneous Innovations," *Journal of Political Economy*, 126, 1374 – 1443.

Akcigit, U. and Q. Liu (2016): "The Role of Information in Innovation and Competition," *Journal of the European Economic Association*, 14, 828–870.

Angeletos, G.-M., G. Lorenzoni, and A. Pavan (2023): "Wall Street and Silicon Valley: A Delicate Interaction," *The Review of Economic Studies*, 90, 1041–1083.

Argente, D., S. Baslandze, D. Hanley, and S. Moreira (2020): "Patents to Products: Product Innovation and Firm Dynamics," *SSRN Electronic Journal*.

Barberis, N. and L. J. Jin (2023): "Model-free and Model-based Learning as Joint Drivers of Investor Behavior," *SSRN Electronic Journal*.

Bergeaud, A., A. B. Jaffe, and D. Papanikolaou (2025): "Natural Language Processing and Innovation Research," .

Berlingieri, G., M. D. Ridder, D. Lashkari, and D. Rigo (2026): "Creative destruction through innovation bursts," .

Bloom, N. (2020): "Are Ideas Getting Harder to Find?" *THE AMERICAN ECONOMIC REVIEW*, 110.

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

Carvalho, V. M. and L. B. Freund (2026): "Firms as Foragers," .

Garcia-Macia, D., C.-T. Hsieh, and P. J. Klenow (2019): "How Destructive Is Innovation?" *Econometrica*, 87, 1507–1541.

Gershman, S. J. and N. D. Daw (2017): "Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework," *Annual Review of Psychology*, 68, 101–128.

Grossman, G. M. and E. Helpman (1991): "Quality Ladders in the Theory of Growth," *The Review of Economic Studies*, 58, 43.

Hodgson, C. and G. Lewis (2025): "You Can Lead a Horse to Water: Spatial Learning and Path Dependence in Consumer Search," *Econometrica*, 93, 1299–1332.

Ilut, C. and R. Valchev (2024): "Learning Optimal Behavior Through Reasoning and Experiences," ArXiv:2403.18185 [econ].

Jaffe, A. (1986): "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value," Tech. Rep. w1815, National Bureau of Economic Research, Cambridge, MA.

Jo, K. and S. Kim (2025): "Heterogeneous Innovations and Growth Under Imperfect Technology Spillovers," *SSRN Electronic Journal*.

Kaplowitz, B. (2026): "Reinforcement Learning and Consumption-Savings Behavior," .

Klette, T. and S. Kortum (2004): "Innovating Firms and Aggregate Innovation," *Journal of Political Economy*, 112, 986–1018.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017): "Technological Innovation, Resource Allocation, and Growth*," *The Quarterly Journal of Economics*, 132, 665–712.

Lei, X. and C. Shi (2024): "Risk, Uncertainty and Entrepreneurship," .

Lentz, R. and D. Mortensen (2008): "An Empirical Model of Growth Through Product Innovation," *Econometrica*, 76, 1317–1373.

Malmendier, U. and S. Nagel (2011): "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?*," *The Quarterly Journal of Economics*, 126, 373–416.

——— (2016): "Learning from Inflation Experiences *," *The Quarterly Journal of Economics*, 131, 53–87.

Moll, B. (2024): "The Trouble with Rational Expectations in Heterogeneous Agent Models: A Challenge for Macroeconomics," *Working Paper*.

Plate, R. C., H. Ham, and A. C. Jenkins (2023): "When uncertainty in social contexts increases exploration and decreases obtained rewards." *Journal of Experimental Psychology: General*, 152, 2463–2478.

Rasmussen, C. E. and C. K. I. Williams (2006): *Gaussian processes for machine learning*, Adaptive computation and machine learning, Cambridge, Mass: MIT Press, oCLC: ocm61285753.

Redding, S. (2002): "Path Dependence, Endogenous Innovation, and Growth*," *International Economic Review*, 43, 1215–1248.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis (2016): "Mastering the game of Go with deep neural networks and tree search," *Nature*, 529, 484–489.

Sutton, R. S. and A. G. Barto (2018): *Reinforcement learning: an introduction*, Adaptive computation and machine learning series, Cambridge, Massachusetts: The MIT Press, second edition ed.

Woodford, M. (2019): "Monetary Policy Analysis When Planning Horizons Are Finite," *NBER Macroeconomics Annual*.

Wu, C. M., E. Schulz, M. Speekenbrink, J. D. Nelson, and B. Meder (2018): "Generalization guides human exploration in vast decision spaces," *Nature Human Behaviour*, 2, 915–924.

Yang, Y., C. Wang, A. Schaab, and B. Moll (2026): "Structural Reinforcement Learning for Heterogeneous Agent Macroeconomics," *Manuscript*.

A Theory Appendix

A.1 Proof of Proposition 1

To save some notations, suppress the field index x and denote $\mu \equiv \mathbb{E}[\lambda(x)]$. The true value function coefficient defined in (2) solves

$$\frac{1}{2}\mu^2 v^2 - rv + \bar{\pi} = 0 \tag{A.1}$$

and the value function coefficient v^* is the smaller root of this quadratic function:

$$v^* = \frac{r - \sqrt{r^2 - 2\bar{\pi}\mu^2}}{\mu^2}. \tag{A.2}$$

Define v^c as the value of the firm in field x if it chooses to have innovation intensity c permanently. Hence v^c satisfy the Bellman equation

$$rv^c = \bar{\pi} - c + z(c)\mu v^c \tag{A.3}$$

By definition, $v^* = \max_c v^c$ and therefore $v^c \leq v^*$ for any feasible c.

Now, consider the value updating step from v^c to $v^{c'}$ where $c' = rv^c - \bar{\pi}$. The new value $v^{c'}$ is

$$rv^{c'} = \bar{\pi} - c' + z(c')\mu v^c$$

and consider $v^{c'} - v^c$:

$$\begin{aligned} r(v^{c'} - v^c) &= \bar{\pi} - c' + \sqrt{2c'}\mu v^c - rv^c \\ &= \sqrt{2c'}\mu v^c - 2c' = \sqrt{2c'}(\mu v^c - \sqrt{2c'}) \geq 0 \end{aligned} \tag{A.4}$$

The second equality follows from the definition of c', and the last inequality follows from the fact that $v^c < v^*$, so that

$$0 = \mu^2(v^*)^2 - 2(rv^* - \bar{\pi}) \leq \mu^2(v^c)^2 - 2(rv^c - \bar{\pi}) = \mu^2(v^c)^2 - 2c' \tag{A.5}$$

Hence, adopting the decision rule $c' = rv^c - \bar{\pi}$ necessarily improves the value of the firm in the field. And since v^c is bounded above by v^*, the algorithm must converge to some upper bound $U \leq v^*$, by Monotone Convergence Theorem.

The last step of the proof is to establish $U = v^*$. This is straightforward. Suppose $U < v^*$, then we could always do the iteration again and end up with $U' \in (U, v^*]$ that is larger than U. Hence, the limit must be $U = v^*$.

A.2 Full Model in Section 4

Household A representative consumer supplies a unit of labor inelastically to the labor market, owns the firms in the economy and consumes C_t units of final goods in each period. His lifetime utility is $\int_0^\infty \exp(-\rho t) \log(C(t))\, dt$.

Final Good Producer Standard. Final goods are produced by labor and a continuum of intermediate goods $x \in [0, 1]$, with technology

$$Y(t) = \frac{L(t)^\theta}{1-\theta} \int_0^1 q(x,t)^\theta y(x,t)^{1-\theta}\, dx \tag{A.6}$$

The final good producer chooses y_x and L that solves

$$\max \frac{L(t)^\theta}{1-\theta} \int_0^1 q(x,t)^\theta y(x,t)^{1-\theta}\, dx - w(t)L(t) - \int_0^1 p(x,t)y(x,t)\, dx \tag{A.7}$$

which implies that

$$L(t)^\theta q(x,t)^\theta y(x,t)^{-\theta} = p(x,t) \tag{A.8}$$

$$\frac{\theta}{1-\theta} L(t)^{\theta-1} \int_0^1 q(x,t)^\theta y(x,t)^{1-\theta}\, dx = w(t) \tag{A.9}$$

Intermediate-good Producer Each field is dominated by one firm which takes the demand function into account. The production technology for intermediate good is

$$y(x,t) = \bar{q}(t) \cdot l(x,t) \tag{A.10}$$

and hence the profit max problem is

$$L(t)^\theta q(x,t)^\theta y(x,t)^{1-\theta} - \frac{w(t)}{\bar{q}} y(x,t) \tag{A.11}$$

Hence FOC becomes

$$L(t)^{\theta} q(x,t)^{\theta} (1-\theta) y(x,t)^{-\theta} = \frac{w}{\bar{q}} \tag{A.12}$$

Hence

$$y(x) = Lq(x)(1-\theta)^{1/\theta} \left(\frac{w}{\bar{q}} \right)^{-1/\theta} \tag{A.13}$$

Plugging this back into the wage equation gives us

$$w = (1-\theta)^{1-2\theta} \theta^{\theta} \bar{q} \tag{A.14}$$

Hence

$$y(x) = q(x) L \theta^{-1} (1-\theta)^2, p(x) = \frac{1}{1-\theta} \frac{w}{\bar{q}} = (1-\theta)^{-2\theta} \theta^{\theta} \tag{A.15}$$

Hence the monopolistic profit is

$$\pi(x) = \theta^{\theta} (1-\theta)^{2-2\theta} L q(x) \tag{A.16}$$

Demand for labor from intermediate good producer is

$$\tilde{L} = \int_0^1 l(x)\, dx = L\theta^{-1}(1-\theta)^2 \tag{A.17}$$

Labor market clearing condition is

$$L + \tilde{L} = 1 \tag{A.18}$$

and hence we get

$$L = \frac{\theta}{\theta + (1-\theta)^2}, \tilde{L} = \frac{(1-\theta)^2}{\theta + (1-\theta)^2} \tag{A.19}$$

Innovation under Rational Expectation Internal innovation: success probability $z(x)$, at a cost

$$R_z(z(x), q(x)) = \chi z(x)^{\psi} q(x) \tag{A.20}$$

In the event of success, the quality is improved by λ:

$$q(x, t + \Delta t) = (1+\lambda) q(x,t) \tag{A.21}$$

External innovation: success probability $e(x)$, at a cost

$$R_e(e(x), \bar{q}) = \tilde{\chi} e(x)^{\tilde{\psi}} \bar{q} \tag{A.22}$$

In the event of success, the quality is improved by $\eta > \lambda$, and the firm steals the field from another firm.

Value of an incumbent firm with quality portfolio $\mathbf{q} \equiv \{q_1, \cdots, q_k\}$ is

$$rV(\mathbf{q}) - \dot{V}(\mathbf{q}) = \max_{\{z\},e} \sum_{i=1}^{k} \bar{\pi} q_i - \sum_{i=1}^{k} \chi z_i^{\psi} q_i - \tilde{\chi} e^{\tilde{\psi}} \bar{q} \tag{A.23}$$

$$+ \sum_{i=1}^{k} z_i \left[V((1+\lambda)q_i) - V(q_i) \right] - \tau \sum_{i=1}^{k} V(q_i) \tag{A.24}$$

$$+ e \cdot \mathbb{E}_j \left[V((1+\eta)q_j) \right] \tag{A.25}$$

Note that the interpretation for the first term of the last line is slightly different from Akcigit and Kerr (2018). Here, I assume that the firms only know the quality of products they own, but not the quality of products owned by other firms. Since V is linear in q (will show below), under rational expectation, the expected payoff of successfully innovating in any field is equal to $(1+\eta)\mathbb{E}[q_j] = (1+\eta)\bar{q}$, and hence the search of external field is undirected.

Since the profit function and the internal innovation cost scales with q_i and the rest of the RHS scales with $\bar{q}$, the value function $V(\mathbf{q})$ takes the form

$$V(\mathbf{q}) = \sum_{i=1}^{k} A q_i + B_k \bar{q} \tag{A.26}$$

where A and B_k are constants.

Optimal innovation intensity is given by the first-order conditions

$$\chi \psi z_i^{\psi-1} q_i = \lambda A q_i \implies z_i = \left(\frac{A\lambda}{\chi\psi} \right)^{1/(\psi-1)} \tag{A.27}$$

$$\tilde{\chi} \tilde{\psi} e^{\tilde{\psi}-1} \bar{q} = A(1+\eta)\bar{q} \implies e = \left(\frac{(1+\eta)A}{\tilde{\chi}\tilde{\psi}} \right)^{1/(\tilde{\psi}-1)} \tag{A.28}$$

Constant B is equal to

$$B = A(1+\eta)e - \tilde{\chi} e^{\tilde{\psi}} \tag{A.29}$$

so that $B = 0$. Given r and τ, the undetermined coefficient A solves

$$rA = \bar{\pi} - \chi \left(\frac{A\lambda}{\chi\psi} \right)^{\psi/(\psi-1)} + A\lambda \left(\frac{A\lambda}{\chi\psi} \right)^{1/(\psi-1)} - \tau A \tag{A.30}$$

Entrants Standard. There is an exogenously given mass of entrants M_0. We have

$$rV_0 - \dot{V}_0 = \max e_0(\mathbb{E}_j[V((1+\eta)q_j)] - V_0) - \nu e_0 \bar{q} \tag{A.31}$$

Free entry requires $V_0 = 0$, and hence

$$A(1+\eta) = \nu \tag{A.32}$$

Equilibrium In equilibrium, we have

$$g = r - \rho = \tau\eta + z\lambda \tag{A.33}$$

The creative destruction rate τ is equal to

$$\tau = Me + M_0 e_0 \tag{A.34}$$

where M is the mass of the incumbent firms.

Lastly we pin down the distribution of firm size:

$$M\mu_1\tau = M_0 e_0$$

$$M\mu_2 2\tau + M_0 e_0 = M\mu_1(e+\tau)$$

$$\cdots$$

$$M\mu_{n+1}(n+1)\tau + M\mu_{n-1}e = M\mu_n(e+n\tau)$$

The distribution of μ_n has a closed-form solution:

$$\mu_n = \frac{M_0 e_0}{Me}\left(\frac{e}{\tau}\right)^n \frac{1}{n!} \tag{A.35}$$

And this distribution needs to be a probability distribution:

$$\sum_{n=1}^{\infty} \mu_n = \frac{M_0 e_0}{Me}[\exp(e/\tau) - 1] = 1 \tag{A.36}$$

which pins down e_0.

Also, the total number of product lines is measure 1, so we have

$$M \sum_{n=1}^{\infty} n\mu_n = 1 \tag{A.37}$$

which pins down M.

A.3 Construction of Targeted Moments

This subsection describes how I construct the three moments used to estimate the learning parameters $\boldsymbol{\vartheta} = (\sigma_0, \sigma_\varepsilon, \phi)$. The construction deliberately applies the same observable transformations to the KPSS data and to the simulated patent histories. Let

$$\mathbf{m}^D = \left(m^D_{\text{sig}}, m^D_{\text{grow}}, m^D_{\text{core}}\right)' \tag{A.38}$$

denote the empirical moments and let $\mathbf{m}^S(\boldsymbol{\vartheta})$ denote their simulated counterparts. I estimate the learning parameters by simulated method of moments:

$$\widehat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} \left[\mathbf{m}^S(\boldsymbol{\vartheta}) - \mathbf{m}^D\right]' \widehat{\Omega}^{-1} \left[\mathbf{m}^S(\boldsymbol{\vartheta}) - \mathbf{m}^D\right], \tag{A.39}$$

where $\widehat{\Omega}$ is the covariance matrix from a joint 300-draw firm-block bootstrap of the KPSS sample. Each bootstrap draw resamples firms with replacement, re-estimates the patent-value first stage, and reconstructs all downstream moments. The simulation uses fixed random-number streams across parameter evaluations, so changes in $\mathbf{m}^S(\boldsymbol{\vartheta})$ reflect parameter changes rather than simulation noise.

Moment 1: Dispersion of market signals. In the data, I use the residualized KPSS market response defined in the empirical section. For patent p, I estimate

$$\log \xi_p = \beta_C \log(1 + C_p) + \alpha_{f(p)} + \gamma_{x(p)\times t^I_p} + \widehat{\varepsilon}_p, \tag{A.40}$$

where ξ_p is real patent value, C_p is the truncation-adjusted forward-citation count, $f(p)$ is the patenting firm, $x(p)$ is its primary CPC subclass, and t^I_p is its issue year. Firm fixed effects remove persistent differences in firms' market responses, while field-by-issue-year fixed effects remove aggregate differences in the valuation of technological fields over time. The

empirical target is the patent-level sample standard deviation

$$m_{\text{sig}}^{D} = \text{sd}_p\left(\widehat{\varepsilon}_p\right). \tag{A.41}$$

The sample contains 2,798,247 patents issued between 1976 and 2020 by 9,063 firms, each with positive KPSS value and a matched firm and primary CPC subclass. The resulting target is $m_{\text{sig}}^{D} = 0.7939$.

The model counterpart uses the signal observed after every successful patent-generating event. After removing the component of the innovation prize known to the firm, the residual signal from event j is

$$s_j^F = v(x_j) + \varepsilon_j, \qquad \varepsilon_j \sim N(0, \sigma_\varepsilon^2). \tag{A.42}$$

The quantitative exercise sets the true latent field value to $v(x) = 0$ in every field, although firms do not know this and continue to update their Gaussian-process beliefs. In the benchmark specification, the econometrician observes the same residual signal as the firm. I therefore compute

$$m_{\text{sig}}^{S}(\boldsymbol{\vartheta}) = \text{sd}_j\left(s_j^F\right) \tag{A.43}$$

over all entrant, internal, and external patent events after the five-year simulation burn-in. This moment primarily disciplines the structural signal-noise parameter σ_ε.

Moment 2: Within-size dispersion of patent growth. The second moment captures the firm-level growth heterogeneity generated by different learning histories. In the KPSS data, let I_{ft} be the number of patents issued to firm f in calendar year t. I measure annual patent growth using the inverse-hyperbolic-sine transformation

$$g_{ft} = \text{asinh}(I_{ft}) - \text{asinh}(I_{f,t-1}), \tag{A.44}$$

which retains firm-years with zero patents. For each firm-year, beginning-of-year size is the number of distinct fields in the firm's historical patent portfolio,

$$B_{ft} = \left|\left\{x : N_{fx,t-1}^{I} > 0\right\}\right|, \tag{A.45}$$

where $N_{fx,t-1}^{I}$ is the cumulative number of firm-f patents issued in field x through $t-1$. I first residualize patent growth on firm and year fixed effects,

$$g_{ft} = \alpha_f + \lambda_t + \widetilde{g}_{ft}, \tag{A.46}$$

and divide observations into the size cells

$$\mathcal{B} = \{1,\ 2,\ 3\text{–}4,\ 5\text{–}9,\ 10+\}\,. \tag{A.47}$$

For each cell $b \in \mathcal{B}$, define the within-cell residual variance

$$D_b^D = \mathrm{Var}\left(\widetilde{g}_{ft} \mid B_{ft} \in b\right). \tag{A.48}$$

The scalar target is the empirical-observation-share-weighted average

$$m_{\text{grow}}^D = \sum_{b\in\mathcal{B}} \omega_b D_b^D, \qquad \omega_b = \frac{n_b^D}{\sum_{b'\in\mathcal{B}} n_{b'}^D}. \tag{A.49}$$

For the five size cells listed above, the weights are $(0.1507, 0.1134, 0.1558, 0.1892, 0.3909)$. The empirical panel spans each firm's first through last observed issue year, retaining intervening zero-patent years, and contains 76,575 firm-year observations from 6,096 firms. The resulting target is $m_{\text{grow}}^D = 0.5909$.

I construct the simulated counterpart in exactly the same way. I aggregate successful simulated patent events to annual firm counts I_{ft}^S, compute the growth rate in equation (A.44), measure beginning-of-year size by the number of fields in which the simulated identity has previously produced a patent, and residualize growth on simulated firm and year fixed effects. A simulated identity-year is retained only if the identity is active for the entire current and preceding calendar years, which removes mechanical exposure differences caused by within-year exit and replacement. Let $D_b^S(\boldsymbol{\vartheta})$ denote the resulting variance in size cell b. The model moment is

$$m_{\text{grow}}^S(\boldsymbol{\vartheta}) = \sum_{b\in\mathcal{B}} \omega_b D_b^S(\boldsymbol{\vartheta}), \tag{A.50}$$

using the fixed KPSS weights ω_b, rather than the endogenous simulated size distribution. This makes the comparison depend on within-size growth dispersion rather than on differences between the empirical and simulated firm-size distributions. Although current field ownership in the model can change through creative destruction, the observable size used for this moment is historical patent size, matching what is measured in the data.

Moment 3: Core share and external step size. The third moment disciplines how prior experience shapes the direction of external innovation. Consider firm f's first patent in a previously unentered field y in issue year t. Define its prior-year historical portfolio and

feasible set of new fields as

$$\mathcal{H}_{f,t-1} = \left\{x : N^I_{fx,t-1} > 0\right\}, \qquad \mathcal{C}_{f,t-1} = \mathcal{X} \setminus \mathcal{H}_{f,t-1}. \tag{A.51}$$

I restrict the sample to events with $|\mathcal{H}_{f,t-1}| \geq 2$. The firm's ex ante core field is the field with the largest cumulative patent stock before the entry,

$$x^*_{ft} = \arg\max_{x \in \mathcal{H}_{f,t-1}} N^I_{fx,t-1}. \tag{A.52}$$

Ties are broken first by the field's earliest entry date and then by its identifier. The core field's share of the prior patent portfolio is

$$S^*_{ft} = \frac{N^I_{fx^*_{ft},t-1}}{\sum_{x \in \mathcal{H}_{f,t-1}} N^I_{fx,t-1}}. \tag{A.53}$$

Let $d_t(y, x^*_{ft})$ denote the technological distance from the entered field to the core field, measured in the predetermined backward-citation network. Entry events begin in 1990; the network constructed from decade d is used for entry decisions in decade $d+1$ (for example, the 1980s network measures distances for 1990s entries). Because the number and geometry of fields differ between the CPC data and the simulated grid, I convert the realized distance into its midrank percentile among all fields that the firm could have entered:

$$R^*_{fyt} = \frac{\#\left\{z \in \mathcal{C}_{f,t-1} : d_t(z, x^*_{ft}) < d_t(y, x^*_{ft})\right\} + \frac{1}{2}\#\left\{z \in \mathcal{C}_{f,t-1} : d_t(z, x^*_{ft}) = d_t(y, x^*_{ft})\right\}}{|\mathcal{C}_{f,t-1}|}. \tag{A.54}$$

The empirical target is the raw event-level correlation

$$m^D_{\text{core}} = \text{Corr}\left(S^*_{ft}, R^*_{fyt}\right). \tag{A.55}$$

The sample contains 43,705 first-entry events from 3,736 firms and yields $m^D_{\text{core}} = -0.2302$. The negative correlation means that when a firm's prior patenting is more concentrated in its core field, its next external entry tends to lie closer to that field in the feasible technological opportunity set.

For the model counterpart, I retain first-ever external entries after the burn-in and reconstruct the simulated identity's historical portfolio using all patent events strictly before the entry year. I select the core field and compute its cumulative patent share using equations (A.52) and (A.53). The feasible set contains all grid fields not previously entered by the

identity. I then replace the CPC distance with shortest-path Euclidean distance on the periodic 10×10 model grid and compute the same midrank percentile in equation (A.54). The simulated moment is

$$m^S_{\text{core}}(\boldsymbol{\vartheta}) = \text{Corr}^S\left(S^*_{ft}, R^*_{fyt}\right), \tag{A.56}$$

again restricting to pre-entry size of at least two fields. The percentile transformation places the data and model outcomes on the same $[0, 1]$ support and handles the many equal-distance destinations on the model grid.

At the estimated benchmark parameters, the empirical target vector is

$$\mathbf{m}^D = (0.7939,\ 0.5909,\ -0.2302)', \tag{A.57}$$

while the simulated vector is

$$\mathbf{m}^S(\widehat{\boldsymbol{\vartheta}}) = (0.7956,\ 0.5907,\ -0.2327)'. \tag{A.58}$$

B Data Appendix

B.1 Empirical Data and Sample Construction

This subsection describes the construction of the patent-level data and the three estimation samples used in Section 3.

Data sources and firm identifiers. I use the 2024 release of the patent-value data accompanying Kogan et al. (2017), which provides patent identifiers, filing and grant dates, estimated patent values, and forward citations. I merge these records by patent number with the accompanying patent–firm crosswalk and identify firms using the company-level CRSP identifier PERMCO. This groups patents at the company level rather than the security-level PERMNO. The patent-level sample retains observations with positive real patent values, nonmissing filing and grant dates, valid firm identifiers, and a matched four-character CPC subclass. Patents must have been granted in 1976 or later. The resulting sample contains 2,798,247 patents granted during 1976–2024, belonging to 9,063 firms and spanning 636 CPC subclasses. The innovation-decision panels use filing years through 2020.

Technological fields and citation networks. I assign each patent to a single primary CPC subclass using PatentsView's current CPC classification. Specifically, I select the four-character subclass with the lowest classification sequence for that patent. Each patent contributes one unit to its assigned field; the baseline does not fractionally allocate patents across their classifications. The current classification provides a common field assignment across historical observations.

The backward-citation networks use the broader PatentsView patent universe rather than only patents matched to publicly traded firms. A patent may have several CPC classifications. To assign each patent to a single technological field, I use the four-character subclass of its first-listed CPC classification in PatentsView. I apply this assignment consistently when constructing patent counts, market-signal histories, firm portfolios, and citation networks. The network files collectively contain 640 subclasses. I use the preceding decade's network for each decision year: the 1980–1989 network serves decisions in 1990–1999, the 1990–1999 network serves decisions in 2000–2009, and the 2000–2009 network serves decisions in 2010–2019. Decisions in 2020 use the 2010–2019 network. Pairs absent from a network receive zero similarity. Appendix B.2 presents the resulting matrices.

Patent stocks and market-signal histories. Filing dates determine patenting outcomes, accumulated knowledge stocks, and first entry into a field. Grant dates determine when market feedback enters the measured signal history. Accordingly, for decisions in year t, the knowledge stock $N_{f,x,t-1}$ counts matched patents filed by firm f in field x through year $t-1$, whereas the accumulated signal includes only patents granted through year $t-1$. I obtain patent-level signals from the regression of log real patent value on log one plus forward citations, firm fixed effects, and field-by-grant-year fixed effects described in Section 3. This regression uses the full patent-level sample. Its residuals, $\hat{\varepsilon}_i$, are then assigned to firms' histories according to their grant dates. Thus, the full sample is used to estimate the residualization coefficients, while subsequent signal aggregation respects the timing of patent grants.

The baseline accumulated signal $\hat{\varepsilon}_{f,x,t-1}$ is the equal-weight average of the residuals on firm f's field-x patents granted through $t-1$. It remains unchanged in years without a new observed signal. A firm-field pair without an observed signal has a missing accumulated signal, rather than a signal of zero.

Test 1: Internal innovation. The internal-innovation panel covers filing years 1981–2020. I begin with firm-years containing at least one matched patent filing and pair them with fields in which the firm had filed before the current year. An observation is eligible if the field has a positive filing stock and an observed accumulated market signal by $t-1$. The outcome is the number of current filings in the field divided by the firm's total matched filings that year. The denominator includes filings in newly entered fields as well as existing fields. Consequently, filing shares over the eligible internal fields need not sum to one. Existing fields with no current filings receive a zero outcome and remain in the panel; firm-years without any observed filings are excluded. The constructed panel contains 1,275,390 firm-field-year observations, of which 76.6 percent have zero current filings. The reported PPML specification uses 1,239,635 observations.

Test 2: Direction of external innovation. The external-entry panel covers filing years 1990–2020. I include firm-years in which the firm enters at least one field for the first time and has at least two previously explored fields with positive patent stocks and observed market signals by the end of the preceding year. For each candidate, I calculate the similarity-weighted signal and knowledge indices in equations (25) and (26). When all similarity weights are zero, both indices are assigned zero. The constructed panel contains 10,014,330

candidate-field observations across 16,742 firm-years, including 38,248 first entries. The reported PPML specification uses 9,763,345 observations. Appendix B.3 describes estimation of the spatial weighting parameter.

Test 3: Distance of external innovation. The distance test also covers filing years 1990–2020. For each first entry, I pair the entered field with every eligible prior field of the firm. A prior field must have a positive filing stock and an observed accumulated signal by $t-1$, and the firm must have at least two such fields. The sample therefore includes all eligible entry–prior-field pairs, rather than selecting only the nearest prior field.

For each pair, the outcome is the angular distance between the entered and prior fields, calculated using the preceding decade's network. The resulting sample contains 2,166,979 pairs associated with 38,251 first-entry events, 16,742 firm-years, and 3,006 firms.

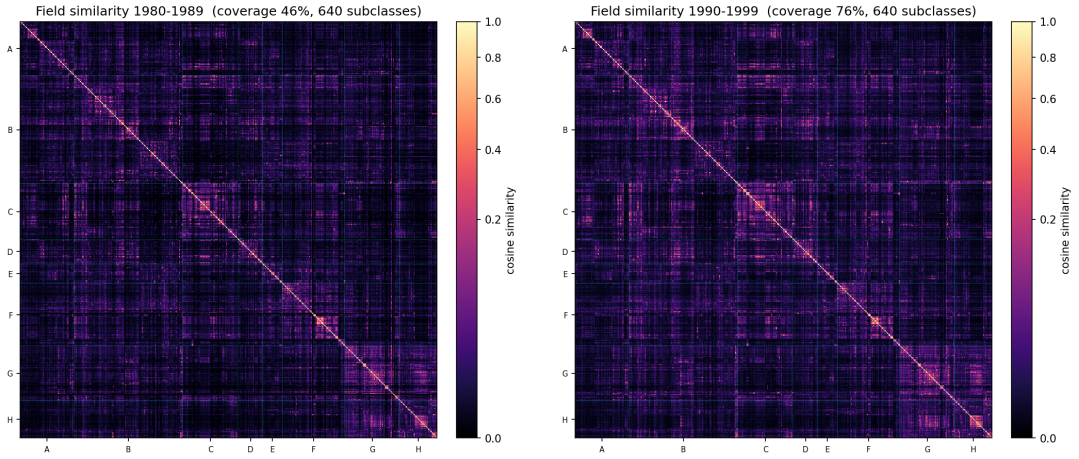
Table B.1: Empirical estimation samples

Specification	Observation unit	Years	Observations
Patent valuation	Patent	1976–2024	2,798,247
Test 1: Internal	Firm–field–year	1981–2020	1,239,635
Test 2: Entry	Firm–candidate–year	1990–2020	9,763,345
Test 3: Distance	Entry–prior-field pair	1990–2020	2,166,979

Notes: Counts refer to observations used in the reported baseline regressions, rather than all rows in the constructed panels. Years refer to grant years for the patent-valuation regression and filing years for the three innovation tests. Earlier observations are used to construct historical stocks, signal histories, and first-entry dates as described in the text.

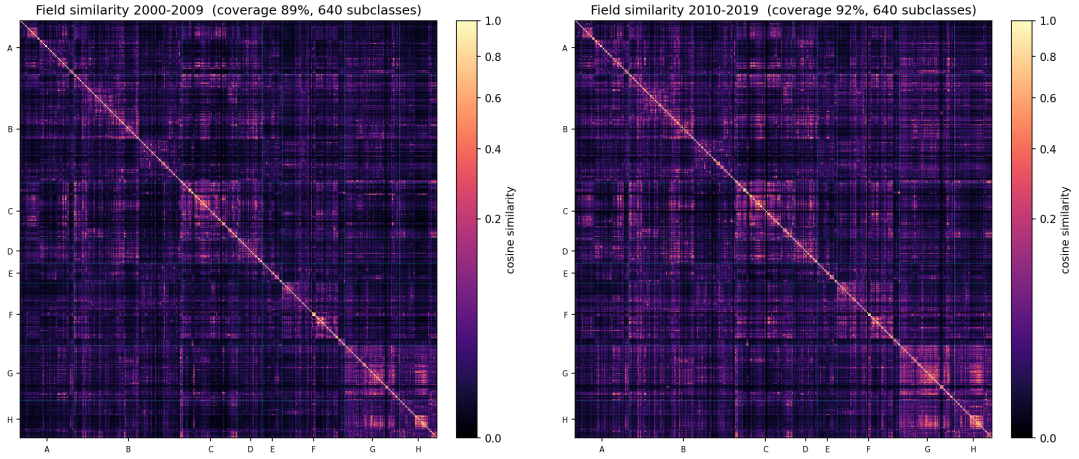
B.2 Similarity matrices by decade

Figure B.1 uses the same ordering of 640 CPC subclasses in all four panels. A brighter cell indicates that two fields cite similar bodies of prior technology. Each decade's citations measure proximity for entries in the following decade. For example, the 1980–1989 matrix serves entries in 1990–1999. Off-diagonal Pearson correlations between successive decades are 0.748, 0.787, and 0.786. I retain pairs with positive similarity in any decade and assign zero where a pair is absent.



(a) 1980–1989. Coverage 46%.

(b) 1990–1999. Coverage 76%.



(c) 2000–2009. Coverage 89%.

(d) 2010–2019. Coverage 92%.

Figure B.1: Backward-citation similarity among 640 CPC subclasses. Coverage is the share of citation links for which both patents can be assigned a subclass.

B.3 Estimation of the spatial weighting parameter in Test 2

A choice occasion is a firm-year $h = (f, t)$ in which the firm first enters at least one field and has at least two prior fields with positive filing stock and an observed market signal by $t-1$. I use historical filings, including those before 1976, to identify first entries. In the estimation reported in Table 2, the candidate set C_h excludes fields with both positive prior filing stock and an observed signal. A previously filed field still awaiting its first grant can therefore appear as a candidate with zero entry. Write $E_{hx} = 1$ if the firm first enters candidate field x in year t, and zero otherwise. The sample contains 38,248 entries across 16,742 firm-years. The number of destinations entered on a single occasion, $J_h = \sum_{x \in C_h} E_{hx}$, can exceed one.

For a trial value of ϕ, I construct the signal index $S_{hx}(\phi)$ and the knowledge index $N_{hx}(\phi)$ in equations (25) and (26) using the preceding decade's citation similarities. If a candidate has zero similarity to every prior field with an observed signal, I assign both indices a value of zero. Let $v_{hx}(\phi) = \beta S_{hx}(\phi) + \gamma N_{hx}(\phi) + \delta_x$, where δ_x is a candidate-field fixed effect. After profiling out the firm-year fixed effects, the Poisson log-likelihood, up to an additive constant, is

$$\ell(\beta, \gamma, \delta, \phi) = \sum_h \left[\sum_{x \in C_h} E_{hx} v_{hx}(\phi) - J_h \log \left(\sum_{x \in C_h} \exp\big(v_{hx}(\phi)\big) \right) \right]. \tag{B.1}$$

The firm-year fixed effect makes the fitted entries sum to J_h. The criterion thus compares where the firm enters, conditional on how many fields it enters that year.

I evaluate 21 logarithmically spaced values of ϕ from 0.05 to 32. At each value I recompute both indices and re-estimate the slopes and candidate-field fixed effects. I then refine the maximum over $\log \phi$ between the grid points on either side of the peak. The estimate is $\hat{\phi} = 0.482$. A prior field four times as similar to a candidate as another prior field receives approximately twice its weight, since $4^{0.482} = 1.95$.

I estimate uncertainty in $\hat{\phi}$ by resampling firms with replacement 100 times, retaining each selected firm's choice occasions and re-estimating ϕ in each draw. The bootstrap holds the patent-level first-stage residuals fixed. The bootstrap standard error is 0.034 and the 95% percentile interval is $[0.424, 0.542]$. Table 2 reports firm-clustered standard errors for the slopes conditional on $\hat{\phi}$. Its asymmetry specification holds ϕ fixed at the baseline estimate.