Small Bus Econ
https://doi.org/10.1007/s11187-023-00829-4

RESEARCH ARTICLE

Why generative AI can make creative destruction more creative but less destructive

Pehr-Johan Norbäck · Lars Persson

Accepted: 24 September 2023

Abstract The application of machine learning (ML) to operational data is becoming increasingly important with the rapid development of artificial intelligence (AI). We propose a model where incumbents have an initial advantage in ML technology and access to (historical) operational data. We show that the increased application of ML for operational data raises entrepreneurial barriers that make the creative destruction process less destructive (less business stealing) if entrepreneurs have only limited access to the incumbent's data. However, this situation induces entrepreneurs to take on more risk and to be more creative. Policies making data generally available may therefore be suboptimal. A complementary policy is one that supports entrepreneurs' access to ML, such as open source initiatives, since doing so would stimulate creative entrepreneurship.

Plain English Summary Why generative AI and big data may make the creative destruction process not only more creative but also less destructive. We show that entrepreneurs should consider that challenging incumbents in the era of ML and big data will be more difficult since incumbents' use of ML for proprietary data makes them more formidable competitors. This fact implies that entrepreneurs need to become riskier and more creative in the future to find a competitive edge. They may therefore seek support from venture capital to become more novel in their ventures. Data protection and privacy issues became a flashpoint in the media due in part to the high-profile exposure of Facebook users' data to Cambridge Analytica in 2016 and 2017. The rapidly expanding adoption of generative AI is seen as a risk of locking in the market dominance of large incumbent technology firms. Partially in response to these events, government regulators have instituted tighter rules on data protection. The results derived in this paper suggest that a complementary policy might be to support entrepreneurs' access to and knowledge of ML technology since doing so would stimulate creative entrepreneurship.

Keywords Machine learning · Big data · Generative AI · Open source · Creative destruction · Entrepreneurship · Operational data

JEL Classification L1 · L2 · M13 · O3

P.-J. Norbäck
Research Institute of Industrial Economics (IFN),
P.O. Box 55665, SE-102 15 Stockholm, Sweden
e-mail: pehr-johan.norback@ifn.se

L. Persson (✉)
Research Institute of Industrial Economics (IFN),
CEPR and CESifo, Stockholm, Sweden
e-mail: lars.persson@ifn.se

1 Introduction

Firms today often collect vast amounts of data through their regular activities, such as data on sales transactions and production processes, which we refer to as "operational data". While operational data have always

Published online: 16 November 2023

been of importance, with the introduction of machine learning (ML), they have become much more informative and important.[1,2] The use of ML on increasing amounts of operational data and unstructured unlabeled data will likely create significant efficiency gains for firms.[3]

However, the use of ML with increasing amounts of operational data is also likely to produce regulatory challenges since a fundamental feature of ML is that the more data there are available to train a system, the better the system becomes (see, e.g., Dutton 2018). The development of more efficient ML applications will create competitive advantages for incumbent firms due to their access to more operational data (see, e.g., Bessen 2018). The rapidly expanding adoption of generative AI is seen as risky in terms of locking in the market dominance of large incumbent technology firms. As Lina Khan, the Chairperson of the Federal Trade Commission (FTC), recently wrote in The New York Times, "A handful of powerful businesses control the necessary raw materials that startups and other companies rely on to develop and deploy AI tools.[4] This includes cloud services and computing power, as well as vast stores of data." This development may increase the barriers to entrepreneurship, with severe implications for the economy's dynamism.[5]

The purpose of this paper is to examine the impact of incumbent firms' application of ML to their (historical) operational data on entrepreneurial activity. The intensified rivalry between small entrepreneurial firms and incumbents using ML technologies is being increasingly observed.[6] As stated by Yoav Shoham, cofounder of the Israeli start-up AI21 Labs, "[t]he future will belong to smaller, specialist generative AI models that are cheaper to train, faster to run and serve a specific use case"[7]

In the model employed, firms use ML for operational data to improve its products and increase customers' willingness to pay.[8] As noted by Varian (2018), there are returns to scale from applications of ML to operational data whereby incumbents can gain a competitive advantage: classical returns to scale in production returns to scale due to demand-side network effects and learning by doing, which leads to quality improvements or cost decreases. Agrawal, Gans, and Goldfarb (2019b) show that a crucial feature of AI is its prediction capabilities, which will therefore have widespread consequences for the business sector. Like the steam engine, electrification, and the internet, AI is a general-purpose technology (GPT) that will significantly impact the whole business sector. We focus on the GPT and learning-by-doing aspect of ML applications for operational data by incorporating incumbent firms that employ ML on (previously) collected proprietary operational data and incoming

[1] Using a survey, Bughin et al. (2017) estimate that businesses—mainly large companies—spent $20–30 billion on AI development in 2016, while venture capital, private equity, and other external sources invested $6–9 billion.

[2] The application of ML took a giant leap with the introduction of ChatGPT in 2023, reaching 100 million users in just a couple of months. Generative AI chatbots are powered by foundation models, i.e., vast neural networks trained on large amounts of unstructured and unlabeled data in various formats, such as text and audio.

[3] In the European Commission's proposal for laying down harmonized rules for AI13, an artificial intelligence system (AI system) is defined as a software that is developed with ML, is logic and knowledge-based, and involves statistical approaches, which can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments with which they interact. However, the term "AI" has been mainly associated with ML algorithms or software containing one or several ML algorithms. (IPOL, 2021).

[4] OPINION, GUEST ESSAY, Lina Khan: We Must Regulate A.I. Here's How, May 3, 2023, The New York Times.

[5] See, e.g., Cohen (2010) for an overview of the research regarding how breakthrough innovations tend to come from smaller firms and startups rather than from large incumbent firms.

[6] Venture capital firms are betting that a fresh wave of generative AI startups, including Anthropic, Cohere, Stability AI, Inflection, and AI21 Labs, can move faster than can the more prominent companies, dominate select market niches, and perhaps ignore costly safety controls. See Opinion, Artificial intelligence, The likely winners of the generative AI gold rush, John Thornhill MAY 11 2023, Financial Times. https://www.ft.com/content/0cbe91ec-0971-4ba6-bdf1-87855aedd34c.

[7] Opinion, Artificial intelligence, The likely winners of the generative AI gold rush, John Thornhill MAY 11, 2023, Financial Times. https://www.ft.com/content/0cbe91ec-0971-4ba6-bdf1-87855aedd34c.

[8] The rapid development of generative AI has opened up many new ML applications in firms' operations. While ChatGPT has received the most attention among such applications, generative AI can improve performance across a broad range of content for firms, including images, video, audio, and computer code. Generative AI can perform several functions within organizations, including classifying, editing, summarizing, answering questions about, and drafting new content (McKinsey, 2023a, b).

sales data to improve their products and services. However, incumbents also face potential competition from entrepreneurial firms that can invest in the same ML (at a fixed cost) but with restricted data access to incumbents' data.

We assume that an entrepreneur needs to innovate successfully to compete with an incumbent firm that applies ML to its (previously) collected proprietary operational data. The entrepreneur chooses between different types of R&D projects, where a project with a lower probability of success is associated with higher efficiency and profitability if the project is successful. As expected, the more extensively the incumbent uses ML on its (previously) collected operational data, the more efficient the incumbent becomes and the higher the entrepreneurial barriers. Importantly, however, the increased barriers to entry will induce entrepreneurs to invest in R&D projects with higher risk and higher potential market value, the mechanism of which as follows: With access to more proprietary operational data, the incumbent becomes more aggressive in the product market, and—for a given project—entry becomes less profitable if the project succeeds. This effect induces the entrepreneur to switch to riskier R&D projects because succeeding with a mediocre project will bring about tiny profits for the entrepreneur since she will now face a stronger incumbent. This feature implies that it becomes profitable for the entrepreneur to take on more risk (and fail more often) in its R&D project since, when she succeeds, she will be sufficiently efficient to face competition from the stronger incumbent. Hence, the model predicts that the uptake of ML technologies will create advantages for incumbents due to their proprietary access to operational data, which will lead entrepreneurs not only to fail more often in their R&D projects but also to develop more transformative new products occasionally.

The appropriate regulatory response to this risk of market domination associated with ML is not easy to determine. Antitrust enforcement officials have already recognized that challenges may arise when large incumbent firms control most of the operational data. For example, FTC Commissioner Terrell McSweeny has noted that "it may be that an incumbent has significant advantages over new entrants when a firm has a database that would be difficult, costly, or time consuming for a new firm to match or replicate." In its new strategy for the digital industry, the European Union (EU) emphasizes the need to ensure that small and medium-sized businesses have adequate access to data and the competence required to implement ML. In the words of EU Commissioner Margrethe Vestager, "The real guarantee of an innovative future comes from keeping markets open so that anyone—big, small—can compete to produce the very best ideas" (Summit, Summit). In June 2022, the Bundeskartellamt initiated a proceeding against the technology company Apple to review its tracking rules and the App Tracking Transparency Framework under competition law. As Andreas Mundt, President of the Bundeskartellamt, stated, "We welcome business models that use data carefully and give users a choice as to how their data are used. A corporation like Apple, which can unilaterally set rules for its ecosystem, in particular for its app store, should make procompetitive rules. We have reason to doubt that this is the case when we see that Apple's rules apply to third parties but not to Apple itself. This would allow Apple to give preference to its own offers or impede those efforts of other companies. Our proceeding is largely based on the new competencies we received as part of the stricter abuse control rules regarding large digital companies, which were introduced last year (Section 19a German Competition Act - GWB). On this basis, we are conducting or have already concluded proceedings against Google/Alphabet, Meta/Facebook and Amazon."[9]

Himel and Seamans (2017) discuss how policymakers might address these issues and describe several policy solutions to consider, including provisions that would institute temporary data monopolies, data portability regimes, and the use of trusted third parties. A key feature of all these suggestions is that incumbents' monopoly access to operational data would be somewhat limited.

To capture this aspect in our model, we assume that the entrepreneurial firm can access a share of the incumbent's operational data to improve its products and increase customers' willingness to pay. Our analysis shows that policymakers should consider how these operational data policies affect not only the amount but also the quality of entrepreneurship. In particular, while policies that make operational data generally available

[9] The Bundeskartellamt. https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2022/14_06_2022_Apple.html.

stimulate the amount of entrepreneurship, this growth could come from entrepreneurs who take on too little risk. These findings suggest that entrepreneurship policies that reduce the cost of becoming an entrepreneur with access to ML technology, such as the support of open-source AI initiatives, might complement policies regarding access to incumbents' operational data.

2 Relation to the literature

This paper contributes to the literature on how the use of ML for big data may affect barriers to entry and entrepreneurship and its implications for intellectual property (IP) and antitrust policy (Bessen, 2018). Farboodi et al. (2019) propose a model where data accumulation increases the skewness of the firm size distribution, as large firms generate more data, but data-savvy small firms can overtake incumbents provided that they can finance their initial money-losing growth. Other scholars contend that operational data alone are unlikely to pose a barrier to entry (Lambrecht and Tucker, 2017; Sokol and Comerford, 2016). Bajari et al. (2019) find that increasing the number of online products that Amazon tracks does not significantly improve ML prediction accuracy after a certain point, suggesting that data quantity may function as only a low-level barrier to entry. We add to this literature by examining the quality of the products or processes with which entry occurs, as well as the likelihood of entry. This qualitative aspect is of fundamental importance since the benefits of industrial restructuring depend not only on the pace at which firms are replaced but also on the nature of the novel products or processes. In particular, we show that the application of ML for incumbents' previously collected proprietary data increases the barrier to entrepreneurship because incumbents' competitive advantage increases due to oligopolistic strategic effects. However, we also show that the increased use of ML by incumbents increases entrepreneurs' willingness to take on risk and lengthens the technological jumps that entrepreneurs provide to society.

There is recent literature examining how the implementation of AI affects firms' decisions under uncertainty. Agrawal et al. (2018); Agrawal et al. (2019b) show that better predictions from firms' implementation of AI will have widespread consequences since predictions are fundamental to decision-making in firms. Gans (2023) proposes a model where the implementation of AI implies a better prediction of demand, allowing firms to match decisions, such as those related to output and employment, with the predicted state. While output might be relatively stable when there is no prediction, the availability of a prediction may cause firms to increase or reduce output accordingly. These better predictions represent an improvement in efficiency. Nevertheless, some of the efficiency gains come from reducing output, which implies that the external effect of AI adoption on other firms is positive rather than negative. Agrawal et al. (2019a) propose a decision-making model under uncertainty where the implementation of AI improves prediction about uncertain states of the world. The above authors show that having more accurate predictions leads to better decisions. Moreover, more accurate prediction makes firms make more risky decisions because it makes risky action less risky. We add to this literature by showing that while AI may, through its prediction capabilities, reduce risk in firms' decision-making, it may also trigger entrepreneurs to seek business ventures with inherently higher risk to overcome the advantage incumbents have due to their proprietary data.

This paper also contributes to the literature on the effects of privacy, data protection policy, and competition (see, for instance, Acquisti et al. 2016). Jia et al. (2021) find that the EU General Data Protection Regime (GDPR) might constitute a barrier to entry for startups. Campbell et al. (2015) propose a model of how regulatory attempts to protect consumers' data privacy affect the structure of competition and find that the consent-based approach may disproportionately benefit firms that offer a larger scope of services, thus most adversely affecting small and new firms. This prediction has also been supported empirically in recent work on the EU GDPR (Batikas et al., 2020; Johnson et al., 2022). What has not been examined, however, is how such regulations affect the quality of the products and services offered by entrant firms. We add to this stream of literature by proposing a model where machine-learning-by-doing mechanisms with entrepreneurial innovations are central. This approach enables us to show that policies designed to reduce incumbents' advantages from using ML on previously collected proprietary data can stimulate the degree of

entrepreneurship but that this can result in too little risk taking from a social point of view.[10]

3 A Primer: generative AI and firms' business models

Here, we briefly describe how generative AI may be used in firms' businesses. The text is based on the description in McKinsey (2023a, b, c). The generative AI tool ChatGPT reached 100 million users within just 2 months of its release and has given rise to many applications. The underlying technology enabling generative AI is a class of artificial neural networks called foundation models, which are trained using deep learning, alluding to the many (deep) layers within neural networks. Deep learning has powered many of the recent advances in AI. However, some characteristics set foundation models apart from previous generations of deep learning models; such foundation models can be trained on vast and varied sets of unstructured data. For example, a foundation model called a large language model can be trained on vast amounts of text that is publicly available on the internet and covers many different topics. Foundation models amass these capabilities by learning patterns and relationships from the broad training data they ingest, enabling them to predict the next word in a sentence.

[10] Our paper also adds to the literature on firm asymmetries and risk behavior in R&D processes. Rosen (1991) and Cabral (2003) show that small firms may have an incentive to choose risky strategies due to strategic output effects in the product market, i.e., small firms do not take on low-risk–low-return projects since they cannot exploit large output improvements. Färnstrand Damsgaard et al. (2017) show that entrepreneurial firms may choose riskier strategies because, unlike incumbents, they would not have already sunk a large share of their entry (commercialization) costs before the outcome of an R&D process is determined. Moreover, Haufler et al. (2014) study the effects of tax policies on entrepreneurs' choice of the riskiness (or quality) of an innovation project. It is shown that limited-loss-offset provisions in the tax system induce entrepreneurs innovating for entry to choose projects with inefficiently low risk but that the same distortion does not arise when entrepreneurs sell their innovation in a competitive bidding process. Henkel et al. (2015) show that independent entrepreneurs who innovate for sales choose riskier R&D projects than do incumbents since the latter has an incentive to opt for safer R&D projects to improve its bargaining power in subsequent acquisitions. We add to this literature by pointing out that the development of ML and the buildup of incumbent proprietary data induce entrepreneurs to take on more risk but that those policies that make operational data generally available may be suboptimal. The reason for this is that it can reduce entrepreneurs' willingness to take on risk.

How generative AI improves the quality of firms' products and reduces their costs Generative AI can speed up, scale, or improve business practices. A specially trained AI model could suggest upselling opportunities to a salesperson. Nevertheless, until now, such studies have usually been based only on static customer data obtained before the start of the call, such as demographics and purchasing patterns. However, a generative AI tool might also suggest upselling opportunities to the salesperson in real time based on the actual content of the conversation, drawing from internal customer data, external market trends, and social media influencer data. In addition, generative AI could offer a first draft of a sales pitch for the salesperson to adapt and personalize.

While generative AI may eventually be used to automate some tasks, much of its value could derive from how software vendors embed the technology into the everyday tools (for example, email or word processing software) used by knowledge workers. Such upgraded tools could substantially increase productivity. Generative AI can enable capabilities across a broad range of content, including images, video, audio, and computer code, and it can perform several functions in organizations, including classifying, editing, summarizing, answering questions, and drafting new content. Each of these actions has the potential to create value by changing how work is carried out at the activity level across business functions and workflows.

Entrepreneurial opportunities in generative AI A value chain is emerging to support the training and use of generative AI. Specialized hardware provides the extensive computer power needed to train the models. Cloud platforms offer the ability to tap this hardware. MLOps and model hub providers deliver the tools, technologies, and practices that an organization needs to adapt and deploy a foundation model within its end-user applications. Many companies are entering the market to offer applications built on top of foundation models that enable them to perform a specific task, such as helping a company's customers with service issues.

Incumbent advantage in generative AI business models The first foundation models required high levels of investment to develop, given the substantial computational resources needed to train them and the human effort needed to refine them. As a result, they were developed primarily by a few tech giants, startups backed by significant investment, and some open-

source research collectives (for example, BigScience). However, work is underway on smaller models that can deliver effective results for some tasks and more efficient training. This development could eventually open the market to more entrants. Some startups have already succeeded in developing their own models—for example, Cohere, Anthropic, and AI21 Labs build and train their own large language models.

4 Model

To examine the effects of generative AI on entrepreneurs' incentives to innovate and enter existing markets, we develop a framework in which firms use ML for operational data to improve their products and increase customers' willingness to pay. As described in Sect. 3, a vital feature of the application of generative AI is that it increases consumer satisfaction and, thereby, consumers' willingness to pay. (See McKinsey (2023a) for examples of how firms use generative AI to increase customers' willingness to pay). The model combines active learning-by-doing mechanisms (see Thompson 2010) with entrepreneurial innovation mechanisms, as modeled in Färnstrand Damsgaard et al. (2017). In this framework, an incumbent firm obtains an advantage from being able to employ ML on previously collected proprietary data and incoming sales data to increase consumers' willingness to pay. We refer to such data as operational data. The incumbent firm also faces potential competition from an entrepreneurial firm that can invest in research to develop new products and use ML to obtain access to operational data to increase consumers' willingness to pay. Note that we could have cast the model so that ML reduced the production cost for firms, reaching similar results.

The new generative AI chatbots use, to a large extent, mainly public data. However, as described in the previous section, these new AI methods will be applied to both firm-specific operational and publicly available data. Indeed, policy makers worry that the rapidly expanding adoption of generative AI risks further locking in the market dominance of large incumbent technology firms, as expressed by Lina Khan, the chair of the Federal Trade Commission (see the description in the "Introduction"). To focus the analysis on the potential anticompetitive problems of ML applied to proprietary operational data, we assume that the incumbent and entrepreneur have access to the same amount of public data and assume this amount to be zero. We also assume that the entrepreneur faces an extra fixed cost, F, to learn ML, and if it takes such a cost, then it will reach the same ML knowledge level as that of the incumbent.

The setting is as follows:

- In Stage 1, the firm faces a fixed cost F to learn ML, and if it takes such a cost, then it will reach the same ML knowledge level as that of the incumbent.
- In Stage 2, the entrepreneur can invest in an R&D project that—if successful—will generate an invention. This invention can take several forms, all of which increase the profits of its owner. The invention can be a new product, a product of higher quality, or a new or improved production process. For simplicity, we assume that the invention is a product innovation that increases consumers' willingness to pay. The entrepreneur chooses among an infinite number of independent R&D projects. There is a cost of running a project, and to capture this cost, we assume that the entrepreneur can undertake only one project at a time.[11] Along the technological frontier, the entrepreneur thus faces a choice between projects that have a high probability of success but deliver a small increase in willingness to pay in case of success and projects that are riskier but also feature a higher increase in willingness to pay if successful. At the end of Stage 2, the outcome of the entrepreneur's R&D project is revealed, where the entrepreneur stays in the market if she is successful and exits otherwise.
- In the final stage, Stage 3, product market interaction takes place, where, for simplicity, competition is modeled as Cournot competition (in differentiated goods or services). The product market profits depend on whether the entrepreneur succeeds with her R&D project and on the type of project undertaken. The key to the model is that both firms use ML in Stage 3 to process sales information to increase consumers' willingness to pay, but the incumbent has an advantage in the form of access to operational data.

[11] See Gilbert (2006) for our motivation.

In what follows, we analyze the equilibrium of the proposed game, following the usual backward-induction procedure.

4.1 Stage 3: product market

4.1.1 Consumers

Consumers have quasilinear quadratic utility and solve the following utility maximization problem:

$$\underset{\{q_E,q_I,q_0\}}{Max} : U = u(q_E, q_I) + q_0 \tag{1}$$

$$s.t : u(q_E, q_I) = a_E \cdot q_E + a_I \cdot q_I - \frac{1}{2} \cdot \left[q_E^2 + q_I^2\right] - q_E \cdot q_I \tag{2}$$

$$s.t : P_E \cdot q_E + P_I \cdot q_I + q_0 = m, \tag{3}$$

where q_E represents the quantity consumed of the entrepreneur's good, q_I is the quantity consumed of the incumbent's good, and q_0 is the quantity consumed of a numeraire good—or outside good. The subutility function $u(q_E, q_I)$ in Eq. 2 over the goods of the entrepreneur and the incumbent is linear quadratic.[12] The consumer budget set as given in Eq. 3, where m is exogenous consumer income, P_E is the price of the entrepreneur's product, and P_I is the price of the incumbent's product. The price of the outside good is normalized to one.

Solving for the amount of the outside good q_0 from the budget constraint Eq. 3 and substituting the quadratic utility $u(q_E, q_I)$ in Eq. 2 into the direct utility in Eq. 1, we can rewrite direct utility as follows:

$$U = [a_E - P_E] \cdot q_E + [a_I - P_I] \cdot q_I - \frac{1}{2} \cdot \left[q_E^2 + q_I^2\right] - q_E \cdot q_I + m. \tag{4}$$

Taking the first-order condition for consumer maximization, $\frac{\partial U}{\partial q_i} = 0$ for $i = E, I$, we obtain the residual demand facing each firm as follows:

$$\frac{\partial u}{\partial q_i} = P_i = a_i - q_i - q_j, \text{ for } i, j = \{E, I\}, i \neq j. \tag{5}$$

From Eq. 5, consumers' willingness to pay for a firm's product $\frac{\partial u}{\partial q_i}$ is shown to be decreasing in the firm's own output q_i and in the rival's output q_j. We now assume that consumers' willingness to pay, as measured by the intercept a_i, can be affected by firms' use of ML.[13]

4.1.2 Residual demand for the incumbent's product and ML

Firms use ML and operational data to increase consumers' willingness to pay. As discussed in detail in Sect. 3, generative AI can be used to increase consumer satisfaction and, thereby, consumers' willingness to pay. Indeed, McKinsey (2023a) provides numerous examples of how firms use generative AI to increase customers' willingness to pay. In the model, this is captured by ML on operational data affecting the demand intercept a_i in Eq. 5 in two distinct ways. For the incumbent firm, the demand intercept is given as

$$a_I = a + \underbrace{\alpha \cdot d_I}_{\text{ML on old data}} + \underbrace{\alpha \cdot q_I}_{\text{ML on new data}}, \tag{6}$$

where a is the part of consumers' willingness to pay that is unaffected by ML.

- The incumbent uses historical customer data d_I (available from previous sales) to increase consumers' willingness to pay. We may also think of these as data as customer data from locked-in consumers in a switching cost type of model. This scenario is illustrated in the upper panel in Fig. 1, starting with the case where the incumbent is a monopolist. Applying ML to preexisting data results in an upward shift of the demand intercept from a to $a + \alpha \cdot d_I$, where we can think of the parameter α as indicating the productivity level of ML (in using data to increase consumers' willingness to pay), which is a function of the (exogenous) state of computer technology.

[12] For a review of quasilinear quadratic utility models, see Choné and Linnemer (2019).

[13] The assumption of linear demand is made for ease of exposition since we can then solve the model analytically. We can introduce second-order effects using quadratic terms in the demand equation. Such an extension would maintain our derived results as long as these second-order effects are sufficiently small. However, allowing for more significant second-order effects appears to be an interesting avenue for future research.

- The incumbent can also use the information on contemporaneous sales, q_I, to increase willingness to pay, where we assume that consumers' willingness to pay also increases at the rate α. This result is also shown in the upper panel in Fig. 1, where ML applied to data on contemporaneous sales makes demand more elastic, shifting the demand curve $a - q_I$ to $a - q_I + \alpha \cdot q_I$. An example of such a situation would be the information gathered from the road mileage and driving patterns of buyers of self-driving cars, where performance and safety in new cars increase from advanced learning with the generation of new data. Here, we thus take the shortcut of assuming that the incumbent learns directly from its Stage 3 sales, similar to the mechanism in the learning-by-doing literature.[14] We also assume that each consumer does not internalize the information that she gives firms with her purchase, captured by the term $\alpha \cdot q_I$ in her consumption choice. This situation does not seem to be at odds with reality, as it seems notoriously difficult for consumers to reap any benefits from information sharing.

Combining these two ML channels, Fig. 1(i) depicts the inverse demand for the incumbent's product when the incumbent faces no competition from the entreprene ur, $P_I^M = a + [\alpha \cdot d_I + \alpha \cdot q_I] - q_I$. In Fig. 1(ii), we then depict the incumbent's residual demand—the demand facing the incumbent when q_E units are supplied by the entrepreneur, $P_I = P_I^M - q_E$, or

$$P_I = P_I^M - q_E = a + [\alpha \cdot d_I + \alpha \cdot q_I] - q_I - q_E. \quad (7)$$

4.1.3 Residual demand for the entrepreneur's product

The demand for the entrepreneur's product is more involved since consumers' willingness to pay for the entrepreneur's product depends on whether the innovation project is successful.

Entrepreneur succeeds with her innovation If the entrepreneur succeeds with her innovation, then her demand intercept is given by

$$a_E|_{Succeed} = a + \underbrace{\gamma \cdot \alpha \cdot d_I}_{\text{ML on incumbent's old data}} + \underbrace{\alpha \cdot q_E}_{\text{ML on new data}} + \underbrace{[b - \beta \cdot \rho_E]}_{\text{Innovation succeeds}} \quad (8)$$

The entrepreneur can also use information from consumers' purchases of her product and apply ML to make the product more attractive, increasing consumers' willingness to pay. This scenario is shown in Fig. 2(i), where ML applied to contemporaneous sales shifts the demand curve without ML, $a - q_E$, to the demand curve under ML, $a - q_E + \alpha q_E$. If the entrepreneur has access to the incumbent's historical data, d_I (whether through a general agreement or by law or regulation), then the entrepreneur can also use this data source to increase consumer's willingness to pay.

In what follows, we shall assume that the entrepreneur is disadvantaged by not having access to her own historical data, i.e., that $d_E = 0$, and by having inferior access to the incumbent's data, where $\gamma \in [0, 1)$ captures the share of the incumbent's data to which the entrepreneur has access. We then define the incumbent's data advantage as follows:

Definition 1 *The incumbent has privileged data access as follows:* $d_I > 0 = d_E$ *and* $\gamma \in [0, 1)$.

While the entrepreneur has an inherent disadvantage in having weaker access to historical data, she can compensate for this by succeeding with her innovation. Adding new features to her product increases consumers' willingness to pay by $b - \beta\rho_E \geq 0$, where $\beta \in [0, b)$ and $\rho_E \in [0, 1]$ is the probability that the project succeeds. Note how consumers' willingness to pay for the entrepreneur's product is higher if she has taken greater risk in her research project, i.e., if she has succeeded with a project with a lower probability of success ρ_E. That is,

$$\frac{\partial a_E}{\partial \rho_E} \cdot d\rho_E = \underset{(-)}{-\beta} \cdot \underset{(-)}{d\rho_E} > 0. \quad (9)$$

This situation reflects a natural tradeoff, where *riskier projects* have a *greater value for consumers if they succeed compared to less risky projects* . We turn to project choice in more detail in the next section.

The upward shift of the demand intercept from a to $a + [b - \beta \cdot \rho_E] + \gamma \cdot \alpha \cdot d_I$ in the upper panel in

[14] An alternative approach would be to assume that ML for old (incumbent) data is less effective. In such a setup, the disadvantage for the entrepreneurial firm would be reduced, and there would be less incentive for the entrepreneurial firm to choose riskier projects. However, our results would be qualitatively the same as long as firms can apply ML to old data with some level of efficiency.

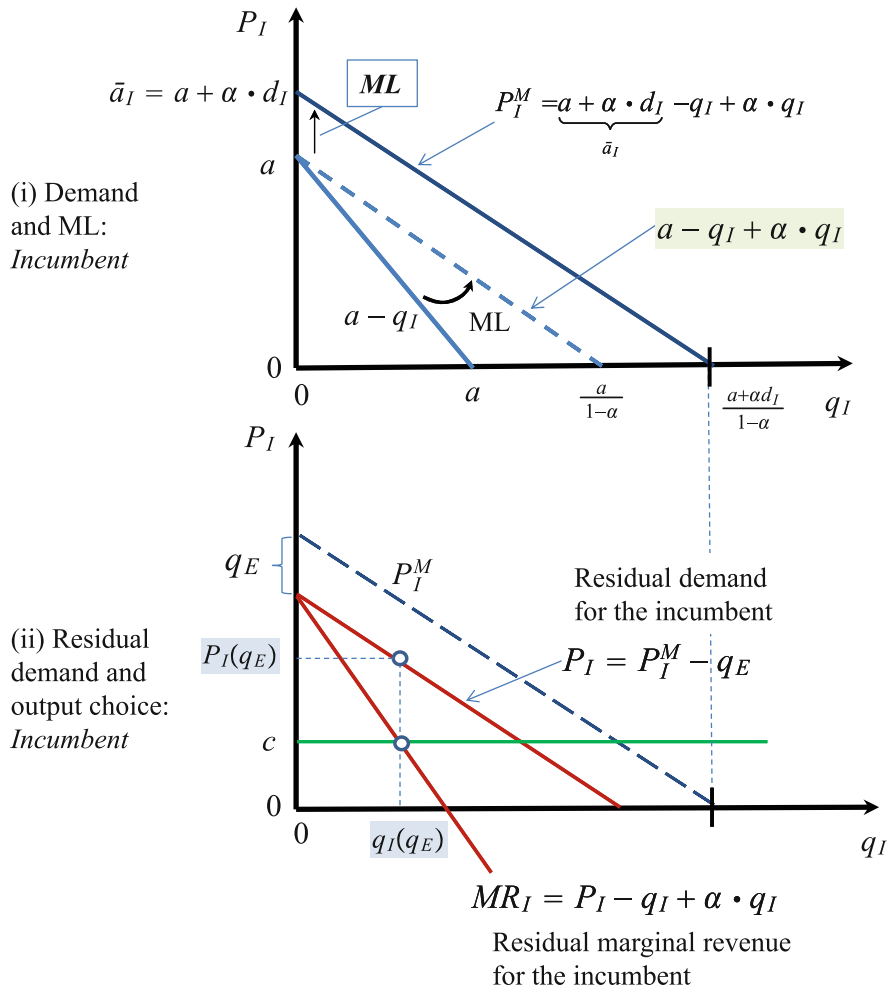

Fig. 1 Panel (**i**) shows how ML affects consumers' willingness to pay (WTP) for the incumbent's product and the incumbent's inverse demand in the absence of competition from the entrepreneur. Panel (**ii**) then derives the incumbent's residual demand and its residual marginal revenue and illustrates its profit-maximizing output choice

Fig. 2 illustrates how the entrepreneur can increase consumers' willingness to pay after succeeding with her innovation project by using ML with limited access to the incumbent's data d_I. When ML for contemporaneous sales data is considered, the inverse demand for the entrepreneur *without* competition from the incumbent $P_E^M = a + [b - \beta \cdot \rho_E] + \gamma \cdot \alpha \cdot d_I + \alpha \cdot q_E - q_E$ is drawn in Fig. 2(i). In Fig. 2(ii), we depict the entrepreneur's residual demand—the demand facing the incumbent when q_I units are supplied by the incumbent—that is,

$$P_E = P_E^M - q_I = a + [(b - \beta \cdot \rho_E) + \gamma \cdot \alpha \cdot d_I + \alpha \cdot q_E] - q_E - q_I. \quad (10)$$

Entrepreneur fails with her innovation If the entrepreneur fails with her innovation, then she is left out of the increase in willingness to pay $b - \beta \cdot \rho_E$ in $a_E|_{Succeed}$ shown in Eq. 8. We shall assume that consumers' willingness to pay for the entrepreneur's product $a_E|_{Fail} = a_E|_{Succeed} - [b - \beta \cdot \rho_E]$ is too low to secure profitable entry into the product market if she fails. This feature can be formalized in several ways. One way is to introduce a sufficiently high entry cost, f, for the entrepreneur before the quantity competition in Stage 2 takes place. Including such an entry cost in the analysis will not affect our results qualitatively, but computations can become more involved. Alternatively, we could assume that failing—in addition to the loss of the increase in willingness to pay $b - \beta \cdot \rho_E$—is also associated with a loss of confidence in the entrepreneur's product with the cost f effectively becoming a penalty for failing: $a_E|_{Fail} = a_E|_{Succeed} - [b - \beta \cdot \rho_E] - f < 0$. A possible third approach would be to focus the analyses on outcomes where the incumbent's data advan-

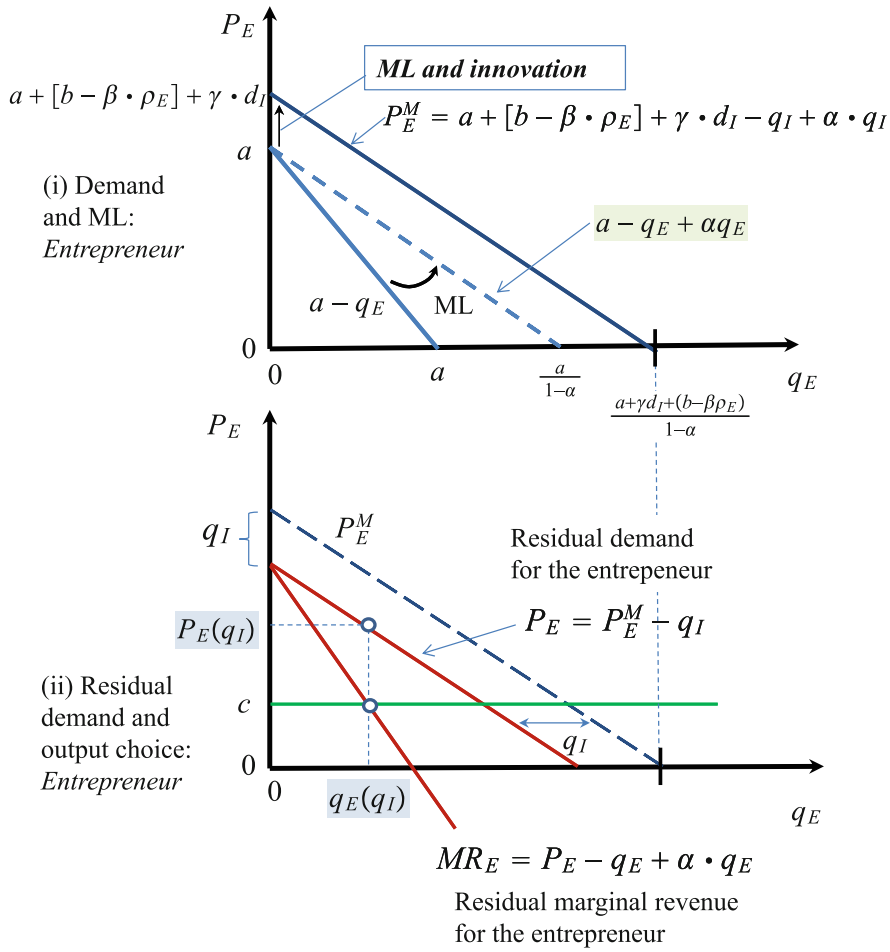

Fig. 2 Panel (**i**) shows how ML affects consumers' WTP for the entrepreneur's product given access to the incumbents data and depicts its inverse demand in the absence of competition from the incumbent. Panel (**i**) also illustrates that a successful innovation project increases consumers' WTP—even more so if she has taken greater risk in her research project. Panel (**ii**) derives the entrepreneur's residual demand and its residual marginal revenue and illustrates its profit maximizing output choice

tage is so significant that the entrepreneur would optimally choose a zero output level in the Cournot competition if she fails with her innovation.

For ease of exposition—but with no loss of generality —we shall assume that the entrepreneur does not enter the product market if she fails with the invention. Thus, when the entrepreneur fails with her innovation, the incumbent is a monopolist in the market with inverse demand $P_I^M = a + [\alpha \cdot d_I + \alpha \cdot q_I] - q_I$.

4.1.4 Optimal quantity

The profit maximization problem of firm i is

$$\max_{\{q_i\}} \pi_i = [P_i - c]\, q_i, \; i = \{I, E\}, \tag{11}$$

where each firm's price P_i is given from the residual demand functions Eqs. 7 and 10 and where we assume that each firm faces a constant marginal cost, c. The first-order conditions $\frac{\partial \pi_i}{\partial q_i} = 0$ imply that each firm chooses its output such that marginal revenue equals marginal cost or such that

$$\underbrace{P_i - (1-\alpha)\, q_i^*}_{MR_i} = \underbrace{c}_{MC_i}, \; i = \{I, E\}, \tag{12}$$

where $dP_i/dq_i = -(1-\alpha)$ from Eqs. 7 and 10 and where $\frac{da_i}{dq_i} = \alpha$ from Eqs. 6 and 8, with the latter expressions capturing how ML increases consumers' willingness to pay from information on contemporaneous sales.[15] These first-order conditions (giving each

[15] The second-order condition is fulfilled since $\frac{\partial^2 \pi_i}{\partial q_i^2} = 2(1-\alpha) < 0$.

firm's best response to its rival) are also illustrated in the lower panels of Figs. 1 and 2.

4.1.5 Nash–Cournot equilibrium

To derive the Nash–Cournot equilibrium, it is useful to derive firms' reaction functions. Using Eqs. 6 and 8 and defining $\Lambda = a - c$, we define consumers' net willingness to pay for each firm's product $\bar{\Lambda}_i$ as

$$\bar{\Lambda}_I(\Lambda, \alpha, d_I) = \Lambda + \alpha \cdot d_I, \tag{13}$$

$$\bar{\Lambda}_E(\Lambda, b, \beta, \rho_E, \alpha, d_I, \gamma) = \Lambda + [b - \beta \cdot \rho_E] + \gamma \cdot \alpha \cdot d_I. \tag{14}$$

As shown, consumers' net willingness to pay increases when ML techniques become more efficient due to the availability of better computers, i.e., when α increases. For a given computer technology, applying ML to more data allows firms to better infer consumer preferences and further increase consumers' willingness to pay. However, since the entrepreneur does not have full access to the incumbent's data, $\gamma \in [0, 1)$, the increase in net willingness to pay is smaller for the entrant than for the incumbent: $\partial \bar{\Lambda}_I(\cdot)/\partial d_I = \alpha > \gamma\alpha = \partial \bar{\Lambda}_E(\cdot)/\partial d_I$. Again, this can be compensated for if the entrepreneur succeeds with her innovation, in which case the net willingness to pay for the entrepreneur's product $\bar{\Lambda}_E(\cdot)$ rises with the features of the new product, $b - \beta \cdot \rho_E > 0$, with the increase in willingness to pay endogenously determined by project choice ρ_E.

From Eqs. 7–14, we can derive firms' reaction functions as follows:

$$R_i(q_j) = \frac{\bar{\Lambda}_i(\cdot) - q_j}{2(1-\alpha)}, I, j = \{E, I\}, i \neq j. \tag{15}$$

The reaction function for firm i, $R_i(q_j)$, gives the optimal output choice q_i for a given choice of output by firm, j, q_j. The reaction function of the incumbent $R_I(q_E) = \frac{\bar{\Lambda}_I(\cdot) - q_E}{2(1-\alpha)}$ is depicted as the downward-sloping dark blue curve in Fig. 3(i). The downward slope captures the fact that firms' quantities are strategic substitutes: if the incumbent believes that the entrepreneur will produce more, then she will expect a lower price for her product and—as shown in Eq. 12—lower marginal revenue, which will induce her to produce less to make marginal revenue equal to marginal cost. Since Fig. 3 is drawn with the output of the incumbent on the vertical axis and the output of the entrepreneur on the x-axis, we use the inverse reaction function of the entrepreneur, $R_E^{-1}(q_E) = \bar{\Lambda}_E(\cdot) - 2(1-\alpha) q_E$. The reaction function of the entrepreneur is yet again downward sloping, displaying the fact that quantities are strategic substitutes: if the entrepreneur expects the incumbent to produce a high level of output, then she will expect a low price for her product and lower marginal revenue, which will induce her to choose a lower output.

Nash–Cournot equilibrium is reached when both firms choose the optimal output and correctly infer the output choice of their rival, i.e., the Nash–Cournot equilibrium is given from the intersection of the reaction functions at point N in Fig. 3(i). It is straightforward to verify that the Nash–Cournot equilibrium is

$$q_I^* = \frac{2(1-\alpha)\Lambda_I - \Lambda_E}{(1-2\alpha)(3-2\alpha)} = \frac{(1-2\alpha)\Lambda - (b - \beta\rho_E) + (2(1-\alpha) - \gamma) d_I \alpha}{(1-2\alpha)(3-2\alpha)} \tag{16}$$

$$q_E^* = \frac{2(1-\alpha)\Lambda_E - \Lambda_I}{(1-2\alpha)(3-2\alpha)} = \frac{(1-2\alpha)\Lambda + 2(b - \beta\rho_E)(1-\alpha) - (1 - 2(1-\alpha)\gamma) d_I \alpha}{(1-2\alpha)(3-2\alpha)}, \tag{17}$$

where $1 - 2\alpha > 0$ ensures the stability of the equilibrium.[16] Stability, i.e., $\alpha \in [0, 1/2)$, ensures that $3 - 2\alpha > 0$ and $2(1-\alpha) - \gamma > 0$.

4.1.6 Data availability and product market outcome

Our main interest lies in exploring how the equilibrium in the product market and the entrepreneur's incentives to innovate are affected by access to data and the use of ML. Let us first explore how the amount of historical data in the hand of the incumbent d_I affects the Nash equilibrium in Eqs. 16 and 17 for a *given project choice* of the entrepreneur ρ_E.

To proceed, we make use of the following definition:

Definition 2 *The entrepreneur has (i) strong access to the incumbent's historical data d_I if and only if $1 - 2(1-\alpha)\gamma < 0$ and (ii) weak access to the incumbent's historical data d_I if and only if $1 - 2(1-\alpha)\gamma > 0$.*

We use this definition to describe how the incumbent's amount of historical data affects the equilibrium behavior of the entrepreneur and, in particular, how the

[16] This approach ensures that the reaction function of the entrant is steeper than that of the incumbent.

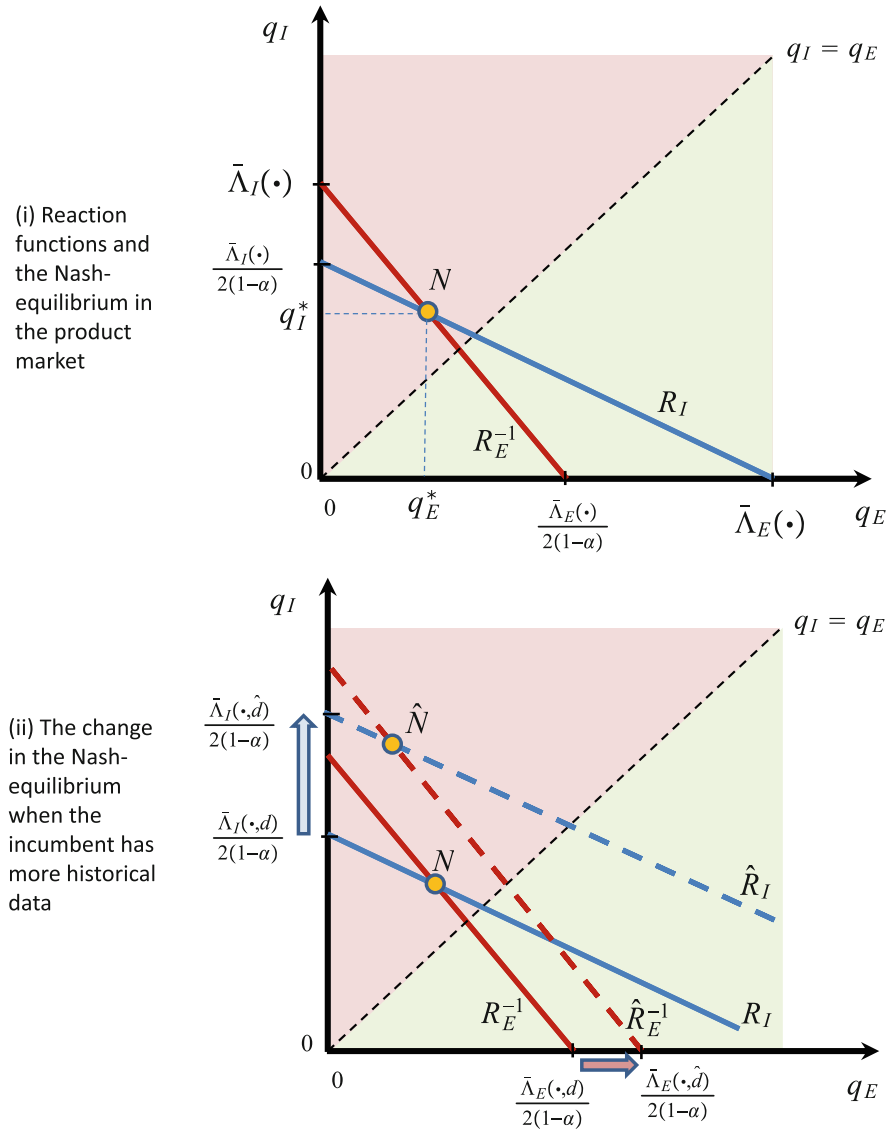

Fig. 3 Illustrating the Nash equilibrium in the product market for a given (successful) innovation project chosen by the entrepreneur. Panel (**i**) illustrates the initial Nash equilibrium with the incumbent assumed to be the larger firm. Panel (**ii**) illustrates the shift in the Nash equilibrium toward reinforced incumbent domination (under weak access to the incumbent's historical data for the entrepreneur)

entrepreneur's output choice is affected by the incumbent having more historical data.

Note that when the strong access condition is fulfilled, $(1-2(1-\alpha)\gamma > 0$, it implies that

$$\gamma > \tilde{\gamma} = \frac{1}{2(1-\alpha)}, \tag{18}$$

where $\tilde{\gamma} \in [0.5, 1)$ since stability requires $\in [0, 1/2]$.

Intuitively, the strong access condition holds when the entrepreneur has sufficient access to the incumbent's historical data, i.e., when $\gamma > \tilde{\gamma}$. From Eq. 18, we also note that

$$\frac{d\tilde{\gamma}}{d\alpha} = \frac{1}{2(1-\alpha)^2} > 0. \tag{19}$$

Thus, to ensure that the strong access condition is fulfilled, $(1-2(1-\alpha)\gamma > 0$, when the efficiency in ML, α, increases, the entrepreneur needs better access to the incumbent's data, i.e., γ needs to increase. Intuitively, more efficient ML will increase the incumbent's data advantage, and better access for the entrepreneur is then needed to compensate for such an increase.

We explore how the entrepreneur's behavior is affected by her access to the incumbent's historical data, γ, in more detail in Sect. 4.3. The Appendix A also illustrates the impact on the entrepreneur's behavior when the effectiveness of ML, α, varies.

We can derive the following results:

Proposition 1 *When the incumbent's amount of historical data d_I increases,*

(i) The incumbent always expands its output $\frac{\partial q_I^*}{\partial d_I} > 0$;

(ii) The entrant expands her output only when she has strong access to the incumbent's historical data d_I: $\frac{\partial q_E^*}{\partial d_I} > 0$ iff $1-2(1-\alpha)\gamma < 0$ and $\frac{\partial q_E^*}{\partial d_I} < 0$ iff $1-2(1-\alpha)\gamma > 0$; and

(iii) The incumbent always expands its output more than does the entrepreneur: $\frac{\partial q_I^*}{\partial d_I} > \frac{\partial q_E^*}{\partial d_I} > 0$.

To prove parts (i) and (ii), we partially differentiate Eqs. 16 and 17 to obtain the following:

$$\frac{\partial q_I^*}{\partial d_I} = \frac{2(1-\alpha)-\gamma}{(1-2\alpha)(3-2\alpha)}\alpha > 0, \tag{20}$$

$$\frac{\partial q_E^*}{\partial d_I} = \begin{cases} -\frac{1-2(1-\alpha)\gamma}{(1-2\alpha)(3-2\alpha)}\alpha > 0, \text{ for } 1-2(1-\alpha)\gamma < 0, \\ -\frac{1-2(1-\alpha)\gamma}{(1-2\alpha)(3-2\alpha)}\alpha < 0, \text{ for } 1-2(1-\alpha)\gamma > 0. \end{cases} \tag{21}$$

where, again, stability, i.e., $\alpha \in [0, 1/2)$, ensures that $2(1-\alpha)-\gamma > 0$.

As shown in Eq. 20, the incumbent strictly increases its output with access to more data, $\frac{\partial q_I^*}{\partial d_I} > 0$true for the entrepreneur when she has strong access to the incumbent's historical data, i.e., $\frac{\partial q_E^*}{\partial d} > 0$ if $\gamma \in (\frac{1}{2(1-\alpha)}, 1]$.[17] However, as shown by the lower line, in Eq. 21, if the entrepreneur has weak access to the incumbent's historical data, $\gamma \in [0, \frac{1}{2(1-\alpha)}]$larger amounts of historical data, $\frac{\partial q_E^*}{\partial d_I} < 0$. Figure 3(ii) provides an illustration of the interaction in the latter case: increases in the amount of historical data held by the incumbent d_I and both firms' (differential) application of ML to these data—and to new data from contemporaneous sales—induce both firms to increase their sales as consumers' willingness to pay increases. Hence, both firms' reaction functions shift outward. However, with access to the incumbent's historical data being suppressed, the entrepreneur's reaction function shifts outward less than does that of the incumbent, and the incumbent reinforces her market dominance.

To prove part (iii), we first note that when the entrepreneur has weak access to the incumbent's historical data, i.e., when $1-2(1-\alpha)\gamma > 0$, it immediately follows that $\frac{\partial q_I^*}{\partial d_I} - \frac{\partial q_E^*}{\partial d_I} > 0$. When the entrepreneur has strong access to the incumbent's historical data, i.e., $1-2(1-\alpha)\gamma < 0$, Eqs. 20 and 21 directly imply $\frac{\partial q_I^*}{\partial d_I} - \frac{\partial q_E^*}{\partial d_I} = \alpha\frac{1-\gamma}{1-2\alpha} > 0$.

However, the amount of historical data d_I held by the incumbent does not only affect the product market equilibrium—the amount of data and access to it by the entrepreneur also affect the entrepreneur's innovation incentives through its effects on the entrepreneur's project choice, ρ_E. This innovation channel—which we have ignored thus far—is the subject of the next section.

4.2 Stage 2: R&D by the entrepreneur

In this stage, the entrepreneur decides on her optimal R&D project. Using the direct profit function Eq. 11, the residual demand Eq. 10, the net willingness to pay Eq. 14, and the Nash quantity in Eq. 17, we can write the reduced-form product market profit for the entrepreneur as follows:

$$\pi_E(\rho_E) = \left(\underbrace{\bar{\Lambda}_E(\cdot, \rho_E) + \alpha \cdot q_E^*(\rho_E) - q_E^*(\rho_E) - q_I^*(\rho_E)}_{P_E - c} \right) \times q_E^*(\rho_E). \tag{22}$$

By assumption, the entrepreneur enters the market only if the selected R&D project is successful in Stage 1.[18] This outcome occurs with probability ρ_E and generates net profit $\pi_E^*(\rho_E)$ for the entrepreneur. The entrepreneur's expected profit is therefore given as follows:

$$\underset{\{\rho_E\}}{Max} : \mathrm{E}[\Pi_E] = \rho_E \times \pi_E(\rho_E), \tag{23}$$

$$s.t : \rho_E \in [0, 1], \tag{24}$$

[17] Note that at the limit $\alpha = 1/2$, $\frac{1}{2(1-1/2)} = 1$ so that $\gamma \in [0, 1]$ is fulfilled.

[18] As explained at the end of Sect. 4.1.3, it is straightforward to formalize that the entrepreneur stays out of the market if the innovation project fails.

$$s.t: \pi_E(\rho_E) > 0. \tag{25}$$

Let us first focus on an interior solution: a solution, ρ_E^*, that fulfills constraints Eqs. 24 and 25. The first-order condition for an interior solution, $\frac{dE[\Pi_E]}{d\rho_E} = 0$, is then

$$\underbrace{\pi_E(\rho_E^*)}_{\text{Securing success (SS):}} = \underbrace{\underbrace{-\rho_E^* \times \frac{d\pi_E(\rho_E^*)}{d\rho}}_{(-)}}_{\text{Cost of going safer (CGS):}} > 0. \tag{26}$$

As shown in Eq. 26, we can understand this first-order condition through the following two distinct effects.

SS effect The left-hand side of Eq. 26 gives the *increase in expected profit from choosing a marginally safer project* and is simply the reduced product market profit from succeeding, $\pi_E(\rho_E)$. We label this the *SS effect*.

Cost-of-going-safer effect The right-hand side of Eq. 26 represents *the reduction in expected profit from choosing a marginally safer project*, which we label the *CGS effect*. The downside of choosing a safer project stems from a lower level of consumer willingness to pay and more aggressive competition from the incumbent. To see this, we use the envelope theorem in Eq. 22 to obtain the following[19]:

$$\frac{d\pi_E(\rho_E)}{d\rho_E} = \left(\overbrace{\underbrace{\frac{\partial a_E}{\partial \rho_E}}_{(-)} + \underbrace{\frac{\partial P_E}{\partial q_I}}_{(-)} \times \underbrace{\frac{dq_I^*}{d\rho_E}}_{(+)}}^{(-)} \right) \times q_E^*(\rho_E) < 0. \tag{27}$$

Direct demand effect; Strategic effect

The first term shows that choosing a project with a marginally higher probability of success reduces consumers' willingness to pay (if the project is successful), $\frac{\partial a_E}{\partial \rho_E} = -\beta < 0$. This reduces the entrepreneur's product market price from Eq. 10. The second term captures that a lower willingness to pay for the entrepreneur's product also induces the rival incumbent to be more aggressive in the product market, which follows since $\frac{dq_I^*}{d\rho_E} = \frac{\beta}{(1-2\alpha)(3-2\alpha)} > 0$ from Eq. 16, thus further reducing the entrepreneur's product market price since $\frac{\partial P_E}{\partial q_I} = -1 < 0$ from the residual demand in Eq. 10. Using the information in Eq. 27, we can then rewrite the *CGS effect* in Eq. 26 as

$$-\rho_E \frac{d\pi_E(\rho_E)}{d\rho_E} = \rho_E \left(1 + \frac{1}{(1-2\alpha)(3-2\alpha)}\right) \beta \times q_E^*(\rho_E) > 0. \tag{28}$$

4.2.1 Optimal project choice

We are now ready to determine the optimal project. First, note that since $P_E - c = (1-\alpha) q_E^*(\rho_E)$ holds from Eq. 12, the reduced product market profit in Eq. 22 is a quadratic function of the Nash output as follows:

$$\pi_E(\rho_E) = (1-\alpha)\left[q_E^*(\rho_E)\right]^2. \tag{29}$$

By inserting Eqs. 28 and 29 into the first-order condition in Eq. 26, we then obtain that

$$\left((1-\alpha) q_E^*(\rho_E) - \left(1 + \frac{1}{(1-2\alpha)(3-2\alpha)}\right) \beta \times \rho_E\right) \times q_E^*(\rho_E) = 0. \tag{30}$$

Note that this first-order condition Eq. 30 holds if the bracketed expression is zero, output is zero, or both of these conditions hold. Thus, we have two candidates for the optimal project, $\hat{\rho}_E$ and ρ_E^*, as follows:

$$q_E^*(\hat{\rho}_E) = 0, \tag{31}$$

$$(1-\alpha) q_E^*(\rho_E^*) - \left(1 + \frac{1}{(1-2\alpha)(3-2\alpha)}\right) \times \beta \times \rho_E^* = 0, \; q_E^*(\rho_E^*) > 0. \tag{32}$$

From Eq. 17, we know that choosing an easier project comes with less consumer appreciation in terms of lower net willingness to pay, which shrinks output, $\frac{\partial q_E^*(\rho_E)}{\partial \rho_E} = -\frac{2\beta(1-\alpha)}{(1-2\alpha)(3-2\alpha)} < 0$. However, choosing ρ_E to achieve zero output cannot be optimal since it would imply that the expected profit is zero, $\Pi_E(\hat{\rho}_E) = \hat{\rho}_E \pi_E(\hat{\rho}_E) = \hat{\rho}_E (1-\alpha)\left[q_E^*(\hat{\rho}_E)\right]^2 = 0$. Thus, only ρ_E^* in Eq. 32 can be a maximum.

To derive ρ_E^*, it is useful to rearrange Eq. 32 to obtain

$$\underbrace{(1-\alpha)\, q_E^*(\rho_E^*)}_{\text{SS:}} = \underbrace{\left(1 + \frac{1}{(1-2\alpha)(3-2\alpha)}\right) \times \beta \times \rho_E^*}_{\text{CGS:}}, \tag{33}$$

[19] Changes in the entrepreneur's own output $q_E^*(\rho_E)$ have only a second-order effect on the reduce-form profit since output is already optimally set from Eq. 12.

where the left-hand side is the SS effect and the right-hand side is the CGS effect, *rewritten* in linear form following Eq. 29.

In Fig. 4(i), we illustrate how the SS and CGS effects shape the equilibrium. The downward-sloping curve labeled SS is the SS effect, showing the benefit from succeeding with a marginally safer project in terms of per-unit profit. The SS curve is downward sloping since the value of SS is lower the more likely the project is to succeed (since the quality of the project is inversely related to its probability of success—see Eq. 9). The upward-sloping curve labeled CGS is the CGS effect and shows the reduction in per-unit profit from a safer project from lower-level consumer willingness to pay and intensified competition from the incumbent. The CGS curve is upward sloping since the higher the CGS is, the more likely the project is to succeed.

The optimal project ρ_E^* is thus given from the intersection of the SS and CGS loci and illustrated at point A in Fig. 4(i). Combining Eqs. 17 and 32, we obtain

$$\rho_E^* = \frac{1}{6\beta}\left(\left(\frac{1-2\alpha}{1-\alpha}\right)\Lambda + 2b - \left(\frac{1-2(1-\alpha)\gamma}{1-\alpha}\right)\alpha d_I\right). \tag{34}$$

In the Appendix A, available upon request from the authors, we (i) verify that ρ_E^* is the unique maximum, i.e., that $\frac{\partial^2 \Pi_E(\rho_E^*)}{\partial \rho_E^2} < 0$, and (ii) derive the conditions under which ρ_E^* satisfies $\rho \in [0, 1]$ and $q_i^*(\rho_E^*) > 0$.

4.2.2 Comparative statics of the entrepreneur's project choice

Let us now explore the comparative statics results of the entrepreneur's project choice.

Amount of incumbent's historical data What is the effect on the entrepreneur's optimal project if the incumbent has access to more historical data?

We put forth the following proposition:

Proposition 2 *If the incumbent firms possesses more historical data d_I, then*

(i) The entrepreneur chooses an R&D project with a lower probability of success when the entrepreneur has weak access to the incumbent's historical data, i.e., when $\frac{d\rho_E^*}{dd_I} < 0$ if $1 - 2\gamma(1-\alpha) > 0$, and

(ii) The entrepreneur chooses an R&D project with a higher probability of success when the entrepreneur has strong access to the incumbent's historical data, i.e., when $\frac{d\rho_E^*}{dd_I} > 0$ if $1 - 2\gamma(1-\alpha) < 0$.

From Eqs. 34 and 8, we have that

$$\frac{d\rho_E^*}{dd_I} = -\frac{\alpha}{\beta} \cdot \frac{(1-2\gamma(1-\alpha))}{(1-\alpha)} \tag{35}$$

Given that the entrepreneur has weak access to the incumbent's historical data, $1 - 2\gamma(1-\alpha) > 0$, a greater amount of historical data held by the incumbent decreases the entrepreneur's output, $\frac{\partial q_E^*}{\partial d_I} < 0$ (as shown in the lower line in Eq. 21). This implies that the value of SS decreases, as illustrated by a downward shift in the SS curve to the curve SS' in Fig. 4(ii). Since the CGS locus is unaffected by Eq. 33, we can infer that the incumbent's possession of more historical data induces the entrepreneur to choose a riskier project, moving from ρ_E^* to $\rho_E^{*\prime} < \rho_E^*$.

In contrast, when the entrepreneur has strong access to the incumbent's historical data, i.e., when if $1 - 2\gamma(1-\alpha) < 0$ holds, the availability of a greater amount of historical data increases the entrepreneur's output, $\frac{\partial q_E^*}{\partial d_I} > 0$ (as shown in the lower line in Eq. 21). This finding implies that the value of SS increases and is illustrated by an upward shift in the SS curve to SS" in Fig. 4(iii). The incumbent's possession of a greater amount of historical data now induces the entrepreneur to choose a safer project, moving from ρ_E^* to $\rho_E^{*\prime\prime} > \rho_E^*$.

Entrepreneur's access to the incumbent's historical data (γ) What is the effect on the entrepreneur's optimal project if the entrepreneur's access to a greater amount of historical operational data is improved? We put forth the following proposition:

Proposition 3 *If the entrepreneur obtains better access to the incumbent's historical data, then the entrepreneur chooses an R&D project with a higher probability of success, i.e., $\frac{d\rho^*}{d\gamma} > 0$.*

From Eq. 34, we have that

$$\frac{\partial \rho_E^*}{\partial \gamma} = \frac{1}{3}\frac{\alpha}{\beta} d_I > 0. \tag{36}$$

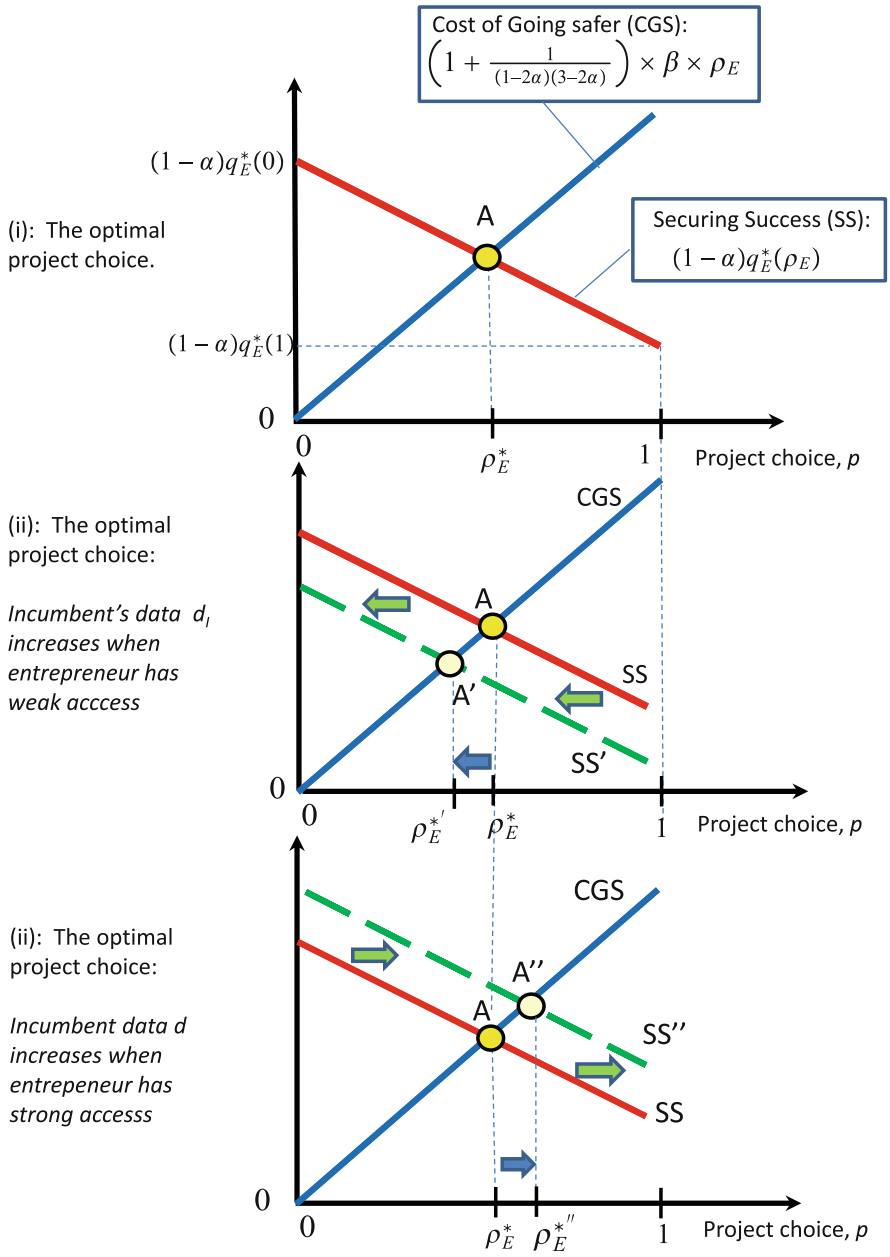

Fig. 4 Panel (**i**) derives the optimal innovation project of the entrepreneur for a given amount of historical data held by the incumbent. Panel (**ii**) illustrates the change in project choice by the entrepreneur when she has weak access to the incumbent's historical data. Panel (**iii**) illustrates the change in project choice by the entrepreneur when she has strong access to the incumbent's historical data

Thus, better access to the incumbent's historical data induces the entrepreneur to choose a safer project. A less risky project then provides less value to consumers if it succeeds.

More efficient ML (α) We can also examine how the entrepreneur's project is affected if ML technology improves, which is captured by an increase in α. We then put forth the following proposition:

Proposition 4 *If ML becomes more effective, then the entrepreneur chooses an R&D project with a lower probability of success and a higher consumer willingness to pay given success, i.e.,* $\frac{d\rho_E^*}{d\alpha}<0$ *if* $\gamma<1/2$.

From Eq. 34, we have that

$$\frac{d\rho_E^*}{d\alpha}=\frac{1}{6}\frac{\left(2\gamma\left(\alpha^2-2\alpha+1\right)-1\right)d_I-\Lambda}{\beta\left(1-\alpha\right)^2}<0 \text{ if } \gamma<1/2. \quad (37)$$

When ML becomes more efficient, it becomes costlier for the entrepreneur to choose a safer project since the strategic effect of a more aggressive incumbent in the product market becomes stronger.

In terms of Fig. 4(i), this situation would cause the CGS locus to twist counterclockwise (not shown). Indeed, by partially differentiating the right-hand side of Eq. 33, we see that the expected CGS increases when α increases:

$$\frac{\partial}{\partial\alpha}\left(\left(1+\frac{1}{(1-2\alpha)(3-2\alpha)}\right)\times\beta\times\rho_E\right)$$
$$=8\frac{1-\alpha}{\left(3+4\alpha^2-8\alpha'\right)^2}\times\beta\times\rho_E>0. \tag{38}$$

The effect of more effective ML on the SS locus is more involved. The entrepreneur chooses output such that profit per unit equals the net reduction in revenues per unit from a unit expansion in sales, $P_E - c = (1-\alpha)q_E^*(\rho_E)$. The more efficient use of data, increasing α, then makes expansion less costly and hence allows the entrepreneur to operate with a lower per-unit profit at an unchanged output level, thus shifting SS downward in Fig. 4(i) (again not shown). Since more efficient ML increases consumers' willingness to pay, this also gives the entrepreneur an incentive to increase her output, which makes the SS effect stronger and shifts the SS' condition further upward. However, as derived above, if the entrepreneur has sufficiently low access to the incumbent's data, then the entrepreneur always responds to more efficient ML by choosing a riskier project.

4.2.3 Why ML and big data may lead to more creation but less destruction

Let us now combine our results and explore the main question of interest in this paper: *What is the impact of more protected big data and ML on the creative destruction process*?

From Eq. 8 and as illustrated in Fig. 2(i), we know that when succeeding with the invention, consumers willingness to pay for the entrepreneur's product will increase:

$$\Delta\, a_E|_{Succeed} = b - \beta\cdot\rho_E^*(d_I). \tag{39}$$

From Eq. 39, a *riskier project* (lower ρ_E^*) then has a *greater value for consumers if it succeeds*

Definition 3 ***Creative entrepreneurship:*** *Entrepreneurship is (more) creative when the entrepreneur takes on more risk and aims for a more innovative invention in her innovation decision.*

If the entrepreneur succeeds and enters the product market, then this will have a business-stealing effect, which will be destructive for the incumbent. We shall then define destructive entrepreneurship as follows:

Definition 4 ***Destructive entrepreneurship:*** *Entrepreneurship is destructive when the entrepreneur, through successful innovation, can enter the market and overtake the incumbent's position as market leader.*

To capture destructive entrepreneurship, we could calculate the market share for each firm. However, without loss of generality, it becomes easier to capture destructive entrepreneurship by comparing profits using reduced-form profits as a function of the incumbent firm's historical data d_I. In the next section, we show that the results also hold with more traditional market shares.

To this end, let $\pi_i(\rho_E^*(d_I), d_I) \equiv \pi_i(q_i^*\left(\rho_E^*(d_I)\right., d_I), q_j^*\left(\rho_E^*(d_I), d_I\right), d_I)$ for $i, j = \{E, I\}$ and $i \neq j$. Then, let the relative reduced-form profit of the entrant be $\varphi_E(d_I)$, or

$$\varphi_E(d_I) = \frac{\pi_E(\rho_E^*(d_I), d_I)}{\pi_I(\rho_E^*(d_I, d_I)} = \frac{(1-\alpha)\left[q_E^*(\rho_E^*(d_I), d_I)\right]^2}{(1-\alpha)\left[q_I^*(\rho_E^*(d_I), d_I)\right]^2}$$
$$= \frac{q_E^*(\rho_E^*(d_I), d_I)}{q_I^*(\rho_E^*(d_I), d_I).} \tag{40}$$

From Eq. 40, it directly follows that destructive entrep-
reneurship can be captured by simply comparing firms' outputs:

$$\varphi_E(d_I) > 1 \Leftrightarrow q_E^*(\rho_E^*(d_I), d_I) > q_I^*(\rho_E^*(d_I), d_I). \tag{41}$$

However, does an equilibrium exist where the entrepreneur overtakes the incumbent, i.e., a postentry market equilibrium, where $\varphi_E(d_I) > 1$? Does this occur when the incumbent has a greater or lesser amount of historical data d_I? How does the entrepreneur's access to the incumbent's historical proprietary data γ affect the likelihood of this equilibrium? How does the higher efficiency of ML α affect the likelihood of this equilibrium?

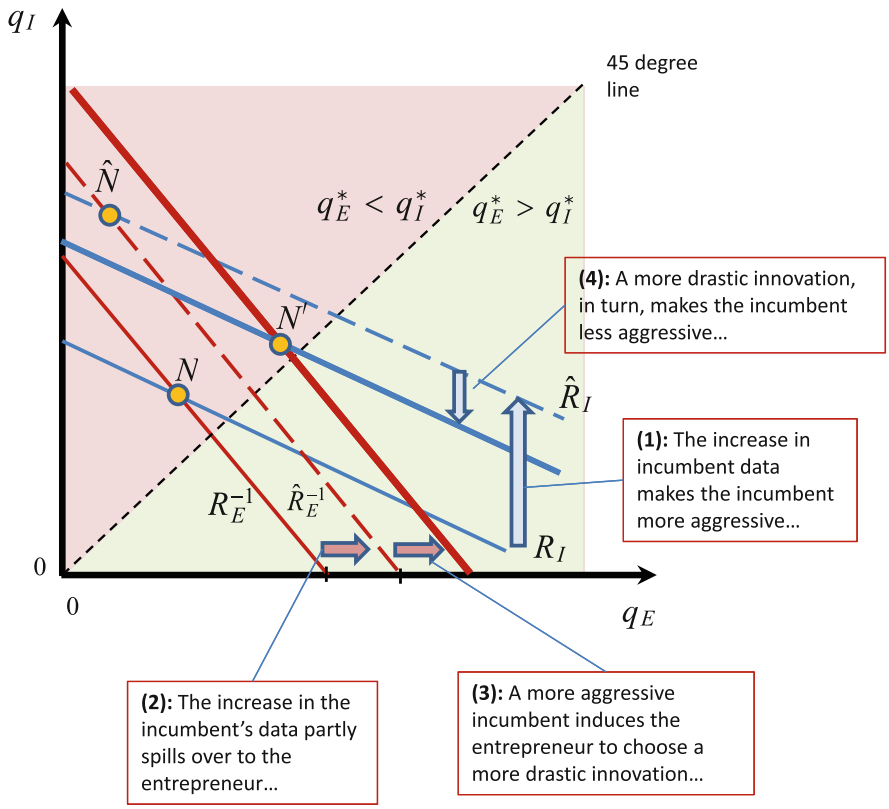

Fig. 5 Illustrating how the Cournot-Nash equilibrium in the product market is affected when the incumbent has access to more historical data and the entrepreneur has weak access to the incumbent's data

To begin our analysis, it useful to return to Fig. 4(i) and (ii). Let us recall that we started in a situation where the data advantage of the incumbent makes the incumbent the market leader: this Nash equilibrium is now reproduced at point N in Fig. 5, where $\varphi_E^N(d_I) < 1$, as point N is above the 45-degree line where $\varphi_E = 1$. Consider what happens if the incumbent has access to a greater amount historical data. In Proposition 1, we show that when the amount of historical data held by the incumbent increases, the incumbent expands more than the entrant given that we hold the project choice of the entrepreneur—and, hence, the quality of the innovation—constant. That is,

$$0 < \frac{\partial q_I^*(\rho_E^*(d_I), d_I)}{\partial d_I} > \frac{\partial q_E^*(\rho_E^*(d_I), d_I)}{\partial d_I} = \begin{cases} < 0; & 1 - 2\gamma\,(1-\alpha) > 0 \\ > 0; & 1 - 2\gamma\,(1-\alpha) < 0. \end{cases} \tag{42}$$

Suppose that the entrepreneur has weak access to the incumbents data in Eq. 42, $1 - 2\gamma\,(1-\alpha) < 0$. The movement from the Nash equilibrium from N to $\hat{N}$ in Fig. 5 involves only the direct change in output from an increase in historical data, i.e., the change in output holding the entrepreneur's innovation project choice constant. However, Proposition 2(i) then shows that the entrepreneur chooses a riskier project when the incumbent has access to a greater amount of historical data, given the weak access of the entrepreneur to these data, i.e., $\frac{d\rho_E^*}{dd_I} < 0$. If such a project succeeds, then it brings with it a higher consumer willingness to pay, and the incumbent therefore faces more aggressive competition from the entrepreneur. This situation is shown by the movement from $\hat{N}$ to N', where the reaction function of the entrepreneur shifts further out in Fig. 5, while the reaction function of the incumbent firm shifts further in. In Fig. 5, the incumbent remains the market leader in the new Nash equilibrium N'. Can this new Nash equilibrium even move from above the 45-degree line (where $\varphi_E < 1$) to a point below the 45-degree line (where $\varphi_E > 1$) as a consequence of the incumbent having access to more historical data?

Let us simplify Eq. 41 further by using Eqs. 16 and 17, and we can obtain an intuitive condition for when

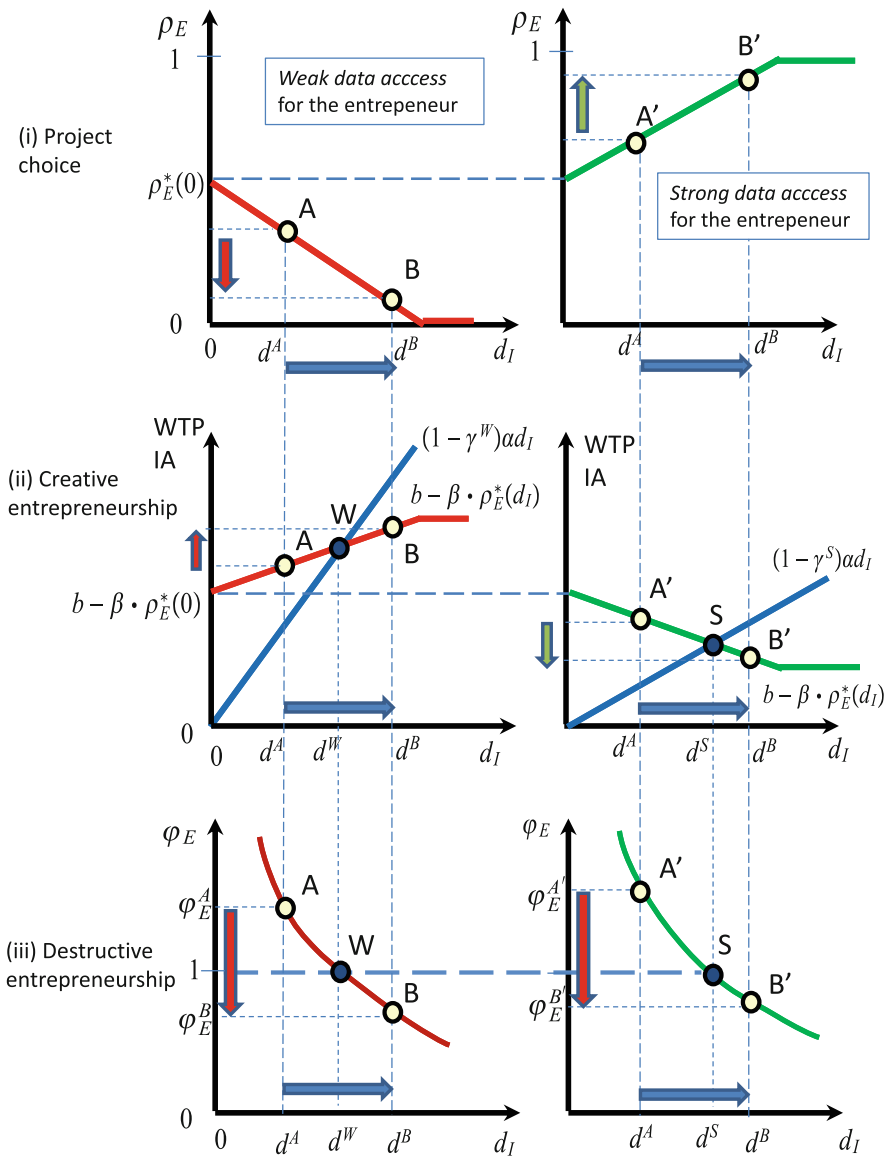

Fig. 6 The left column depicts the case of the weak access of the entrepreneur to the incumbent's historical data, $1-2\gamma\,(1-\alpha)>0$. The right column depicts the case of the strong access of the entrepreneur to the incumbent's historical data, $1-2\gamma\,(1-\alpha)<0$. The panels in (**i**) show the optimal project choice in terms of the probability of success as a function of the amount of historic data held by the incumbent, $\rho_E^*(d_I)$. The panels in (**ii**) show the extra WTP that a successful project brings consumers of the entrepreneurial product, i.e., our measure of creative entrepreneurship $b-\beta\rho_E^*(d_I)$, as well as the advantage had by the incumbent in terms of better access to data (IA), $(1-\gamma)\alpha d_I$. The panels in (**iii**) depict our measure of destructive entrepreneurship, i.e., the relative profit of the entrepreneur if she succeeds with her innovation, $\varphi_E(d_I)$

entrepreneurship is destructive:

$$\varphi_E(d_I)>1 \Leftrightarrow \underbrace{b-\beta\rho_E^*(d_I)}_{\text{Creation effect}} > \underbrace{(1-\gamma)\alpha d_I}_{\text{Big data incumbent advantage effect}}. \tag{43}$$

The left-hand side captures how creative the invention is given that it succeeds, while the right-hand side captures how strong the big data advantage is for the incumbent when it is able to use ML for all its historical data. Note that if the creation effect of the innovation dominates the incumbent advantage effect, then the invention is considered destructive, i.e., $\varphi_E(d_I)>1$. Figure 6 illustrates how the amount of historical data possessed by incumbent d_I affects the equilibrium market share of the entrepreneur, $\varphi_E(d_I)$. The left-hand side of the figure describes the case of weak access to the incumbent's historical data on the part of the entrepreneur. The right-hand side of the figure describes the case of strong access to the incumbent's historical data on the part of the entrepreneur.

The left column depicts the case of the weak access of the entrepreneur to the incumbent's historical data,

$1-2\gamma\,(1-\alpha) > 0$. The right column depicts the case of the strong access of the entrepreneur to the incumbent's historical data, $1-2\gamma\,(1-\alpha) < 0$. The panels in (i) show the optimal project choice in terms of the probability of success as a function of the amount of historical data held by the incumbent, $\rho_E^*(d_I)$. The panels in (ii) show the extra willingness to pay that a successful project brings consumers of the entrepreneur's product, i.e., our measure of creative entrepreneurship $b-\beta\rho_E^*(d_I)$, as well as the advantage had by the incumbent in terms of better access to data (IA), $(1-\gamma)\alpha d_I$. The panels in (iii) depict our measure of destructive entrepreneurship, i.e., the relative profit of the entrepreneur if she succeeds with her innovation, that is, $\varphi_E(d_I)$.

Strong access to the incumbent's historical data From Proposition 2, we know that if the entrepreneur has strong access to the incumbent's historical data, then she chooses a less risky project (a higher success probability ρ_E^*) in response to the incumbent having more data. This finding is illustrated by the upward-sloping green curve in the right diagram in Fig. 6(i). The intuition is that when the entrepreneur has strong access to the incumbent's historical data, an increase in the amount of the incumbent's data will strengthen the so-called SS effect, that is, increase the cost of failure for the entrepreneur. This situation induces the entrepreneur to take on less risk (as illustrated in Fig. 4(iii)).

In the right diagram of Fig. 6(ii), we then depict the creation effect, $b-\beta\rho_E^*(d_I)$, and the big data incumbent advantage effect, $(1-\gamma^S)\alpha d_I$, where γ^S indicates strong access, $1-2\gamma^S\,(1-\alpha) < 0$. The blue curve is the big data incumbent advantage effect, which naturally increases with the amount of historical data d_I for the incumbent. The green curve is the creation effect, which is decreasing with the amount of historical data d_I for the incumbent since—as shown in panel (i)—strong access to the incumbent's data induces the entrepreneur to take on less risk, which (when the innovation succeeds) creates less valuable innovation, leading to a lesser increase in consumers' willingness to pay.

Let d^S be the level of incumbent data that equalizes the incumbent advantage effect and the entrepreneur's creation effect. For low levels of historical data, $d_I \in [0, d^S)$, the creation effect dominates the big data incumbent advantage effect; the entrepreneur steals business from the incumbent in the product market and becomes the market leader. This situation is shown in panel (iii) in the right-hand diagram, where $\varphi_E(d_I) > 1$ for $d_I \in [0, d^S)$. At larger amounts of historical incumbent data, $d_I > d^S$, the incumbent advantage effect dominates the entrepreneur's creation effect, and the incumbent remains the market leader, which is as shown in panel (iii) in the right-hand diagram, where $\varphi_E(d_I) < 1$ for $d_I > d^S$.

Weak access to the incumbent's historical data What if the entrepreneur has weak access to the incumbent's historical data? This case is shown on the left side of Fig. 6. The entrepreneur then responds to an increase in the incumbent's historical data with a riskier project (a decrease in ρ_E), as shown by the upward-sloping red curve in the left diagram of Fig. 6(i). The intuition is now that weak access to incumbents' increasing amount of data worsens the entrepreneur's position in the product market, which, in turn, softens the SS effect, as failure in this case is less costly.

Turning to Fig. 6(ii), the blue upward-sloping curve in the left diagram depicts the big data incumbent advantage effect, $(1-\gamma^S)\alpha d_I$, where γ^W indicates weak access, $1-2\gamma^W\,(1-\alpha) < 0$. The red curve is the creation effect, $b-\beta\rho_E^*(d_I)$, which is now increasing in the amount of historical data d_I for the incumbent since, as panel (i) shows, weak access to incumbents' data induces the entrepreneur to take on more risk, which (if innovation succeeds) creates more valuable innovation, leading to a stronger increase in consumers' willingness to pay.

Let d^W be the level of incumbent data that equalizes the incumbent advantage effect and the entrepreneur's creation effect. For low levels of incumbent data, $d_I \in [0, d^W)$, the creation effect again dominates the incumbent advantage effect, and the entrepreneur becomes the market leader. This situation is shown in panel (iii) in the left-hand diagram, where $\varphi_E(d_I) > 1$ for $d_I \in [0, d^W)$. At higher amounts of historical incumbent data, $d_I > d^S$, the incumbent advantage effect again dominates the entrepreneur's creation effect, and the incumbent remains the market leader. This situation is as shown in panel (iii) in the left-hand diagram, where $\varphi_E(d_I) > 1$ for $d_I > d^W$.

Our analyses show that regardless of the type of access the entrepreneur has to the incumbent's data, she cannot become the market leader postentry when the incumbent has access to a sufficient amount of histori-

cal data: in other words, when the incumbent has gathered enough data, entry is less likely to be destructive. However, the analysis also shows that the entrepreneur responds differently in her choice of innovation project, which leads her to pursue more creative inventions under weak access to the incumbent's data and less creative inventions when she has stronger access.

We can summarize the above points using the following proposition:

Proposition 5 *Let d^S and d^W be defined from $b-\beta\cdot\rho_E^*(d^l)=(1-\alpha)\,d^l$ for $l=\{S,W\}$. Suppose that $d^A<\max(d^S,d^W)<d^B$. Then, if the incumbent's data d_I increases from $d_I=d_A$ to $d_I=d_B$, then the following holds:*

*(i) Under weak access, $1-2\gamma\,(1-\alpha)>0$, entrepreneurship becomes **more creative**, i.e., $b-\beta\cdot\rho_E^*(d_B)>b-\beta\cdot\rho_E^*(d_A)$ and **less destructive**, i.e., $\varphi_E(d_B)<1<\varphi_E(d_A)$*

*(ii) Under strong access, $1-2\gamma\,(1-\alpha)<0$, entrepreneurship becomes **less creative**, i.e., $b-\beta\cdot\rho_E^*(d_B)<b-\beta\cdot\rho_E^*(d_A)$ and **less destructive**, i.e., $\varphi_E(d_B)<1<\varphi_E(d_A)$*

4.2.4 Creative and destructive entrepreneurship and big data: a parametric example

In the Appendix A, we provide a simple parametric example, where we illustrate the effects of the model and numerical illustrations of Proposition 5. We then show how the amount of historical data possessed by the incumbent and the efficiency of ML affects (i) the R&D project choice of the entrepreneur, ρ_E^*; (ii) our measure of *creative entrepreneurship*, i.e., the increase in willingness to pay for the entrepreneur's product if the entrepreneur succeeds with her R&D; $\Delta\, a_E|_{Succeed}=b-\beta\cdot\rho_E^*$, and (iii) our measure of *destructive entrepreneurship*, φ_E.

4.3 Stage 1: becoming an AI entrepreneur and the open-source community

Let us now close the model and examine the incentive to become an AI-based entrepreneur and the way in which this incentive depends on the increasing importance of the operational data possessed by the incumbent. To this end, we assume that the entrepreneur faces a fixed R&D cost or investment cost, F, to become an entrepreneur. This cost can consist of the cost of evaluating different types of possible business opportunities, the cost of setting up the basics of the business, the opportunity cost of becoming an entrepreneur in the form of forgone wage earnings, etc. In our setting, we can also think of F as a fixed cost for entrepreneurs to obtain access to and knowledge of ML technology. One possibility to reduce such cost would be to share knowledge and codes from the open-source community.

Since fixed cost F is incurred before the entrepreneur makes her R&D decision, the expected profit for an entrepreneurial venture $\mathrm{E}[\Pi_E]$ becomes

$$\mathrm{E}[\Pi_E]=\rho_E^*\times\pi_E(\rho_E^*)-F. \tag{44}$$

We can then examine how the expected profit of the entrepreneur depends on the amount of historical data possessed by the incumbent by differentiating $\mathrm{E}[\Pi_E]$ w.r.t. from d_I:

$$\frac{d\mathrm{E}[\Pi_E]}{dd_I}=\underbrace{\left[\pi_E(\rho_E^*)+\rho_E^*\frac{\partial\pi_E(\rho_E^*)}{\partial\rho_E}\right]}_{=0}\frac{d\rho_E^*}{dd_I}+\rho_E^*\times\frac{\partial\pi_E(\rho_E^*)}{\partial d_I}=\rho_E^*\times\frac{\partial\pi_E(\rho_E^*)}{\partial d_I}, \tag{45}$$

where we use the fact that the f.o.c. for the project choice in period 1 implies that $\pi_E(\rho_E^*)+\rho_E^*\frac{\partial\pi_E(\rho_E^*)}{\partial\rho_E}=0$. Using Eqs. 29 and 21, we can rewrite Eq. 45 as follows:

$$\frac{d\mathrm{E}[\Pi_E]}{dd_I}=-2\rho_E^*\times(1-\alpha)\,q_E^*(\rho_E^*)\times\frac{(1-2(1-\alpha)\gamma)\alpha}{(1-2\alpha)(3-2\alpha)}=\begin{cases}<0;\quad 1-2\,(1-\alpha)\,\gamma>0:\ \text{weak access}\\ >0;\quad 1-2\,(1-\alpha)\,\gamma<0:\ \text{strong access.}\end{cases} \tag{46}$$

We can thus put forth the following proposition:

Proposition 6 *When the amount of historical data possessed by the incumbent increases, the incentive to become an entrepreneur increases when the entrepreneur has strong access to the incumbent's historical data, i.e., when $1-2\,(1-\alpha)\,\gamma<0$, and decreases when the entrepreneur has weak access to the incumbent's historical data, i.e., when $1-2\,(1-\alpha)\,\gamma>0$.*

Propositions 5 and 6 point to a policy dilemma. It is likely that the entrepreneur will have limited access to the incumbent's data if data access is unregulated. Proposition 5 then suggests that the trend toward the

greater availability of big data and greater use of ML will lead to less destructive entrepreneurship—with sustained incumbent market power—but also to less entrepreneurship in general, as those entering markets in which incumbents have big data advantages will be less profitable. This finding suggests a policy that levels the playing field between entrepreneurs and incumbents by forcing incumbents to give entrepreneurs access to their data. Indeed, such a policy would not only encourage more entrepreneurship in general but would also give rise to more destructive entrepreneurship in particular.

To see this, let us first differentiate Eq. 44 in γ to obtain

$$\frac{d\mathrm{E}\left[\Pi_E\right]}{d\gamma}=2\rho_E^*\times(1-\alpha)\,q_E^*(\rho_E^*)\times\tfrac{2(1-\alpha)\alpha d_I}{(1-2\alpha)(3-2\alpha)}>0. \tag{47}$$

That is, better access for the entrepreneur to the incumbent's data will increase her expected profit from becoming an entrepreneur, which will make her more likely to invest fixed cost F to take the chance to become a successful entrepreneur.

Better data access will also lead to more destructive entrepreneurship. To see this, recall from Eq. 43 that entrepreneurship is destructive, $\varphi_E(d_I)>1$, when the entrepreneur's creation effect, $b-\beta\cdot\rho_E^*$, is greater the incumbent big data advantage effect, $(1-\gamma)\alpha d_I$. From Eq. 36, it follows that better data access will weaken the creation effect since, under such conditions, the entrepreneur will choose a safer project. However, the decline in creative entrepreneurship is being dominated by lesser incumbent data advantage, that is,

$$\underbrace{\left|\frac{\partial\left[b-\beta\cdot\rho_E^*\right]}{\partial\gamma}\right|=\frac{1}{3}\alpha d_I}_{\text{Decline in creative entrepreneurship}}<\underbrace{\alpha d_I=\left|\frac{\partial\left[(1-\gamma)\alpha d_I\right]}{\partial\gamma}\right|}_{\text{Decline in incumbent advantage}}. \tag{48}$$

The left-hand side in inequality in Eq. 48 reveals the drawback in trying to level the playing field between entrepreneurs and incumbents by giving the entrepreneur better data access—this policy weakens the incentives for creative entrepreneurship.

These findings suggest that policies supporting early entrepreneurial ventures might instead be warranted. Subsidizing the fixed cost to become an entrepreneur F would increase $\mathrm{E}\left[\Pi_E\right]$ in Eq. 44 and make entry by the entrepreneur more likely without reducing her incentive to be creative (i.e., to take on more risk). In summary, we can put forth the following proposition:

Proposition 7 *A policy that makes operational data generally available (increasing γ) may be suboptimal: while it may make entrepreneurial entry more likely and increase the level of destructive entrepreneurship (make the entrepreneur the new market leader), it may also reduce the level of creative entrepreneurship (create less value for consumers). An alternative or complementary policy might be to subsidize the fixed R&D or investment cost F (i.e., reduce the cost of becoming an entrepreneur with access to ML technology). This policy will promote entrepreneurial entry without reducing creative entrepreneurship.*

How can the above proposition become operational in real policy terms? One way for policy makers to improve entrepreneurs' access to ML technology is to increase resources into programming in the education system. This strategy will increase the number of potential entrepreneurs with programming skills—but also the pool of employees that entrepreneurs could hire in their startups. This situation will likely benefit entrepreneurial firms more than it will incumbents since incumbents are more likely to have opportunities for job training in programming for their employees.

Another policy to increase access to ML for entrepreneurs would be to support open-source communities, where software libraries and algorithms are freely available to developers and entrepreneurs alike. The open-source software community has produced innovations like the Firefox web browser, Apache server software, and Linux operating system. In terms of the development of applications of generative AI, the open-source community has been very active. At the beginning of March 2023, the open-source community obtained access to their first competent foundation model, as Meta's LLaMA was leaked to the public.

A few months later, variants with instruction tuning, quantization, quality improvements, etc., emerged[20]

An active policy discussion is also being carried out on how open-source AI should be supported and monitored. The IPOL (2021) analyzes how EU policy could support and monitor open-source AI in a report requested by the European Parliament's Special Committee on Artificial Intelligence in a Digital Age (AIDA), examining the main open-source AI pros and cons and proposing different policy measures to support AI open source. The findings in our analysis suggest that such an effort can stimulate AI-based entrepreneurship, particularly creative AI-based entrepreneurship.

5 Concluding remarks

This paper investigates how ML applications and increased incumbent operational data affect entrepreneurship incentives. In a model where incumbents have an initial advantage in ML technology and access to (historical) operational data, we show how increased ML applications on operational data affect entrepreneurial entry and the type of entrepreneurship. In particular, we show that limited access to operational data can induce entrepreneurs to take on more risk, thereby increasing their probability of developing transformative products. Thus, increased ML for incumbents' operational data may make the creative destruction process not only more creative but also less destructive.

Our model gives rise to several testable predictions. First, the model predicts that successful entry will be less frequent in ML propriety data-intensive industries, but that when entry appears, it will involve more novel products. Second, the model also has testable predictions on country levels: we should expect entrepreneurial firms in ML-intensive industries in countries with a stronger protection of propriety data to be fewer in number but more productive than those in countries with a weaker protection of propriety data. We should expect entrepreneurial firms in ML propriety data-intensive industries in countries with many individuals participating in open-source communities to be more productive than those in countries with few individuals participating in such communities.

Policy implications This paper has important implications for both entrepreneurs and incumbents. First, entrepreneurs should consider that challenging incumbents in the era of ML will be more difficult since incumbents' use of ML for proprietary data makes them more formidable competitors. This fact implies that entrepreneurs need to become riskier and more creative in the future to obtain a competitive edge. They may therefore seek support from angels or venture capital firms and use their financing and experience to become more novel in their ventures. Incumbents conversely have the incentive to employ ML applications on their operational data so that they become so efficient that they force entrepreneurs to take on so much risk that entrepreneurship will seldom pose a (destructive) threat.

Data protection and privacy issues became a flashpoint in the media due in part to high-profile data breaches such as that at Equifax in 2017 and in part to the high-profile exposure of Facebook users' data to Cambridge Analytica in 2016 and 2017. Partially in response to these events, government regulators have instituted tighter rules on data protection. This development has most notably manifested in Europe in the form of the GDPR. The results derived in the paper suggest that a complementary policy might be to support entrepreneurs' access to and knowledge of ML technology since it stimulates creative entrepreneurship. The subsidization of R&D by small entrepreneurial firms will increase effort but not reduce risk taking.

Limitations The model has several limitations. The result that entrepreneurs choose more creative (riskier) projects when ML becomes more efficient and that incumbents' proprietary data becomes more critical depends on the assumption that entrepreneurs do not face substantial financial restrictions. If they do, then our results would be less relevant since ML will block entrepreneurship entirely if it becomes sufficiently efficient. However, the growing venture capital and angel market might relax such financial restrictions.

[20] Lerner and Tirole (2005) present an empirical analysis of the determinants of license choice using the SourceForge database, a compilation of nearly 40,000 open-source projects, and find that projects with unrestricted licenses attract more contributors. Engelhardt and Freytag (2013) compile data on the worldwide allocation of the activities of developers registered at SourceForge, finding that interpersonal trust has a positive impact on the number of OSS developers as well as on the OSS activity level. Moreover, they find that the enforceability of IP rights (IPRs) positively affects OSS activities.

A second limitation is that we assume that the incumbent cannot acquire the entrepreneurial firm. A particular case would be if a potential acquisition by the incumbent serves the sole purpose of shutting down the invention. While this feature is sometimes relevant, there is also evidence that many (leading) firms, such as Microsoft, Google, and Ericsson, acquire startups to incorporate them highly efficiently into their businesses. In fact, entrepreneurs must still employ a sufficiently creative (risky) strategy to become interesting enough to be a target.

A third limitation is that we have allowed only for two incumbent advantages—access to historical data and sunk investment in ML—and only one advantage for the entrepreneur—the opportunity to engage in R&D. There are other advantages and disadvantages for the different agents that seem relevant for investigation. For instance, the incumbent might have more efficient production processes, face lower variable costs, and have better access to external finance. These advantages might reinforce the incentive for entrepreneurs to take risks in the R&D phase, and investigating this in detail seems to be a promising avenue for future research.

Finally, we have assumed that firms interact in an oligopolistic setting. Our oligopolistic setup is essential in many markets where ML and generative AI are applied. However, although the risk decision of R&D will also be affected by increased ML and generative AI in other types of market structures, this issue is left to future research.

Future research Our analysis treats the entrepreneur's human capital as a constant. Varian (2018) notes that in the traditional form of learning by doing, learning is passive, but in practice, learning requires active investment in ML machinery and human capital. Thus, human and financial capital quality likely affect ML applications and the R&D process. Therefore, endogenizing the human capital level in the analysis is a fruitful avenue for future research.

Acknowledgements We gratefully acknowledge financial support from the Jan Wallander and Tom Hedelius Research Foundation and the Marianne and Marcus Wallenberg Foundation (grant number 2020.0049). We have benefitted from the feedback provided by participants at numerous seminars.

Appendix A

A.1 Creative and destructive entrepreneurship and big data: a parametric example

We here provide numerical illustrations of Proposition 5. We start with the R&D project choice of the entrepreneur, ρ_E^*. We then turn to our measure *creative entrepreneurship*, i.e., the increase in willingness to pay for the entrepreneur's product if the entrepreneur succeeds with her R&D, $\Delta\, a_E|_{Succeed} = b - \beta \cdot \rho_E^*$. Finally, we look at our measure of *destructive entrepreneurship*, where we illustrate that results are qualitatively the same when we use our relative profit as measure of destructive entrepreneurship, φ_E, and when we use a more traditional market share measure. We study how each outcome varies with the amount of historical incumbent data d_I and the state of ML captured by the effectiveness parameter α. As prompted by Proposition 5, for each measure, we compare the two cases of weak and strong access for the entrepreneur to the incumbent's data.

A.1.1 Risk-taking behavior and big data

We begin by examining how the risk behavior of entrepreneurs in the innovation process depends on the amount of historical data that the incumbent possesses and on the efficiency of ML. Start with the upper panel in Fig. 7 with weak access. Note that risk-taking increases as the entrepreneur is choosing a lower success probability ρ_E^* when we move in north-east direction. From Proposition 5(i), she takes on more risk when incumbent data increases as successful entry is associated with less value from a weaker SS effect: From Proposition 4, she also take on more risk when

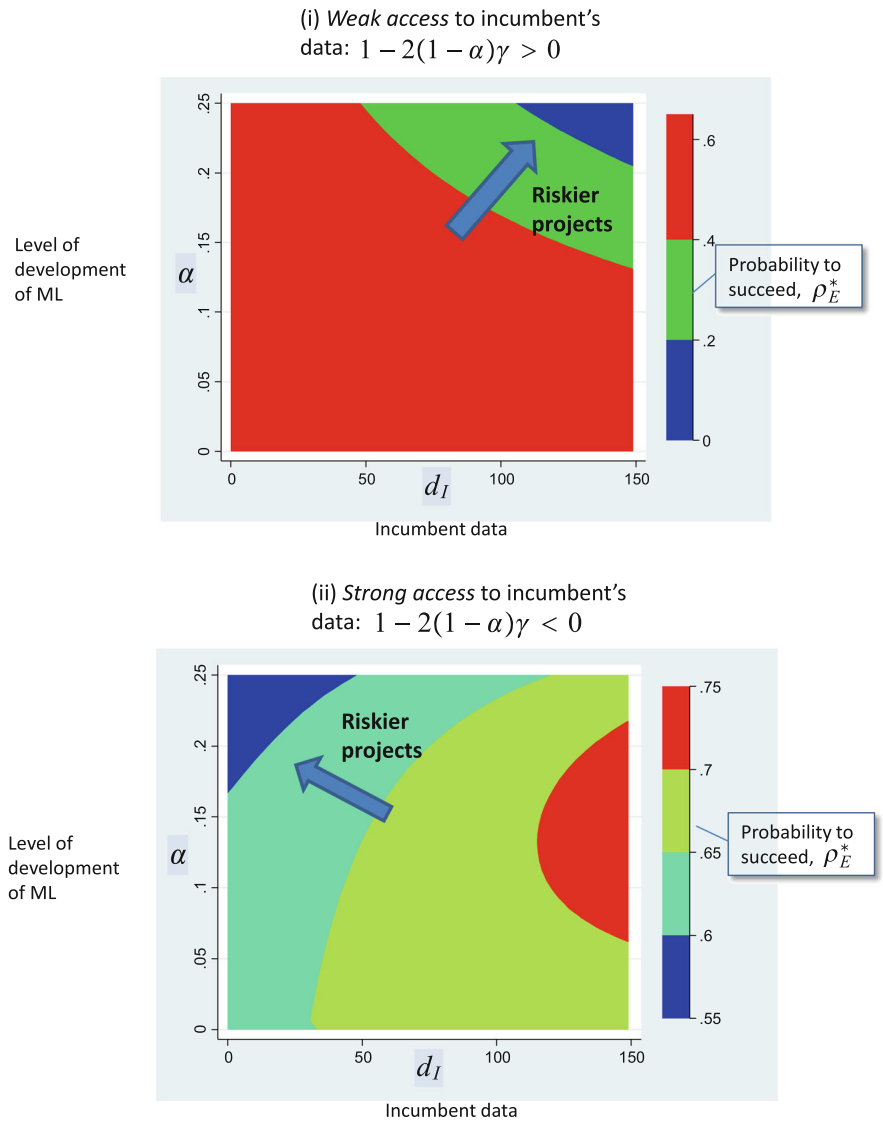

Fig. 7 Illustrating risk-taking by the entrepreneur through her choice of success probability ρ_E^*. Panel (**i**) shows contours of ρ_E^* as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has weak access to the incumbent's data. Panel (**ii**) shows contours of ρ_E^* as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has strong access to the incumbent's data. Parameter values set at $\Lambda = 15, b = 12, \beta = 10$ combined with $\gamma = 0.25$ in panel (**i**) and $\gamma = 0.75$ in panel (**ii**)

ML becomes more efficient in using these data. When ML becomes more efficient under weak access, the CGS effect is strengthened—it becomes more costly for the entrepreneur to choose a safer project due to the strategic effect of a more aggressive incumbent in the product market.

Then, turn to the lower panel with strong access. From Proposition 5(ii), we know that more incumbent data induces the entrepreneur to reduce her risk-taking as the value of entry increases which strengthens the SS effect. As shown in the lower panel in Fig. 7, risk-taking now increases in the north-west direction. As indicated by Eq. 37, the entrepreneur will respond to more efficient ML by choosing more risky R&D projects if the amount of historical data is not too large, again due to a stronger CGS effect. However, from Eq. 37, we also note that when the incumbent has abundant historical data, we there may be a nonlinear effect on risk-taking by the entrepreneur. Indeed, we can see that when d_I is sufficiently large, an increase in α first induces the entrepreneur to go for safer projects which is then reversed when α becomes sufficiently large.

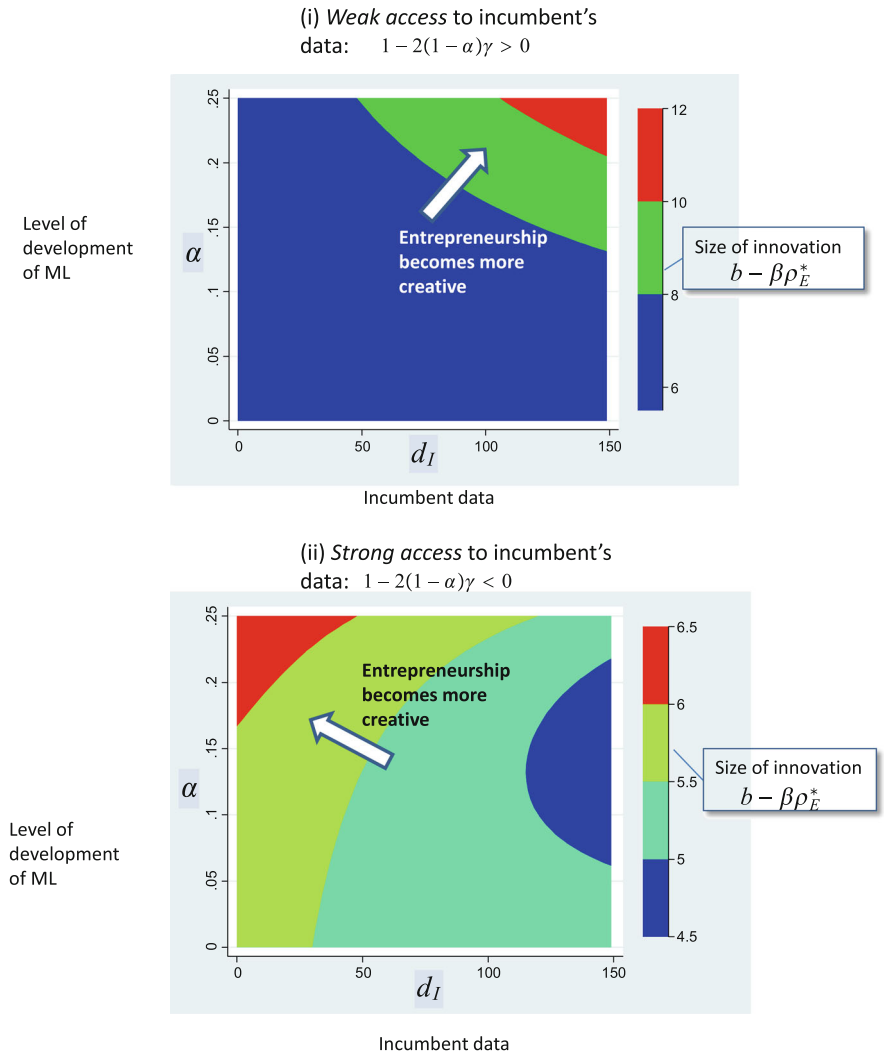

Fig. 8 Illustrating creative entrepreneurship as measured the size of a successful innovation $b-\beta\cdot\rho_E^*$ in terms if the increase in consumers willingness to pay. Panel (**i**) shows contours of $b-\beta\cdot\rho_E^*$ as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has weak access to the incumbent's data. Panel (**ii**) shows contours of $b-\beta\cdot\rho_E^*$ as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α,when the entrepreneur has strong access to the incumbent's data. Parameter values set at $\Lambda=15, b=12, \beta=10$ combined with $\gamma=0.25$ in panel (**i**) and $\gamma=0.75$ in panel (**ii**)

A.1.2 Creative entrepreneurship and big data

Having illustrated how the amount of historical incumbent data, d_I, and the effectiveness with which this data can be used by ML, α, affect the entrepreneur's risk-taking through her R&D project choice, ρ_E^*, we now turn to have these choices affect creative entrepreneurship. Recall that Definition 3 defines creative entrepreneurship in terms of how much consumer willingness to pay (willingness to pay) increases when the entrepreneur succeeds with her R&D project, $\Delta\, a_E|_{Succeed} = b-\beta\cdot\rho_E^*$. Since the increase in willingness to pay for a successful invention is larger if the entrepreneur succeeds with a more risky project, i.e., with a project with a lower success probability ρ_E^*, the size of the creation effect will be a direct mapping of the entrepreneur's level of risk-taking.

Indeed, comparing Figs. 7 and 8, we observe that creativity in entrepreneurship maps the R&D risk behavior of the entrepreneur: Under weak access in the top panel in Fig. 8, entrepreneurship is more creative in the north-east direction. Under strong access in the bottom panel in Fig. 8, in contrast, entrepreneurship is more creative in the north-west direction.

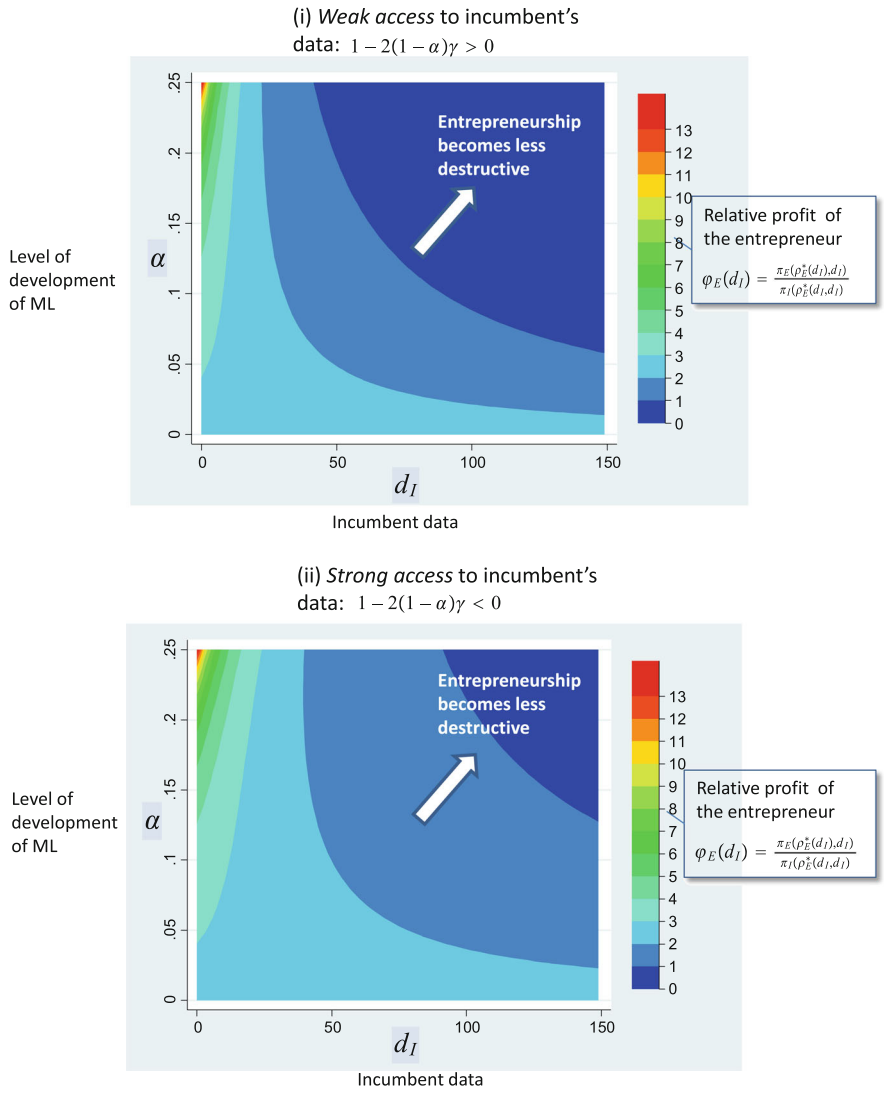

Fig. 9 Illustrating destructive entrepreneurship measured as the relative size of the entrepreneur's profit, φ_E. Panel (**i**) shows contours of φ_E as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has weak access to the incumbent's data. Panel (**ii**) shows contours of φ_E as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has strong access to the incumbent's data. Parameter values set at $\Lambda = 15, b = 12, \beta = 10$ combined with $\gamma = 0.25$ in panel (**i**) and $\gamma = 0.75$ in panel (**ii**)

A.1.3 Destructive entrepreneurship and big data

Recall that Definition 4 defines destructive entrepreneurship as entrepreneurial entry with successful R&D where the entrepreneur becomes the market leader. Proposition 5 showed that entrepreneurship will be less destructive when the incumbent gets access to more historical data—regardless if the entrepreneur has weak or strong access to these data. This is illustrated in Fig. 9 which uses our relative profitability measure φ_E in Eq. 40.The top panel shows the case of weak access, and the bottom panel shows the case of strong access. More abundant incumbent historical data eventually leads to less destructive entrepreneurship. The impact of more efficient use of these data through more efficient ML again depends on the amount of data that the incumbent has access to. In both panels, we see a nonlinear pattern where increasing the ML parameter α leading destructive entrepreneurship when the amount of historical data at the incumbent d_I is low, whereas the opposite is true when the amount of data possessed by the incumbent is high. In Fig. 10, we see the same pattern when we measure destructive entrepreneurship using the more conventional market share measure.

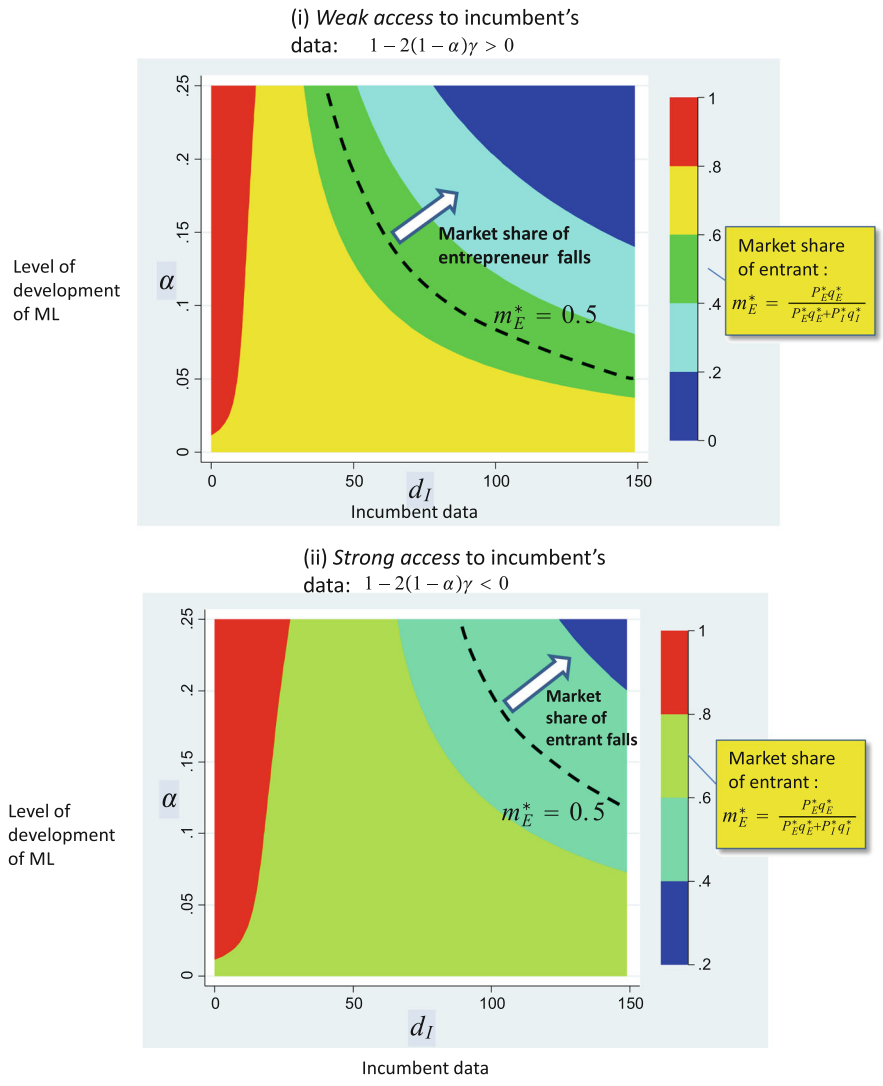

Fig. 10 Illustrating destructive entrepreneurship measured as the market share of the entrepreneur's m_E. Panel (**i**) shows contours of m_E as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has weak access to the incumbent's data. Panel (**ii**) shows contours of m_E as a function of the amount of historical data held by the incumbent, d_I, and the effectiveness of machine learning (ML), α, when the entrepreneur has strong access to the incumbent's data. Parameter values set at $\Lambda = 15, b = 12, \beta = 10$ combined with $\gamma = 0.25$ in panel (**i**) and $\gamma = 0.75$ in panel (**ii**)

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature, 54*(2), 442–492.

Agrawal, A., J. Gans, and A. Goldfarb (2018). Prediction machines: The simple economics of artificial intelligence. Harvard Business Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1–6.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives, 33*(2), 31–50.

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. *AEA Papers and Proceedings, 109*, 33–37.

Batikas, M., S. Bechtold, T. Kretschmer, and C. Peukert (2020). European privacy law and global markets for data. Technical Report 14475, London, Center for Economic Policy Research Discussion Paper.

Bessen, J. (2018). *The policy challenge of artificial intelligence* (pp. 18–16). Law and Economics Research Paper No: CPI Antitrust Chronicle. Boston Univ. School of Law.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., & Trench, M. (2017). *Artificial intelligence the next digital frontier?* McKinsey Global Institute Discussion Paper: Technical report.

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

Campbell, J., Goldfarb, A., & Tucker, C. (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy, 24*(1), 47–73.

Choné, P. and L. Linnemer (2019). The quasilinear quadratic utility model: An overview. HAL Working Papers. hal-02318633.

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (Vol. 1, pp. 129–213). North-Holland.

Dutton, T. (2018). *An overview of national AI strategies*. Medium: Politics+ AI.

Engelhardt, S., & Freytag, A. (2013). Institutions, culture, and open source. *Journal of Economic Behavior & Organization, 95*, 90–110.

Farboodi, M., Mihet, R., Philippon, T., & Veldkamp, L. (2019). Big data and firm dynamics. *AEA Papers and Proceedings, 109*, 38–42.

Färnstrand Damsgaard, E., Hjertstrand, P., Norbäck, P.-J., Persson, L., & Vasconcelos, H. (2017). Why entrepreneurs choose risky r&d projects - but still not risky enough. *The Economic Journal, 127*(605), F164–F199.

Gans, J. (2023). Artificial intelligence adoption in a competitive market. *Economica, 90*(358), 690–705.

Gilbert, R. (2006). Looking for Mr. Schumpeter: Where are we in the competition-innovation debate? *Innovation Policy and the Economy, 6*, 159–215.

Haufler, A., Norbäck, P.-J., & Persson, L. (2014). Entrepreneurial innovations and taxation. *Journal of Public Economics, 113*, 13–31.

Henkel, J., T. Rø nde, and M. Wagner (2015). And the winner is–acquired. entrepreneurship as a contest yielding radical innovations. Research Policy 44(2), 295–310.

Himel, S. and R. Seamans (2017). Artificial intelligence, incentives to innovate, and competition policy. Antitrust Chronicle 1(3).

IPOL (2021). Challenges and limits of an open source approach to artificial intelligence. Technical report, Policy Department for Economic, Scientific and Quality of Life Policies Directorate-General for Internal Policies.

Jia, J., Jin, G. Z., & Wagman, L. (2021). The short-run effects of the general data protection regulation on technology venture investment. *Marketing Science, 40*(4), 661–684.

Johnson, G., S. Shriver, and S. Goldberg (2022). Privacy & market concentration: Intended & unintended consequences of the gdpr.

Lambrecht, A. and C. E. Tucker (2017). Can big data protect a firm from competition? CPI Chronicle.

Lerner, J., & Tirole, J. (2005). The scope of open source licensing. *The Journal of Law, Economics, and Organization, 21*(1), 20–56.

McKinsey (2023a). The economic potential of generative AI the next productivity frontier.

McKinsey (2023b). What every CEO should know about generative.

McKinsey (2023c). What is generative AI?

Rosen, R. (1991). Research and development with asymmetric firm sizes. *The RAND Journal of Economics, 22*(3), 411–429.

Sokol, D. D., & Comerford, R. E. (2016). *Does antitrust have a role to play in regulating big data?* Intellectual Property and High Tech: Cambridge Handbook of Antitrust.

Summit, W. (2019). https://websummit.com/blog/highlights-web-summit-2019.

Thompson, P. (2010). Learning by doing. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (Vol. 1, pp. 429–476). North-Holland.

Varian, H. (2018). Artificial intelligence, economics, and industrial organization. National Bureau of Economic Research (NBER) Working Paper (24839).