The Limits of AI-based Growth:

The Scale, Scope and Boundary of Digital Platforms

June 12th, 2024

Gwendolyn K. Lee
Warrington College of Business
University of Florida
gwendolyn.lee@warrington.ufl.edu

Xun (Brian) Wu
Ross School of Business
University of Michigan
wux@umich.edu

Authors' note:

We would like to submit this paper to join the conversation on "*What should strategic management's dependent variable be*?". To answer this question, we develop a theory about how time-specific causal knowledge augments the firm's strategic decision-making, in a setting where new technologies change the way firms are organized as digital platforms and compete with tacit knowledge. Many of the platform offerings are deployments of AI tools such as recommender systems. The recommender systems cause us to re-evaluate existing theories on the limits of firm growth. As advocated by researchers of recommender systems Joachims et al. (2021), the next generation of recommender and decision-support systems should be viewed as policies that decide what interventions to make in order to optimize a desired outcome. These policies and their associated managerial interventions are central to the firm's strategic decision-making. In this paper, we connect the AI literature from the field of computer science to the field of strategic management with an objective of re-evaluating fundamental strategy research on the scale, scope, and boundary of the firm.

The Limits of AI-based Growth:

The Scale, Scope and Boundary of Digital Platforms

ABSTRACT

Tacit knowledge is a key construct in existing strategy theories that juxtapose knowledge transfer and competitive imitation as blades of the same scissor. Low tacitness enables, on the one hand, the transfer of knowledge inside the firm, but, on the other hand, the leakage of knowledge outside the firm. As an example, the firm's proprietary knowledge is the causal effect of how users of the firm's products/services respond to the recommendations made by artificial intelligence (AI), a system of statistical inference derived from collected data. If the knowledge is more explainable and therefore less tacit, when is it less easily transferred outside the firm? We answer this question by re-evaluating the paradox of tacit knowledge, as "Principles of Explainable AI" demand explanations for the reasons behind AI predictions/recommendations. Our re-evaluation has implications for theories on firm growth and factors that limit the scale, scope and boundary of digital platforms.

Primary area: Knowledge & Innovation

Secondary area: Corporate Strategy

Keyword 1: Tacit knowledge

Keyword 2: Explainable AI

Keyword 3: Digital platforms

Keyword 4: Recommender systems

Keyword 5: Firm growth

INTRODUCTION

Theories on the limits of firm growth posit that a firm's rate of growth is limited by the rate at which the firm's management is capable of rendering services from productive resources that are unused (Penrose, 1959) and the sharing and transfer of the knowledge of individuals and groups within the firm (Kogut and Zander, 1992). A key assumption that both theories have in common is that some managerial capabilities and knowledge are inherently tacit (Polanyi, 1966; Nelson and Winter, 1982). Tacitness constrains the transfer of knowledge and capabilities across organizational units, market segments and industry sectors. Viewed through these theories, a firm's rate of growth is limited by the rate at which capabilities and knowledge can be transferred for deployment inside the firm. Why tacit managerial capabilities and knowledge are more easily transferred inside the firm but difficult to imitate outside the firm is a central research question and a paradox.

The paradox of tacitness is particularly salient when humans use the predictions generated by artificial intelligence (AI) as inputs to make and support decisions that are consequential for everyday life: decisions about what to consume, believe and do. We refer to AI as a system of statistical inference, using machine learning (ML) techniques to find patterns or to identify key variables in data with the purpose of making predictions and taking actions by changing some of those variables. The predictions generated by ML are made not by human-legible rules, but by less scrutable statistical techniques. As an example, a credit-scoring ML algorithm may predict a high probability of an applicant defaulting on a loan. Yet, the loan applicant cannot easily understand whether there were errors in the data that ML used about her, whether there are biases in the historic loan payment data such that algorithms are trained to perpetuate some demographic groups receiving fewer loans compared to others, or what she can do to increase her chance of loan approval in the future.

While the opacity of predictive models keeps a firm's intellectual property hidden, it can have (and has had) severe consequences that deeply impact firm performance and human lives (Rudin, 2019). Recent changes in the institutional environment have sought to make AI more transparent and explainable. For example, under the European Union's General Data Protection Regulation (GDPR),

which is the most important change in data privacy regulation in the last 20 years, automated decisions with legal implications—such as who qualifies for a loan, insurance coverage, or a job—should be transparent and explainable. As another example, the United States National Institute of Standards and Technology (Phillips et al., 2020) introduced Four Principles of Explainable Artificial Intelligence: (1) the system produces an explanation, (2) the explanation is meaningful to humans, (3) the explanation reflects the system's processes accurately, and (4) the system expresses its knowledge limits. The Knowledge Limits Principle states that systems identify cases they were not designed or approved to operate, or their answers are not reliable. By identifying and declaring knowledge limits, this practice safeguards answers so that a judgment is not provided when it may be inappropriate to do so.

However, what are the implications of such institutional change on competitive imitation when AI becomes more transparent and explainable? If the firm's proprietary knowledge is more explainable and therefore less tacit, when is ML, which is a prediction technology, less easily transferred and imitated outside the firm? Our central thesis is that AI is more easily transferred inside the firm but difficult to imitate outside the firm when the knowledge about the system is causal and time specific, for a given level of explainability. Creating time-specific causal knowledge about AI requires not only computation and domain expertise in generating on-demand predictions, but also the firm as a community where the organizing of social relationships and interactions has persistence and regularity in creating such knowledge.

We posit that the persistence and regularity of creating such knowledge results from the process of making predictions and designing interventions. The process involves strategic decision-making by management about whether and how to intervene in the system, such as nudging users to explore, taking action to prevent and correct prediction biases, and re-allocating decision authority from AI to humans. Having time-specific causal knowledge that augments the firm's strategic decision-making is changing the way firms are organized and how they compete. Going back to the example of explaining to bank loan applicants why rejection was recommended by AI, the strategic decision-making by management takes the recommendations made by AI as an input. If managers can use the system of statistical inference to

augment the process of loan approval such that the decisions can be made fairly while reducing default rates, predictions can inform managerial interventions through the firm's strategic decision-making.

Our focus is on the firm's strategic decision-making, whereas existing literature on AI has emphasized the functional aspects of management, featuring operations, marketing, talent acquisition, etc. (Allen and Choudhury, 2022; Choudhury, Starr, and Agarwal, 2020; Iansiti and Lakhani, 2020; Kellogg, Valentine, and Christin, 2020; Raj and Seamans, 2019; Tong, Jia, Luo, and Fang, 2021). Our focus on strategic decision-making, which is a core research area in the field of strategic management, pinpoints how AI is changing the way firms organize ecosystem value chains and compete as disruptive digital platforms. Many of the platform offerings are deployments of AI tools such as recommender systems. The recommender systems cause us to re-evaluate existing theories on the limits of firm growth. As advocated by researchers of recommender systems Joachims et al. (2021), the next generation of recommender and decision-support systems should be viewed as policies that decide what interventions to make in order to optimize a desired outcome. These policies and their associated managerial interventions are central to the firm's strategic decision-making.

Suppose the knowledge is a manager's mental model about how users of the firm's products/services respond to the recommendations made by AI. As an example of ML's time specificity, consider how the video streaming platform Netflix wins "moments of truth" with its recommender system. Within a few seconds, the personalized recommendation of items has to keep a user engaged and prevent the user from switching to an alternative entertainment option, as reported by Netflix researchers (Gomez-Uribe and Hunt, 2015: 13.6). Personalized recommendation has time specificity because new items are added, the user picks up new interests, popularity of items trends temporarily, and the data that the system uses for recommendations change over time (Basilico and Raimond, 2017). The knowledge about recommendations' effect on user response is specific to the time during which users interact with items. The importance of time has been reported by AI researchers at Netflix (Steck et al., 2021). The researchers reveal that adding time as contextual information to user-item interaction data—by incorporating raw, continuous timestamps indicating the time when the user played a video in the past,

along with the current time when making a prediction—is a key reason why Netflix's deep learning algorithm has significant improvements in performance over other ML techniques. Using data of user-item interactions, Netflix researchers build mathematical representations of users and items, predict each individual user's interest in each item available at the time, and generate on-demand recommendations of next items that are personalized to the user based on time-specific causal knowledge about user response. The new technologies change the way the firm is organized and how it competes, as time-specific causal knowledge augments the firm's strategic decision-making.

The subsequent sections of our paper are organized as follows. First, we introduce the concept of time-specific tacitness of AI-based knowledge and explain how this new concept leads to a re-evaluation of existing strategy theories on the limits of firm growth. Specifically, we examine the relevance of the new concept to the growth of digital platforms, which are firms that have disrupted a wide variety of industries by creating new transactions for goods, services, and information across multiple sides of a market (e.g., consumers, producers, and advertisers). Then, we theorize how the time specificity of causal knowledge enhances our understanding about what factors limit the scale, scope, and boundary of digital platforms. Finally, we analyze in what ways theories of firm growth can be revised and revitalized with the concept of time-specific tacitness of AI-based knowledge.

TIME-SPECIFIC TACITNESS OF AI-BASED KNOWLEDGE

The paradox of tacit knowledge: A challenge to the development and deployment of AI

Tacit knowledge, Michael Polanyi's (1966) well-known idea stating that individuals appear to know more than they can explain, is a key construct in existing strategy theories that juxtapose knowledge transfer and competitive imitation as blades of the same scissor, a paradox raised by Winter (1987). Polanyi argued that expertise, or a high degree of skills, is a precondition for articulate knowledge in general, and scientific knowledge in particular. An important part of expertise is tacit. Philosopher Hubert Dreyfus argued that an important part of the expert knowledge is tacit and therefore cannot be articulated and incorporated in a computer program. Hubert and Stuart Dreyfus's (1986) model of skill acquisition in the

development of expert systems posits that systems that enable a computer to simulate expert performance (for example medical diagnostics) are not able to capture the skills of an expert performer.

Extending Polanyi's idea, a theory of organizational knowledge, as advanced by Kogut and Zander (1992), examines when efforts by a firm to grow by the replication of its technology may inadvertently elevate the probability of imitation by competitors. The theory maintains that technology is less easily transmitted and replicated outside the firm, because no organizing principles exist to efficiently transfer and redeploy tacit knowledge. By contrast, inside the firm, the firm provides a social community in which individual and social expertise is transformed into economically useful products and services. Coded knowledge is alienable from the individual who wrote the code, although the firm may codify knowledge, to a certain extent, into a set of identifiable rules and procedures. The persistence and regularity in the organizing of social relationships and interactions is why tacit knowledge is more easily transferred inside the firm.

Tacit knowledge poses a challenge to the inter-disciplinary literature on AI governance that studies the ethical, legal and technical challenges in the development and deployment of AI. The AI workforce appears to know more than they can explain. Humans are involved in the ML training and testing process, but ML does not need the categorizations of our world as prescribed by humans to infer the underlying decision rules from historical data based on statistical analysis. However, quite different from Polanyi's well-known idea, here knowledge is tacit in the sense that the AI workforce built a system to imitate/simulate expert performance but can see only the system's output. The AI workforce cannot easily explain to users why a specific item or diagnosis was recommended, especially when deep neural networks employ hidden weights and activations that are generally noninterpretable, thus limiting explainability.

As an example, a tumor-detection system can diagnose more accurately than radiologists regarding the likelihood of cancer being present, as reported by McKinney et al. (2020). However, the AI workforce cannot explain the inferred decision rules because they cannot be directly observed or manipulated. The statistical inference is made with black box models and sophisticated algorithms for

deep learning. For instance, artificial neural networks store models as weights that do not have any correspondence to real-life objects. Moreover, the system's expertise is limited to a narrow, specific domain. For instance, the tumor-detection system cannot have a hunch about a peculiar shape of a tumor. A peculiar shape reflects some extenuating circumstances of the patient, so it is outside the space of the data the system was trained on. The system may detect a pattern in image pixels, but it will not be able to work reliably in unfamiliar situations. ML is fundamentally a form of context-dependent statistical inference and therefore its application is limited to a narrow, specific domain.

Given the limits of ML and the AI workforce, the crucial question is whether the firm's management can explain the reasons behind the predictions or recommendations made by AI: How is the system developed and deployed? In which domain of application? To what ends? With what benefits, and to whom? And with what risks? The firm's management faces pressures from the board of directors, shareholders, and regulators when the use of AI is seen as biased against certain groups. The management's explanations are important in designing interventions for preventing and correcting biases in predictions. For example, Amazon discontinued an AI recruiting tool because the system taught itself that male candidates were preferable (Dastin, 2018). Yet, as argued by Cowgill and Tucker (2019), algorithm users (principal) and algorithm developers (agent) have asymmetric ability to evade responsibility for mistakes, so if self-serving interests lead to the avoidance of responsibility, then AI applications cannot be trusted.

Time specificity of causal knowledge in digital platforms

GDPR regulations on data collection and storage, the protection of data security, and the concerns about privacy and transparency have led the AI workforce to take advantage of the temporal features of the data in session-based recommender systems. The temporal dynamics and sequential patterns in user attention and item evolution within a current session of user-item interactions are modeled with deep neural networks (see Zhang et al., 2019 for a review). Although the importance of time and temporal dynamics is disclosed through academic publications, competitors cannot easily reproduce the time-specific causal

knowledge, even if some of the firm's prediction technology is copied or reverse-engineered by competitive imitation. The creation of such knowledge results from the firm's intricate process of making predictions and designing interventions, which is deeply rooted in the firm as a social community and therefore not easily transferred by competitive imitation. Yet, the system can be explained by diagnosing the causal effect of personalized recommendations on user response such as satisfaction, addictive behaviors, probability of subscription cancellations, and complaints about unfair treatments.

We submit that strategy theories on the limits of firm growth need to be revised and revitalized by incorporating time specificity of causal knowledge. In proposing the revision and revitalization, we focus on digital platforms and their use of recommender systems. Recommender systems are one of the most successful applications of artificial intelligence (Jannach et al., 2019, 2021; Quandrana et al., 2018; Zhang et al., 2019). Matching users and items with a recommender system, a digital platform increases item availability on one side, attracts a large user base on another side, and monetizes from the knowledge about user response to personalized recommendations and experiences. Platforms, as Kenney, Bearson, and Zysman (2019) argued, "are an emblem and embodiment of the digital era just as factories were of the industrial revolution."

We are particularly interested in digital platforms because their growth is reshaping the playing field upon which competition and entrepreneurship take place. The growth of digital platforms is increasingly subject to regulatory scrutiny. Anti-trust violations, AI bias and algorithmic harm have alerted public authorities and standards bodies (Cutolo and Kenney, 2021). The regulations are intended to ensure a "fair, transparent and predictable business environment for smaller businesses and traders on online platforms" (European Commission, 2018, 2019). For instance, the EU regulation requires that platforms provide an account of the main factors used in their online ranking systems. Yet, disclosing an ML model would be revealing a trade secret (Rudin, 2019). Therefore, the revision to strategy theories that we propose focuses on incorporating time specificity of causal knowledge in the theories under a changing institutional environment where explainable AI has implications for competitive imitation.

As we incorporate time specificity of causal knowledge in strategy theories on the limits of firm growth, we connect how a digital platform generates knowledge to how the platform monetizes from the knowledge. The monetization could be fees for accessing and transacting on the platform (e.g., users pay subscription fees, app developers pay commissions, campaign sponsors pay advertisement and marketing service fees). The revision that we propose focuses on the rate at which the platform's management is capable of growing user engagement by making predictions and designing interventions about the probability of a user responding to on-demand personalized recommendations.

Another way to monetize from the knowledge could be fees from providing AI as a service that helps clients understand how users of the clients' products/services respond to personalized recommendations and experiences. Providing AI as a service faces similar challenges as the transfer of tacit knowledge when the transfer occurs outside the firm. Monetizing from providing AI as a service, however, invokes a conflict of responsibility between the clients who are held responsible for the quality of predictions and the service provider who is not. In her critique of black box ML models for high stakes decisions, Rudin (2019) argued "the fact that the model was complicated and proprietary allowed the company to profit from it." We focus our analysis on the time specificity of causal knowledge, not the transaction between AI-as-a-service provider and clients, so the theoretical revision is parsimonious.

Monetization could also be revenue from fulfilling a user's request through the platform's integration of personalized recommendations with distribution and delivery of a physical item. However, monetization that requires integration with non-scale-free assets for physical distribution and delivery faces capacity constraints of logistics and operations, in addition to the time specificity of user response to personalized recommendations and experiences. We focus our analysis on time specificity as the limiting factor, not the capacity constraints of logistics and operations, so the theoretical revision is parsimonious.

In revising and revitalizing strategy theories on the limits of firm growth, we focus on how the time specificity of causal knowledge enhances our understanding about what factors limit the scale, scope, and boundary of the firm. The temporal aspect that we emphasize acknowledges a key observation we make about digital platforms. That is, users interact with items online, including the items that were

posted/generated by other users in earlier time period. The contemporaneous interaction in real time means that a digital platform's rate of growth has a tempo completely different from a traditional firm's rate of growth, which is the change in firm size typically measured in assets or headcount. The online contemporaneous interaction also means that the knowledge gained from offline experiments may not apply. AI researchers at Netflix (Steck et al., 2021) reported that the performance improvement that is observed offline using historical data would sometimes disappear or, in rare cases, result in worse performance when the recommendations are presented to users in an A/B-test online. This suggests that correlates of performance are not causes of performance.

The causal aspect that we emphasize highlights another key observation we make about digital platforms. That is, users are subjects of the platform for data collection and experimental studies. Platforms collect data about users and how they interact with items. Platforms conduct experiments to study how users respond to personalized recommendations. Platforms train the ML algorithms with the data they collect and revise the algorithms based on the results of the experimental studies.

Strategic decision-making with time-specific causal knowledge

Data-augmented decision-making tools have changed the way managers make decisions. They now rely more on data and less on intuition (Brynjolfsson and McElheran, 2016). These tools enable managers to expand the search space of existing knowledge (Wu et al., 2020). However, causal knowledge is required to answer "what-if" questions in decision making. Causal knowledge is a mental model that link actions to consequences.[1] Yet, ML cannot foresee future consequences as humans can (Balasubramanian et al., 2020). Humans exercise judgment in assessing causal effects about where, how and why something happened. The causal knowledge that humans derive is used to generate and evaluate alternatives for decision making. The implication for the firm is that there has to be a good understanding of and

1 The idea that managers operate on the basis of inaccurate information is one of the hallmarks of bounded rationality; according to Simon (1997: 17), "bounded rationality [...] assumes that the decision maker [...] has egregiously incomplete and inaccurate knowledge about the consequences of actions."

explanation for where the data come from, what has influenced the data, and the causal relation between input and output data (Asatiani et al., 2020: 270).

We add to the importance of causal knowledge by highlighting the need to make accurate predictions in a dynamic environment where the temporal aspect connects predictions to decisions. Most commonly used ML algorithms, including decision trees, support vector machines, and deep learning, rely on correlations between inputs and outputs and are able to make accurate predictions only in a static environment. The algorithms do not generate causal knowledge. As AI pioneer Judea Pearl (2019) argued, "The dramatic success in machine learning has led to ... increasing expectations for autonomous systems that exhibit human-level intelligence. These expectations have, however, met with fundamental obstacles that cut across many application areas. Machine learning researchers have noted current systems lack the ability to recognize or react to new circumstances they have not been specifically programmed or trained for."

In a dynamic environment, time plays a key role in personalized recommendations, because users, items, and systems change over time. Time specificity is not as critical when using only historical data in a static environment. When recommendations are based on historical co-occurrence, there is no time specificity and so one cannot anticipate the changes. An item that is new to the system may start cold, but user interest in the item may change and fluctuate with temporal trends such as external events and seasonality. The recommender system also changes because feedback loops, where users are influenced by the output of the system, cause the data that the system uses to change over time. The feedback loops make it difficult to tease apart the cases where a user chooses an item because the item was displayed prominently and the cases where a user chooses an item independently. Different components of the system may change and thus affect the data that are used by other components. Designing experimentation that can handle time and infer causality is a key research area at Netflix (Basilico and Raimond, 2017).

The firm's intricate process of making predictions and designing interventions

As we submitted earlier, the creation of time-specific causal knowledge results from the firm's process of making predictions and designing interventions. As an illustration, we describe the process with how Netflix improved its key task of personalized ranking of movies and TV shows with deep learning models. AI researchers at Netflix conducted a series of online tests and offline analyses to understand why offline performance (when evaluated on held-out historical data) is not reflective of online performance (when evaluated in an A/B-test where the recommendations are presented to users) when trying deep-learning models (Steck et al., 2021).

"In the early 2010s, deep learning was taking off in the machine-learning community fueled by impressive results on a variety of tasks in different domains including computer vision, speech recognition, and natural language processing (NLP). At that time there was a stir in the air within the recommender-systems research community: Will the wave of deep learning also wash over recommenders to deliver tremendous improvements? As with many others, we at Netflix were intrigued by this question and the potential of deep learning to improve our recommendations. While the answer is now quite clear that deep learning is useful for recommender systems, the path to understand where deep learning is beneficial over existing recommendation approaches was an arduous one. This is evidenced by how many years it took for such methods to get traction in the research community."

Steck and colleagues observed that, if a deep-learning model is given the wrong problem to solve, it will solve it more accurately than less powerful models would. So, a major challenge is to figure out how to train on short-term behavior (e.g., clicks or plays) with an objective of optimizing long-term behavior (e.g., user satisfaction). Short-term behavior can be quite noisy in the sense that subtle changes in the definition of the (short-term) training objective can lead to big changes in the produced recommendations. Another challenge is distribution mismatch. "This is in general true whenever machine learning models are deployed in the real world, that is, the data which are used to train machine-learning models are not reflective of the population for which the model will be used. Covariate shift is a concrete example of distribution mismatch in which the distribution of input features is different between the training data and the real world. Traditional techniques to fix distribution mismatch like importance

sampling have been shown to be less effective with powerful deep-learning models" (Byrd and Lipton 2019).

Steck and colleagues also highlighted interventions for fairness and explainability. "When a deep-learning model (or any machine-learning model) is deployed, we need to be careful of how it may treat real-world entities (in the case of Netflix, members and videos for example), and whether there are any unintentional biases that cause the model to treat some entities in an unfair way. It is again related to the issue of offline–online mismatch as it may not be possible to easily evaluate a model from a fairness perspective as we may not have the appropriate offline evaluation data. A simple example is a model doing well for the majority of the data and poorly on a minority. We found techniques like LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017) and Integrated Gradients (Sundararajan, Taly, and Yan 2017) to be particularly helpful in explaining deep-learning models."

Factors limiting digital platforms' rate of growth: Revision and revitalization of strategy theories

We approach the revision and revitalization of strategy theories by examining the conditions under which the firm's knowledge has external validity. Specifically, we examine user/item/session heterogeneity, domain heterogeneity, and system complexity as the three conditions that determine the validity of the firm's experiment results and their applicability in new contexts. We connect these three conditions that constrain ML algorithm's applicability outside the context for which it is trained to the factors limiting the time-specific causal knowledge's applicability across organizational units, market segments, and industry sectors. We present in Table 1 three conditions that limit digital platforms' rate of growth by incorporating time specificity of causal knowledge along three dimensions of growth: scale, scope, and boundary. Each condition constrains ML algorithm's applicability outside the context for which it is trained.

[TABLE 1]

<u>Scale</u>

For scale, user/item/session heterogeneity limits the applicability of the time-specific causal knowledge as digital platforms grow their users, items and user-item interactions. Adding new users, new items, and

new sessions of user-item interactions in a domain may change the accuracy and cost of predictions when making on-demand personalized recommendations. Each session is composed of multiple user-item interactions in a continuous period of time, ranging from several minutes to several hours. Session-based recommender systems take each session as the basic input unit. These systems generate more accurate and timely recommendations by capturing a user's short-term preference from her recent sessions and the change in preferences from one session to another.

The time-specific causal knowledge can be used to increase the scale of a digital platform when the knowledge can be generalized to additional users, items, and user engagement with items on demand. However, the scale of a digital platform is bounded by ML, which is trained with user-item interaction data that are time specific (e.g., response to on-demand personalized recommendations). Outside the training data, the knowledge has not external validity. Generalization across populations (e.g., differences in user demographics) and time (e.g., differences between sessions of user-item interactions) is a fundamental problem in causal inference. When heterogeneous users are added, the accuracy of predictions decreases and the cost of making predictions increases.

As a limiting factor of scale growth, user/item/session heterogeneity is intellectually connected to the existing strategy theory on the role of knowledge replication in the rate of growing the scale of operations. Winter and Szulanski (2001) linked replication to scale-free knowledge, while Knudsen, Levinthal, and Winter (2014) explained replication errors when scaling. We build on both papers by theorizing the types of errors as the microfoundation in linking replication to AI-based knowledge. Whereas Knudsen et al. (2014) highlighted error-prone transmission, or replication, of firm-specific knowledge, time specificity was not theorized as a source of error in explaining why the rate of growing scale is limited. Our emphasis on time specificity is also intellectually connected to time compression diseconomy, which is a central strategy construct. Pacheco-de-Almeida and Zemsky (2007) as well as Wibbens (2021) formalized the challenges of time compression diseconomies. Our paper extends this central construct by theorizing time specificity as a source of error in the AI setting. Prediction errors—

bias and variance—were not theorized in the existing literature either, as the traditional industry context did not involve prediction technology.

Scope

For scope, domain heterogeneity limits the applicability of the time-specific causal knowledge as digital platforms grow their domains of application. One domain may complement or conflict with another domain when making on-demand personalized recommendations. Recommender systems focusing on a single domain tend to suffer problems of data sparsity and item cold-start that make it hard to model user preferences accurately and efficiently, because the data are restricted to a small fraction of past transactions (Li and Tuzhilin, 2020). By contrast, recommender systems covering multiple domains can transfer learning from one domain to another, if a common knowledge structure defines domain relatedness and user preferences have symmetrical correlation between source domain and target domain for cross-domain learning transfer (Khan et al., 2017; Pan et al., 2010).

Time specificity may limit the scope of a digital platform when the knowledge cannot be generalized to new domains. Continuing with the example of Netflix's recommender system for the ease of exposition, potential new domains may add music and podcast as items for entertainment. Could new domains be advertisements, news and infomercials? Could new domains extend beyond digital content that is produced by professional studios to include user-generated content? Could new domains venture farther into financial products (e.g., loans, investments, insurance policies), physical products (e.g., grocery), services (e.g., ride sharing), providers (e.g., personal fitness, healthcare), medical treatments (e.g., personalized medicine), driving directions, romantic dates, college applicants, job applicants, etc.?

Also, time specificity limits scope when the recommender system does not improve prediction accuracy or reduce cost by learning from the data that are aggregated across domains. When heterogenous domains are added, having knowledge from another domain may not create more value (higher accuracy at lower cost of prediction) for time-specific personalized recommendations. Cross-domain learning requires a common knowledge structure that defines domain relatedness and user preferences that have symmetrical correlation between source domain and target domain. If two users have similar preferences

in one domain, their preferences are assumed to be similar in other domains as well (e.g., music, books, movies). If the assumption holds, user preferences can be transferred across domain simultaneously. Having more than one domain at a time improves the recommendations through a better understanding of user preferences and the simultaneity of learning transfer across domains.

Our emphasis on time specificity contrasts with the existing strategy theory on a firm's horizontal scope by emphasizing that domain data are sources of related diversification. Domain relatedness is different from either input relatedness (supply-side factors including raw materials, technological component and human resource) or market relatedness (demand-side factors including user profile and geography). Existing strategy theory on horizontal scope made a connection to AI, with a focus on dominant logic (Prahalad and Bettis, 1986; Bettis and Prahalad, 1995). The research on dominant logic, however, was cast during the time when AI was primitive and data were limited. We revive the link between AI and horizontal scope, by providing the richness available only nowadays.

Our emphasis on time specificity also contrasts with the existing strategy theory on vertical scope. Digital firms tend to be both narrow in their vertical scope and large in their scale. One reason supporting this empirical observation is that, as argued by Giustiziero, Kretschmer, Somaya, and Wu (2022), scalability—how the value derived from a firm's resource bundle in a focal activity changes as the size of the bundle increases—affects the firm's opportunity costs of integration. Integration requires allocating resources to multiple value-adding activities, rather than using them more intensively to grow within the focal activity. When a firm's resource bundle is scalable, it is more likely to pursue "hyperspecialization" and "hyperscaling" simultaneously. It is more likely to outsource value-adding activities when they require resource bundles that entail significant opportunity costs. Our paper adds to this research by specifying the conditions under which digital platforms, a special type of digital firms, are limited in growing both scale and scope.

Boundary

For boundary, system complexity is a condition that determines the validity of the firm's experiment results and their applicability in new contexts. System complexity increases when knowledge components interact in interdependent ways. As digital platforms grow the components that are interdependent in the system, time specificity in terms of synchronization and inter-operability in the ML ecosystem may limit the boundary of a digital platform. For instance, the user interface of a recommender system can be a virtual assistant, which is an AI system interacting with users via voice recognition, face recognition and automated chatbot. A virtual assistant answers a user's questions and carries out a user's request such as making shopping lists based on the user's preferences, previous choices, and behavioral patterns. Using speech recognition, natural language processing, and robotic process automation, the system interacts with users in real-time, makes predictions about user behavior in context, and learns from each interaction. When a virtual assistant interconnects a recommender system, the boundary of a digital platform encompasses two interdependent knowledge components: the knowledge about how users respond to personalized recommendations is combined with the knowledge about how users respond to personalized experiences created by a virtual assistant. The personalized recommendations and experiences are time specific and need to be synchronized.

In addition to the interdependence between knowledge components, the interdependence between hardware and software in an ML ecosystem also has time specificity. Similar to a kernel in an operating system, an ML model requires an ecosystem of software and hardware beyond the model itself. An ecosystem of tools, libraries, community resources and professional support for AI workforce, especially for the training and deploying of deep neural networks, has emerged for recommender systems. Software modules available as cloud-based computing (e.g., inter-operable tools such as Tensorflow and Keras) can be combined to create complex systems of neural structures and build composite recommenders.[2] A hardware infrastructure can also be combined with software modules to accelerate the training and deploying of large-scale deep learning recommender systems. For example, NVIDIA Merlin™ is a deep

[2] https://www.tensorflow.org/recommenders/examples/basic_retrieval

neural network training framework that is used to predict a user's next action within a short time period, particularly for anonymous users or when users' interests are contextual and change within a session.[3] For session-based recommender systems, time specificity, in terms of synchronization and inter-operability, imposes a limit on what can be combined for the growth of a digital platform.

Our emphasis on time specificity in terms of synchronization and inter-operability as a source of limit to the firm's boundary is intellectually connected to the foundational research on system complexity (Arrow, 1974; Baldwin and Clark, 2000; Brusoni, Prencipe, and Pavitt, 2001; Sanchez and Mahoney, 1996; Simon, 1969). While the foundational research partitions system architecture spatially, our paper makes the partition temporally. In the foundational research, modularity is the main structural feature of system architecture, so the rate of firm growth is limited by how quickly components can be interconnected. Building on the foundational research, we submit temporal partition as another way modules of a complex system interconnect. Linking organizational complexity and AI is a budding and growing literature (e.g., Raj and Seamans, 2019; Shrestha, He, Puranam, and von Krogh, 2020) that we are joining.

In our paper, synchronization and inter-operability highlight the temporal aspect of interconnection. The more the firm simultaneously interconnects its AI with software (e.g., developing the firm's AI with the cloud-based Tensorflow) and hardware (e.g., deploying the firm's AI with neural network boosted with specialized graphics-processing-units [GPU]) in the business ecosystem, in which the firm partners with software and hardware providers, the higher the rate of growing its AI's boundary. When an ecosystem of software and hardware is available as general-purpose technologies that level the playing field, the firm can grow faster than its competitive rivals, who also can partner with the same software and hardware providers simultaneously, if the firm generates time-specific AI-based knowledge.

CONCLUSION

[3] https://developer.nvidia.com/nvidia-merlin

In this paper, we introduce the concept of time-specific tacitness of AI-based knowledge and theorize how the time specificity of causal knowledge enhances our understanding about what factors limit the scale, scope, and boundary of the firm. As a limiting factor of scale growth, user/item/session heterogeneity is intellectually connected to the existing strategy theory on the roles of knowledge replication and time compression diseconomy in the rate of growing the scale of operations. As a limiting factor of scope growth, domain heterogeneity is intellectually connected to the existing strategy theory on dominant logic for horizontal scope and that on hyperspecialization for vertical scope. As a limiting factor of boundary growth, system complexity is intellectually connected to the existing strategy theory on modularity, which is the main structural feature of system architecture that defines which components of the software and hardware ecosystem are inside versus outside the firm's boundary.

The concept that we introduce emerge from our observations of how new technologies (AI and digitalization) change the way firms are organized as digital platforms and compete with tacit knowledge. These new technologies cause us to re-evaluate existing strategy theory on the limit of firm growth with a focus on the scale, scope, and boundary of the firm. The re-evaluation leads to a revision and revitalization of theories that encompass central constructs in strategy research: knowledge replication, time compression diseconomy, dominant logic, hyperspecialization, and modularity. The time specificity of causal knowledge that is inferred with AI is the key in revising and revitalizing these existing theories. The revision and revitalization focus on strategic decision-making. By contrast, the existing literature on AI in the field of strategic management focuses on the functional aspects of management, featuring operations, marketing, talent acquisition, etc. Our focus on strategic decision-making has not only theoretical implications, but also managerial implications, where managers can view AI, particularly recommender systems, as decision-support systems. Time-specific causal knowledge is central to the firm's strategic decision-making as it informs whether and how to intervene. Managers can decide what interventions to make in order to optimize a desired outcome.

Our central thesis that creating time-specific causal knowledge about how to intervene with AI requires firm-specific process of making predictions and designing interventions is connected to three

broader bodies of literature. One body of literature is on firm knowledge and capabilities (Kogut and Zander, 1992; Nickerson and Zenger, 2004). Our paper extends this body by suggesting why and how causal inference might emerge as a type of organizational capability. Causal inference is a scale-free capability that has not been addressed in corporate strategy. Another body is the growing literature on the strategy of digital firms (Adner, Puranam, and Zhu, 2019; Giustiziero et al., 2022). Our paper extends this body by explaining the factoring limiting the scale, scope, and boundary growth of digital firms. The third body is on the limit of ML (Athey, 2017; Athey et al., 2020; Pearl, 2019). Our paper extends this body by linking fundamental obstacles that limit ML applications as well as the limitations of pure prediction methods to the firm's strategic decision-making.

REFERENCES

Adner, R., Puranam, P., and Zhu, F. (2019). What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strategy Science* 4(4), 253-261.

Allen, R., and Choudhury, P. (2022). Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion. *Organization Science*, 33(1), 149-169.

Arrow, K. (1974). *The Limits of Organization*. New York: Norton – Chap. 1 and 2.

Asatiani, A., Malo, P., Nagbøl, P. R., and Penttinen, E. (2020). Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive*, 19(4): 15.

Athey, S. C. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483–485.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020, May). The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings* (Vol. 110, pp. 80-84).

Balasubramanian, N., Ye, Y., and Xu, M. (2020). Substituting human decision-making with machine learning: Implications for organizational learning. Academy of Management Review, (ja) https://doi.org/10.5465/amr.2019.0470.

Baldwin, C., and Clark, K. (2000). *Design Rules: Volume 1. The Power of Modularity*. Cambridge, MA: The MIT Press. Chap. 1-3.

Basilico, Justin, and Raimond, Yves. (2017). Déja vu: The importance of time and causality in recommender systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems*. Page 342. https://doi.org/10.1145/3109859.3109922.

Bettis, R. A., and Prahalad, C. K. (1995). The Dominant Logic: Retrospective and Extension. *Strategic Management Journal*, 16(1), 5–14.

Brusoni, S., Prencipe, A., and Pavitt, K. (2001). Knowledge specialization, organizational coupling, and the boundaries of the firm: Why do firms know more than they make? *Administrative Science Quarterly* 46(4): 597-621.

Brynjolfsson, E., and McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5): 133–39.

Byrd, J., and Z. Lipton. (2019). What is the Effect of Importance Weighting in Deep Learning? In International Conference on Machine Learning (ICML).

Cutolo, D., and Kenney, M. (2021). Platform-dependent entrepreneurs: Power asymmetries, risks, and strategies in the platform economy. *Academy of Management Perspectives*, 35(4), 584-605.Choudhury, P., Starr, E., and Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41(8), 1381-1411.

Cowgill, B., and Tucker, C. E. (2019). Economics, Fairness and Algorithmic Bias. In preparation for *The Journal of Economic Perspectives*. SSRN.

Dastin, J. (2018) "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dreyfus, H. L., and Dreyfus, S. E. (1986). From Socrates to expert systems: The limits of calculative rationality. In *Philosophy and Technology* II (pp. 111-130). Springer, Dordrecht.

European Commission. (2018). Commission Staff Working document—Impact assessment and executive sum- mary accompanying the document proposal for a regulation of the European Parliament and of the Council on Promoting Fairness and Transparency for business users of online intermediation services. https://ec.europa.eu/digital-single-market/en/news/ impact-assessment-proposal–promoting-fairness- transparency-online-platforms

European Commission. (2019). February. Digital single mar- ket: EU negotiators agree to set up new European rules to improve fairness of online platforms' trading practi- ces. Press Release. Retrieved from http://europa.eu/ rapid/press-release_IP-19-1168_en.htm

Giustiziero, G., Kretschmer, T., Somaya, D., and Wu, B. (2022). Hyperspecialization and Hyperscaling: A Resource-based Theory of the Digital Firm. *Strategic Management Journal*. Special issue forthcoming.

Gomez-Uribe, C. A., and Hunt, N. (2015). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS),* 6(4), 1-19.

Iansiti, M., and Lakhani, K. R. (2020). *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*. Harvard Business Press, Boston, MA.

Jannach, D., and Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4), 1-23.

Jannach, D., Pu, P., Ricci, F., and Zanker, M. (2021). Recommender systems: Past, present, future. *AI Magazine*, 42(3), 3-6.

Joachims, T., London, B., Su, Y., Swaminathan, A., and Wang, L. (2021). Recommendations as treatments. *AI Magazine*, *42*(3), 19-30.

Kellogg, K. C., Valentine, M. A., and Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), 366-410.

Kenney, M., Bearson, D., and Zysman, J. (2019). "The Platform Economy Matures: Pervasive Power, Private Regulation, and Dependent Entrepreneurs." *BRIE Working Paper Series*. Accessed July 12, 2020. https://brie.berkeley.edu/sites/default/files/platform_economy_matures_final.pdf.

Khan, M. M., Ibrahim, R., and Ghani, I. (2017). Cross domain recommender systems: A systematic literature review. *ACM Computing Surveys (CSUR),* 50(3), 1-34.

Knudsen, T., Levinthal, D.A., and Winter, S.G. (2014). Hidden but in plain sight: The role of scale adjustment in industry dynamics. *Strategic Management Journal* 35(11): 1569-1584.

Kogut, B., and Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization science*, 3(3), 383-397.

Li, P., and Tuzhilin, A. (2020, January). DDTCDR: Deep dual transfer cross domain recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 331-339).

Lundberg, S. M., and S.-I. Lee. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).

McKinney, S.M., Sieniek, M., Godbole, V., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Nickerson, J.A., and Zenger, T. R. (2004). A Knowledge-Based Theory of the Firm—The Problem-Solving Perspective. *Organization Science* 15(6):617-632.

Pacheco-de-Almeida, G., and Zemsky, P. (2007). The timing of resource development and sustainable competitive advantage. *Management Science* 53(4): 651–666.

Pan, W., Xiang, E., Liu, N., and Yang, Q. (2010, July). Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 24, No. 1, pp. 230-235).

Pearl, J. (2019). The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Communications of the ACM*. 62(3):54–60.

Penrose, E. T. (1959/1995). *The Theory of the Growth of the Firm*, 3rd ed. Oxford University Press, Oxford, UK.

Phillips, P.J., Hahn, C.A., Fontana. P.C., Broniatowski, D.A., and Przybocki, M.A. (2020). Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology, U. S. Department of Commerce, August 2020. https://doi.org/10.6028/NIST.IR.8312-draft.

Polanyi, M. (1966). *The Tacit Dimension*, New York: Anchor Day Books.

Prahalad, C. K., and Bettis, Richard A. (1986) The dominant logic: A new linkage between diversity and performance. *Strategic Management Journal* 7(6): 485-501.

Quadrana, M., Cremonesi, P., and Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys (CSUR),* 51(4), 1-36.

Raj, Manav, and Seamans, Robert. (2019). Primer on artificial intelligence and robotics. *Journal of Organization Design,* 8(11): 1-14.
Ribeiro, M. T., S. Singh, and C. Guestrin. (2016). Why should I Trust You? Explaining the Predictions of any Classifier. In ACM Conference on Knowledge Discovery and Data Mining (KDD).
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215.
Sanchez, R., and Mahoney, J.T. (1996). Modularity, flexibility, and knowledge management in product and organization design. *Strategic Management Journal*, 17(S2): 63-76.
Shrestha, Y. R., He, V. F., Puranam, P., and von Krogh, G. (2020). Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize? *Organization Science* 32(3):856-880.
Simon, H. (1969). The architecture of complexity. In *The Sciences of the Artificial*. Boston, MA: MIT Press.
Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., and Basilico, J. (2021). Deep learning for recommender systems: A Netflix case study. *AI Magazine*, 42(3), 7-18.
Sundararajan, M., A. Taly, and Q. Yan. (2017). Axiomatic Attribution for Deep Networks. In International Conference on Machine Learning (ICML).
Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, 42(9), 1600-1631.
Wibbens, P. D. (2021). The role of competitive amplification in explaining sustained performance heterogeneity. *Strategic Management Journal* 42(10): 1769–1792.
Winter, S. (1987). Knowledge and Competence as Strategic Assets. in *The Competitive Challenge-Strategies for Industrial Innovation and Renewal*, D. Teece (Ed.), Cambridge, MA: Ballinger.
Winter, S., and Szulanski, G. (2001). Replication as Strategy. *Organization Science* 12(6):730-743.
Wu, L., Hitt, L., and Lou, B. (2020). Data analytics, innovation, and firm productivity. *Management Science,* 66(5): 2017–2039.
Zhang, S., Yao, L., Sun, A., and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, *52*(1), 1-38.

TABLE 1: Factors Limiting Digital Platforms' Rate of Growth in Scale, Scope, and Boundary

Factors Limiting Digital Platforms' Rate of Growth in Scale, Scope, and Boundary	**Sources of Knowledge Limits**	**Time Specificity of Causal Knowledge**
Scale **Deepening knowledge within domain**	User heterogeneity Item heterogeneity Session heterogeneity	Adding new users, new items, and new user-item interactions in a domain may change the accuracy and cost of predictions when making on-demand personalized recommendations.
Scope **Broadening knowledge across domains**	Domain heterogeneity	One domain may complement or conflict with another domain when making on-demand personalized recommendations.
Boundary **Combining knowledge as components of a complex system**	System complexity	One component of an AI system may complement or conflict with another component when making on-demand personalized recommendations.