EPFL

Thèse n° 10 037

Technology Strategy in Dynamic Environments: A Computational Analysis of the Automation of Routines, the Organization of AI, and the Evaluation of Technology Risk

Présentée le 29 août 2022

Collège du management de la technologie
Chaire de technologie et stratégies d'innovation
Programme doctoral en management de la technologie

pour l'obtention du grade de Docteur ès Sciences

par

Maximilian Wieland HOFER

Acceptée sur proposition du jury

Prof. D. Kuhn, président du jury
Prof. K. Younge, directeur de thèse
Prof. C. L. Tucci, rapporteur
Prof. Y. Rosokha, rapporteur
Prof. R. West, rapporteur

École polytechnique fédérale de Lausanne

2022

I dedicate this dissertation to my parents, Monika and Wieland, and my fiancée, Giulia.

Acknowledgements

I have benefited greatly from the support of many people during the EDMT doctoral program. First and foremost, I would like to thank my supervisor, Kenneth Younge. Ken's deep knowledge of the literature, his broad expertise with computational methods, and his general wisdom have formed the foundation of my education at EPFL. He guided me through complex problems with clarity and patience, and taught me lessons that I will fondly remember. I would like to thank Ken for collaborating with me on Chapter 2 and Chapter 4 of the dissertation. Ken, I owe you my enduring gratitude.

Second, I would like to acknowledge the vital support of my dissertation committee: Professor Daniel Kuhn, Professor Christopher Tucci, Professor Robert West, and Professor Yaroslav Rosokha. Moreover, I would like to thank the participants of the 2019 ETHZ and UZH Zurich Text as Data workshop for their feedback on Chapter 4 of the dissertation.

Third, I am profoundly grateful for my friends and colleagues who supported me along the way. Mike, thank you for brightening our office with your humor. George, thank you for the enjoyable discussions on research and teaching. Gabriel, thank you for your academic mentoring and meticulous feedback. Raphi, thank you for your friendship and kindness that have lightened up my days.

Moreover, I would like to wholeheartedly thank my family for supporting me throughout the highs and lows of the doctoral program. Thank you, Monika, Wieland, and Alexander, for tirelessly encouraging me when I could not see the light at the end of the tunnel while celebrating the milestones along the way. You mean the world to me.

Finally, I will forever be grateful to Giulia, my fiancée and best friend, for her unlimited support, thoughtfulness, and love. Thank you for always standing by my side.

Abstract

In this dissertation, I develop theory and evidence to argue that new technologies are central to how firms organize to create and capture value. I use computational methods such as reinforcement learning and probabilistic topic modeling to investigate three topics: the automation of routines, the organization of artificial intelligence (AI), and the evaluation of technology risk. Overall, I argue that new technologies are not a panacea for the firm but require deliberate strategic planning to manage the potential downsides of myopic automation, AI interdependencies, and the disclosure of technology risks.

In the first essay, I argue that while automation can increase productivity by reducing the costs of coordinating individuals, the automation of routines can also incur an indirect opportunity cost due to slow adaptation to environmental change. I develop a reinforcement learning simulation to model the impact of automation on the returns from the division of labor in dynamic environments and to show how automation incurs opportunity costs through lost learning and slow adaptation. Moreover, automation can be suboptimal when it brings about myopic behavior, i.e. high returns from the division of labor in the short term, but negative returns in the long term. Given the simulation results, I argue that firms need dynamic routines to simultaneously balance learning and automation. I open-source the simulation platform as *OrgSim-RL* on GitHub.

In the second essay, I argue that a data-driven culture – what I define as a *Data Clan* – can help to coordinate complex interdependencies between AI components within a firm. I analyze in-depth semi-structured interview data with a hierarchical stochastic block model (hSBM) and hand-coding to find that managers focus primarily on building a strong culture and establishing high-quality data assets when allocating resources to AI initiatives. Given the results, I inductively develop implications for theory and argue that the emergence of a *Data Clan* can be a governance

mechanism to reduce coordination frictions and build a competitive advantage in the age of AI.

In the third essay, I argue that investors require a higher initial return to take on more technology risk disclosure during an IPO. I quantify the magnitude of disclosed risk and the risk disclosure topics based on a latent Dirichlet allocation (LDA) topic model of IPO prospectus text and find a *return-for-risk* association between text-based technology risk disclosure and underpricing. The study also finds evidence that owning granted patents is associated with a lower *return-for-risk* association, suggesting that intellectual property allows the disclosure of risk without losing the competitive advantage. I open-source the code for quantifying risk disclosure as *RiskyData-LDA* on GitHub.

In summary, this dissertation develops theory and finds evidence across three essays to argue that leveraging new technologies requires deliberate strategic planning to manage potential downsides of new technologies, such as the opportunity costs of automation, coordination costs, and costs associated with raising capital. The results suggest three mitigating solutions: dynamic routines to balance learning and automation, a *Data Clan* to improve coordination, and disclosure through patents to reduce underpricing.

Keywords: Strategic management, technology, automation, artificial intelligence, risk disclosure, dynamic environments, organizational learning, initial public offerings, reinforcement learning, topic modeling

Résumé

Dans cette thèse, je développe des analyses théoriques et des indications qui proposent que les nouvelles technologies soient au cœur de l'organisation des entreprises pour créer et capturer de la valeur. J'utilise des méthodes informatiques telles que l'apprentissage par renforcement (*reinforcement learning*) et la modélisation thématique (*topic modeling*) pour analyser trois sujets: l'automatisation des routines, l'organisation de l'intelligence artificielle (IA) et l'évaluation des risques technologiques. Dans l'ensemble, je soutiens que les nouvelles technologies ne sont pas une panacée pour les entreprises, mais qu'elles nécessitent une planification stratégique délibérée pour gérer les inconvénients de l'automatisation myope, les interdépendances d'IA et la divulgation des risques technologiques.

Dans le premier essai, je soutiens que si l'automatisation peut augmenter la productivité en réduisant les coûts de coordination des individus, les routines d'automatisation peuvent également entraîner un coût d'opportunité indirect dû à une adaptation lente aux changements environnementaux. Je développe une simulation d'apprentissage par renforcement pour modéliser l'impact de l'automatisation sur les rendements de la division du travail dans des environnements dynamiques et pour montrer *comment* l'automatisation entraîne des coûts d'opportunité en raison de la perte d'apprentissage et de la lenteur de l'adaptation. En plus, l'automatisation peut être sous-optimale lorsqu'elle entraîne un comportement myope, c'est-à-dire des rendements élevés de la division du travail à court terme, mais des rendements négatifs à long terme. Compte tenu des résultats de la simulation, je soutiens que les entreprises ont besoin de routines dynamiques pour équilibrer simultanément l'apprentissage et l'automatisation. Je mets la plate-forme de simulation en libre accès sous le nom de *OrgSim-RL* sur GitHub.

Dans le deuxième essai, je soutiens qu'une culture centrée sur les données – ce que je définis comme un *Data Clan* – peut aider à coordonner les interdépendances complexes entre les composants de l'IA. J'analyse les données d'entretiens semi-structurés avec un *hierarchical stochastic block model* (hSBM) et un codage manuel pour constater que les managers se concentrent princi-

palement sur la construction d'une culture forte et l'établissement des ressources de données lors de l'allocation des ressources aux initiatives d'IA. Compte tenu de ces résultats, j'examine les implications inductives pour la théorie afin de soutenir que l'émergence d'un *Data Clan* peut être un mécanisme de gouvernance important pour réduire les frictions de coordination et de créer un avantage concurrentiel à l'ère de l'IA.

Dans le troisième essai, je soutiens que les investisseurs exigent un rendement initial plus élevé pour assumer davantage de divulgation des risques technologiques lors de l'introduction en bourse (*IPO*). Je quantifie l'ampleur du risque divulgué et les sujets de divulgation des risques sur la base d'un modèle thématique du texte du prospectus d'introduction en bourse. Je trouve une corrélation *return-for-risk* de la divulgation des risques technologiques basée sur le texte. L'essai trouve également que la possession d'un plus grand nombre de brevets délivrés est associée à une corrélation *return-for-risk* plus faible, ce qui suggère que la propriété intellectuelle formelle permet la divulgation des risques sans perdre l'avantage concurrentiel. Je mets en libre accès le code de quantification de la divulgation des risques sous le nom de *RiskyData-LDA* sur GitHub.

En résumé, cette thèse développe une théorie et découvre des indications dans trois essais pour soutenir que l'exploitation des nouvelles technologies nécessite une planification stratégique délibérée pour atténuer les inconvénients potentiels des nouvelles technologies, tels que les coûts d'opportunité de l'automatisation, les coûts de coordination et les coûts associés à la mobilisation de capitaux. Les résultats suggèrent trois solutions d'atténuation : des routines dynamiques pour équilibrer l'apprentissage et l'automatisation, un *Data Clan* pour améliorer la coordination et la divulgation de brevets pour accroître l'efficacité de la mobilisation de capitaux lors de l'introduction en bourse

Mots clés: Stratégie d'entreprise, technologie, automatisation, intelligence artificielle, divulgation des risques, environnements dynamiques, apprentissage organisationnel, introduction en bourse, apprentissage par renforcement, modèle thématique

Zusammenfassung

In dieser Dissertation argumentiere ich basierend auf der Entwicklung von Theorien und der Analyse von Evidenz, dass neue Technologien von zentraler Bedeutung für unternehmerische Prozesse sind. Ich verwende computergestützte Methoden wie bestärkendes Lernen (*reinforcement learning*) und probabilistische Themenmodellierung (*topic modeling*), um drei Aspekte zu analysieren: die Automatisierung von Prozessen, die Organisation künstlicher Intelligenz (KI) und die Evaluierung von Technologierisiken. Übergreifend argumentiere ich, dass neue Technologien kein Allheilmittel für Unternehmen sind, sondern eine aktive strategische Planung erfordern, um potenzielle Nachteile kurzsichtiger Automatisierung zu bewältigen, KI-Ressourcen zu koordinieren und Technologierisiken vorteilhaft offenzulegen.

Im ersten Artikel argumentiere ich, dass Automatisierung zwar die Produktivität steigern kann, indem sie die Kosten für die Koordination von Einzelpersonen senkt, automatisierte Prozesse jedoch auch indirekte Opportunitätskosten aufgrund der langsamen Anpassung an Veränderungen im Umfeld verursachen können. Ich entwickle eine *reinforcement-learning*-Simulation, um die Auswirkungen der Automatisierung auf die Erträge aus der Arbeitsteilung in dynamischen Umgebungen zu modellieren und zu zeigen, wie Automatisierung Opportunitätskosten durch verlorenes Lernen und langsame Anpassung verursacht. Zudem kann Automatisierung suboptimal sein, wenn sie kurzsichtiges Verhalten bewirkt, d.h. kurzfristig hohe Renditen aus der Arbeitsteilung generiert, langfristig aber negative Renditen. Angesichts der Simulationsergebnisse argumentiere ich, dass Unternehmen dynamische Routinen benötigen, um Lern- und Automatisierungsprozesse zu kombinieren. Ich stelle die Simulationsplattform unter *OrgSim-RL* auf GitHub frei zur Verfügung.

Im zweiten Artikel argumentiere ich, dass eine datenzentrierte Kultur – die ich als *Data Clan* definiere – helfen kann, komplexe Abhängigkeiten zwischen KI-Komponenten zu koordinieren. Ich analysiere Interviewdaten mit einem *hierarchical stochastic block model* (hSBM) und kodiere Unternehmensressourcen, um zu zeigen, dass Manager primär über die Allokation von Ressourcen für den Aufbau einer starken Unternehmenskultur und den Aufbau hochwertiger Datensätze sprechen.

Ich interpretiere die komplexen Wechselbeziehungen zwischen KI-Ressourcen und argumentiere, dass ein *Data Clan* ein wichtiger Governance-Mechanismus sein kann, um im Zeitalter der KI Koordinationskosten zu verringern und einen Wettbewerbsvorteil aufzubauen.

Im dritten Artikel argumentiere ich, dass Investoren einen höheren kurzfristig Ertrag erwarten, um mehr Technologierisiken bei einem Börsengang zu tragen. Ich quantifiziere das Ausmaß des offengelegten Risikos und die Risikothemen mit einem *latent Dirichlet allocation* (LDA) Themenmodell. Die Resultate zeigen eine *return-for-risk*-Korrelation von textbasierten Technologierisiken. Des Weiteren zeigt der Artikel, dass Patente die *return-for-risk*-Korrelation abschwächen können, was darauf hindeutet, dass formelles geistiges Eigentum die Offenlegung von Risiken ermöglicht, ohne den Wettbewerbsvorteil zu verlieren. Ich stelle den Code zur Risikoquantifizierung unter *RiskyData-LDA* auf GitHub frei zur Verfügung.

Zusammenfassend sollen die entwickelten Theorien und die analysierte Evidenz in dieser Dissertation zeigen, dass die Nutzung neuer Technologien eine bewusste strategische Planung erfordert, um potenzielle Nachteile neuer Technologien wie die Opportunitätskosten der Automatisierung, Koordinationskosten und Kosten im Zusammenhang mit der Kapitalbeschaffung zu mindern. Die Ergebnisse schlagen drei Lösungen vor: (1) dynamische Routinen, um Lernen und Automatisierung in Einklang zu bringen, (2) ein *Data Clan*, um die Koordination zu verbessern, und (3) Offenlegung von Patenten, um die Effizienz der Kapitalbeschaffung beim Börsengang zu steigern.

Schlagwörter: Strategisches Management, Technologie, Automatisierung, künstliche Intelligenz, Offenlegung von Risiken, dynamische Umgebungen, organisatorisches Lernen, Börsengänge, Verstärkungslernen, Themenmodellierung

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Arguably the most powerful firms today are not those with industry or resource positions but those with technology positions." — Furr (2021)

How can firms manage new technologies to create and capture value? In this dissertation, I explore this question by focusing on three topics – the automation of routines, the organization of artificial intelligence (AI), and the evaluation of technology risk – to develop new theory and insights to argue that technology[1] is a central consideration in how firms operate to create and capture value. Across the three essays, I investigate the impact of automation on a firm's returns from the division of labor in dynamic environments (first essay), the coordination of AI components in incumbent firms (second essay), and the disclosure of text-based technology risks when undertaking an initial public offering (IPO) (third essay). The results and theoretical developments have implications for strategic management research on dynamic capabilities in the age of automation (first essay), coordination mechanisms for organizing interdependent AI assets (second essay), and text-based risk disclosure at IPO (third essay). Moreover, this dissertation can help practitioners adopt an informed approach to managing new technologies in dynamic environments.

In this introductory section, I position my dissertation in the broader context of the strategic management of new technologies. Managing how firms can exploit technology is becoming increasingly important as more and more products, services, and business models directly depend on technology (Furr, 2021). However, I argue that technologies such as automation and AI are

[1] I refer to "technology" as a phenomenon put to use (Arthur, 2009). The internet, the smartphone, CRISPR, and artificial intelligence (AI) are all examples of such technologies (Furr, 2021).

not a panacea for firms but instead require deliberate strategic planning to capture the upsides while mitigating potential downsides. The interdependent and intangible nature of many technologies can complicate the exploitation of productivity benefits due to automation (first essay), the development of AI capabilities (second essay), and the disclosure of technology-related risks (third essay). The dissertation establishes the following insights. First, while recent research has examined a potential downside of automation in the equilibrium (Dogan and Yildirim, 2021), I find evidence that automated routines can incur *hidden* opportunity costs due to slow adaptation during nonlinear learning trajectories in dynamic environments. Firms benefit from dynamic routines that simultaneously balance learning and automation to mitigate opportunity costs. Second, I collect qualitative evidence and develop theory to suggest that a strong, data-centric organizational culture – what I define as a *Data Clan* – can reduce intraorganizational coordination costs; as such, a *Data Clan* can represent a valuable strategic asset for firms in the age of AI. Third, while existing research has investigated *general* risk disclosures (e.g., Loughran and McDonald (2013)), I develop new measures of text-based risk disclosure and find that technology-related risk disclosures are significantly positively associated with IPO underpricing in a *return-for-risk* association. The results show that patents can attenuate the *return-for-risk* association to suggest that intellectual property can allow disclosure without threatening competitive advantage.

New Technologies, Firms, and Environmental Change

As the following observations suggest, technology is central to the organizational processes for creating and capturing value. First, the most valuable public companies today, including Alphabet, Apple, Amazon, and Microsoft, directly depend on advanced digital technologies for their products and services (Brynjolfsson and McAfee, 2011). For example, approximately 81% of Alphabet's 2021 revenue, according to the company's 2021 annual report, comes from online advertising, which

builds on the Google AdSense software that matches ads with the most relevant audience (Vise and Malseed, 2006). Second, the composition of corporate investments in the US and Europe has shifted from tangible to intangible assets that often include technology-related assets such as technical know-how, custom software, and patents (Haskel and Westlake, 2018). Corrado and Hulten (2010) developed a new measure of intangible capital to find that total investments into intangible assets have exceeded investments into tangible assets in the US since approximately 1995. Today, Bailey et al. (2022) argue that new technologies enable new interdependencies within and beyond firm boundaries that shape all aspects of organizing and are, therefore, central to organizational scholars.

Two technologies of organizational relevance include automation (e.g., Autor (2015)) and artificial intelligence (AI) (e.g., Von Krogh (2018)). While robots have been able to automate repetitive physical tasks for a considerable period of time, new technologies can increasingly automate cognitive tasks previously reserved for the human domain (Autor et al., 2003; Brynjolfsson and McAfee, 2014). Consequently, Frey and Osborne (2017) estimate that the computerization of work could automate at least 40% of tasks across various occupations by 2030. Moreover, new technologies for implementing AI solutions can detect patterns in large-scale data to make predictions or decisions relevant to tasks such as hiring personnel, performing financial controlling, and scheduling complex logistics (Raj and Seamans, 2019). Recent technological breakthroughs across a range of complex learning problems as described in the 2022 AI Index Report by the Stanford Institute for Human-Centered AI (Zhang et al., 2022) suggest that AI shares characteristics with existing general-purpose technologies (GPTs) such as the steam engine or the internet (Cockburn et al., 2018). However, exploiting new technologies can be challenging for firms, as automation can generate unexpected costs (Dogan and Yildirim, 2021) and GPTs can complicate performance evaluation due to measurement delays (Brynjolfsson et al., 2021). In short, I argue that technologies such as

automation and AI occupy a central role in organizing processes, making the study of technology central to strategic management research (Bailey et al., 2022). At the same time, relatively little is known about the potential trade-offs and downsides that firms can face when adopting automation (Autor, 2015; Dogan and Yildirim, 2021), building AI capabilities (Von Krogh, 2018), or disclosing risks related to technological innovation at IPO (Loughran and McDonald, 2013).

Today, firms often operate in unpredictable and rapidly-changing environments. Due to the recent Covid-19 pandemic, for example, Bloom et al. (2021) found a shift in patent applications toward technologies that support working from home, and Bloom et al. (2020) argue that the pandemic might decrease longer-term total factor productivity due to diminished R&D expenditures and diverted managerial attention. In response to such fundamental changes, the organizational literature suggests that managers focus on coordination by mutual adjustment and reaction through feedback (March and Simon, 1958), re-configuring capabilities to adapt to the new demands of the dynamic environment (Teece et al., 1997), and engaging in search and learning activities to develop and maintain organizational capabilities (Zollo and Winter, 2002). In addition, recent developments toward explicitly integrating uncertainty into foundational strategic management theory (e.g., Furr and Eisenhardt (2021)) emphasize the importance of dynamic environments to management research.

Creating Value with New Technologies

The resource-based view (RBV) suggests that firms invest in creating and curating unique and valuable asset positions to implement a strategy (Barney, 1986b). RBV theory suggests that superior asset positions can establish competitive advantage and, consequently, enable superior financial performance (Wernerfelt, 1984; Barney, 1986b; Peteraf, 1993). Furr (2021) proposes that successful firms *today* are not those with industry positions (Porter, 1980) or resource positions (Barney,

1986b) but those with technology positions. Even though technology positions can become resource positions, Furr (2021, p.205) argues that jumping straight to a resource-based view is like developing a theory of vegetable markets without a theory of farming. To continue the analogy, the resource position that a farmer can establish in the vegetable market depends on its ability to leverage agriculture technologies such as algorithmically optimized planting, computerized quality assurance, and online platforms to purchase crops and sell fresh vegetables. More broadly, technology positions seem particularly central to how firms operate given that "emerging technologies have the potential to fundamentally shape all aspects of organizing" (Bailey et al., 2022). In short, one can view *technology strategy* as a firm's strategy for creating and capturing value with technology in an environment shaped by technology.

Strategically acquiring resources to accumulate valuable asset positions typically requires luck or superior insight (Barney, 1986b). One way to generate superior insight is to engage in deliberate learning to generate new knowledge (Zollo and Winter, 2002). However, despite superior knowledge, accumulating technology-related assets can be challenging due to asset erosion as a consequence of technological obsolescence (Dierickx and Cool, 1989), asset interdependence (Teece, 1986), and the immobility of invisible resources such as technical production skills or corporate culture (Itami and Roehl, 1991). Ex-post limitations to imitating a stock of interdependent and often intangible assets can enable firms with such asset stocks to gain superior returns; such limitations include impediments to asset accumulation (Dierickx and Cool, 1989), causal ambiguity (Lippman and Rumelt, 1982), and tacit knowledge (Polanyi, 1962). Taken together, I argue that accumulating and organizing technology positions is strategically important to create and capture value with new technologies but does not come without challenges.

Managers are central to implementing a strategy, in part due to their selection ability in allocating resources. For example, Mollick (2012) finds that variation among individual innovators

and managers impacts firm performance, using data from the electronic games industry. Similarly, Burgelman (1991) conducts a field study on Intel Corporation to find that a manager's role in selecting strategic initiatives can be important for enabling organizational change in a dynamic environment such as the semiconductor industry. The importance of managerial selection for firm performance and organizational change in technology-driven environments suggests that managers take a central role in shaping technology positions. Accordingly, Aguinis et al. (2022) argue for closely integrating managers when developing and analyzing propositions and when discussing potential implications of research results.

For firms preparing to raise capital through an IPO, it is essential that potential investors can value their technology positions as the firm might otherwise select a different path to raise capital. Often, technology-related assets such as patents, licenses, and databases are intangible (Itami and Roehl, 1991). Intangible assets can be difficult to evaluate due to various factors, including outdated accounting methodologies for assets associated with the computerization of the economy (Yang and Brynjolfsson, 2001). Consequently, the number of public firms might decrease as Kahle and Stulz (2017) observe in the period between 1975 and 2015, and, when relatively young firms decide to go public, potential investors tap into alternative sources of information such as text data (e.g., Loughran and McDonald (2013)). Whereas existing research has focused on the valuation effect of technology-related aspects such as R&D expenditures (Griliches, 1981), patents (Heeley et al., 2007), and commercialization strategies (Morricone et al., 2017), I investigate the disclosure of risks related to an IPO firm's technology asset stock and how investors might evaluate such text-based technology risk disclosures for firms with and without granted patents.

Computational Methods for Strategic Management

A secondary but complementary objective of this dissertation is to advance computational methods for strategic management research. I exploit computational methods in the dissertation for the following reasons. First, machine learning methods can detect patterns in unstructured data that researchers can then use in traditional econometric models (Mullainathan and Spiess, 2017). The framework of "Text as Data" describes the process of turning textual data into tabular data for further econometric analysis in the social sciences (Grimmer et al., 2022). Second, computational methods such as probabilistic topic models (Hannigan et al., 2019) can uncover nonlinear patterns in data (Choudhury et al., 2020) and enable the identification of novel perspectives of the data with minimal researcher-driven interpretation. Third, simulation approaches can be beneficial for developing management theory by modeling nonlinear interactions among variables of interest (Davis et al., 2007), especially when challenging empirical data restrictions exist (Zott, 2003). I create a computational representation of an automation routine to simulate how automation interacts with the division of labor, coordination between individuals, and the organizational learning process; such interactions can be highly nonlinear and challenging to investigate empirically. Finally, computational methods can generate artifacts, in the form of open-source code repositories, that facilitate replication (Ethiraj et al., 2016) and support future work in the field.

Summary of Results

In the first essay (Chapter Two), I argue that automating can incur a *hidden* opportunity cost due to its limited ability to learn and adapt. While automating tasks can reduce the coordination costs associated with the division of labor, the rigid nature of automation routines can result in opportunity costs due to slow adaptation to the changing environment and increased return variability. The reinforcement learning (RL) simulation results show nonlinear myopic automation

behavior: the static policy for automated routines can generate high initial returns to the division of labor but negative long-term returns. Given these results, I argue that organizations can reduce the opportunity costs of automation by developing dynamic routines that balance the learning ability of humans with the productivity benefit of automation.

In the second essay (Chapter Three), I argue that a data-driven organizational culture – a *Data Clan* – can reduce intraorganizational coordination costs that arise when accumulating interdependent assets required for developing AI capabilities. The semi-structured interviews show that managers in incumbent organizations frequently talk about assets related to culture (e.g., trust, curiosity, mindset) and data (e.g., accessibility, reliability, quality). In the theoretical interpretation of results, I argue that a *Data Clan* can be an effective governance mechanism to reduce coordination costs in the face of complex asset interdependencies and unpredictable environments. Therefore, a *Data Clan* might be a valuable asset for firms aiming to build a competitive advantage in the age of AI. While the first two essays focus on the internal organizing processes, the third essay examines how potential investors evaluate a firm's technology-related risk disclosures at IPO.

In the third essay (Chapter Four), I argue that investors require compensation for taking on technology risks disclosed by an IPO firm. I quantify risks disclosed in IPO prospectuses using a latent Dirichlet allocation (LDA) topic model to investigate the magnitude of disclosed risk and the disclosed risk topics. Using this new measure, I extract risk disclosures related to technology to find a significant positive association between the magnitude of technology risk disclosure and underpricing in a *return-for-risk* association. In other words, the results suggest that investors demand higher short-term returns in the form of IPO underpricing for taking on more disclosed technology risk. Moreover, the results show that granted patents can attenuate the positive *return-for-risk* association of technology risk disclosure, suggesting that intellectual property can allow the disclosure of technology information without threatening competitive advantage.

Contributions

Across the three essays, I develop theory and insights to deepen our understanding of how automation can impact the returns from the division of labor, what AI asset stocks incumbent organizations accumulate, and how text-based technology risk disclosure relates to underpricing at the IPO. Together, the results in this dissertation suggest that "technology occupies a central and constitutive role in the organizing process" (Bailey et al., 2022) and, by extension, that technology is central to strategic management research.

The findings in the first essay contribute to the research on the impact of automation on the organization of labor (Dogan and Yildirim, 2021) by linking value capture from the division of labor (Becker and Murphy, 1992) with organizational search and learning in dynamic environments (e.g., March (1991); Zollo and Winter (2002)). The results suggest that firms might benefit from dynamic routines, integrating learning with automation. In the second essay, I contribute to the literature investigating AI in organizations (e.g., Raj and Seamans (2019); Raisch and Krakowski (2020); Von Krogh et al. (2021)) by examining the assets that incumbent organizations accumulate to build AI capabilities. The results and theoretical development suggest that strong data cultures – *Data Clans* – are a governance mechanism to coordinate interdependencies between AI assets. Finally, in the third essay, I contribute to the literature on the role of text-based risk disclosure at IPO (Loughran and McDonald, 2013; Hanley and Hoberg, 2019) and the role of disclosing technological innovations for firms conducting an IPO (Heeley et al., 2007; Morricone et al., 2017). The results suggest that investors require short-term compensation for taking on technology risk disclosures in a *return-for-risk* association and that patents mitigate the *return-for-risk* association as they allow the disclosure of technology information without threatening competitive advantage. Taken together, the above findings contribute to the body of research at the intersection of new

technologies and organizing processes in dynamic environments (Furr, 2021; Bailey et al., 2022).

The dissertation also provides insights for managers. First, by investigating automation initiatives in rapidly changing environments, I point to a potential downside of using aggressive automation strategies: the opportunity costs of slow adaptation. As such, I aim to help managers understand the environmental conditions under which automated routines are more or less likely to pay off. Second, by examining the types of asset stocks for building AI capabilities, I provide evidence of the importance of human capital for real-world AI projects. In doing so, I propose that a *Data Clan* can reduce coordination frictions and, therefore, can be a valuable component in building AI capabilities. Finally, by developing a new approach to measuring text-based risk, I find that text-based technology risk disclosures contain economically and statistically significant signal that investors assess when evaluating an IPO firm. The results further show that patents allow disclosing technology information without threatening competitive advantage as they attenuate the *return-for-risk* association. For managers in firms preparing for an IPO, these insights can help understand the potential consequences of risk disclosures and patenting activities. Overall, I aim to support managers in adopting a realistic, informed, and prudent approach to governing new technologies.

Finally, I attempt to make the following methodological contributions. Whenever appropriate, I use computational methods, including RL and topic modeling, to develop new insights and advance theory (Davis et al., 2007; Hannigan et al., 2019; Choudhury et al., 2020). More specifically, I develop an extension of the *Dyna-Q* RL algorithm (Sutton and Barto, 2018, p. 164) with multiple agents and organizational parameters to simulate the returns to the division of labor with automation under varying environmental conditions. The algorithm enables simulating nonlinear organizational dynamics over time that would be challenging to investigate empirically (Davis et al., 2007). I also contribute to the management literature using topic models (Hannigan et al., 2019)

with the following two applications. First, I apply a hierarchical stochastic block model (hSBM) (Gerlach et al., 2018) for analyzing and visualizing interview transcripts, a valuable technique to triangulate qualitative analyses (Molina-Azorin, 2012). Second, while existing research has investigated the risk magnitude and risk topics separately (e.g., Loughran and McDonald (2013); Bao and Datta (2014)), I develop a new approach to quantify a firm's disclosed risk magnitude and risk topics relative to its year and industry group. Overall, I intend to facilitate future work using computational methods in strategic management through open-sourcing the RL simulation as *OrgSim-RL*[2] and the quantification of risk disclosures as *RiskyData-LDA*[3] on GitHub.

[2] *OrgSim-RL* platform: `https://github.com/mxhofer/OrgSim-RL`, accessed 22 June 2022.
[3] *RiskyData-LDA* platform: `https://github.com/mxhofer/RiskyData-LDA`, accessed 22 June 2022.

Chapter 2

Dynamic Routines: The Opportunity Costs of Automation in Dynamic Environments*

2.1 Introduction

"The greatest improvement in the productive powers of labour . . . seem to have been the effects of the division of labour." — Adam Smith (1965)

What impacts the returns that organizations gain from the division of labor? Smith (1965) argues that the division of labor can raise productivity due to higher returns from time spent on specialized tasks. While Smith points to the market as a limiting factor in the division of labor, Becker and Murphy (1992) identify the costs of coordinating specialized knowledge as a limiting factor. Coordination costs can occur due to, for example, principal-agent conflicts such as free-riding and skirting (Jensen and Meckling, 1976) and transaction costs as a consequence of complex economic organization, opportunistic behavior, and incomplete contracting (Williamson, 2002).

While management practices are often important for efficiently coordinating individuals, new technologies are rapidly disrupting a range of tasks. Frey and Osborne (2017) estimate that 40% of jobs might be automated by 2030. More specifically, automation can substitute for a wide range of manual tasks (e.g., robotics technology in manufacturing) and cognitive tasks (e.g., machine learning in accounting and sales) (Autor et al., 2003). On the one hand, new technologies for automation can increase productivity by reducing costs that arise when coordinating tasks across individuals (Acemoglu and Restrepo, 2019). On the other hand, there are reasons to suspect

*The content of this chapter is based on: Hofer, M. W. and Younge, K. A. (2022). Dynamic Routines: The Opportunity Costs of Automation in Dynamic Environments. Under review at *Organization Science*.

that automation can incur indirect costs (Dogan and Yildirim, 2021). In particular, the challenge of automating tasks that require flexibility and adaptability remains substantial (Autor, 2015), especially for organizations in dynamic environments (Eisenhardt, 1989b). It is often unclear, a priori, what the net effect of automation might be.

The purpose of this study is to examine how automation can shape the returns from the division of labor under varying environmental conditions. Our unit of analysis is an organization in the broad sense (e.g., a team, a business unit, an entire organization) with a given division of labor and a given tendency to automate tasks. We design a two-dimensional grid-world representation of an exploration-exploitation problem with reinforcement learning agents, where automation makes an agent more efficient, but automation also restricts its flexibility to adapt to the environment.

Our simulation model provides the following results. First, organizations can use automation in both stable and highly-dynamic environments to partially escape the limitations to the division of labor set by coordination costs (Becker and Murphy, 1992). Second, automation can increase the returns to the division of labor, but doing so comes at an indirect opportunity cost of slow adaptation and increased return variability. Third, automation myopia can have high returns in the short run but detrimental performance in the long run due to lost learning. Taken together, our simulation generates findings on the dynamics of opportunity costs as a potential downside of automation in dynamic environments.

We aim to contribute to the literature on automation with a novel perspective on the impact of automation in the context of organizational learning (Dogan and Yildirim, 2021). Moreover, we provide a new methodological approach for studying the automation of specialized knowledge in organizations. Open-sourcing *OrgSim-RL*[1] can support future investigations into related topics of organizational significance. For more details, see Appendix B.

[1] *OrgSim-RL* on Github: `https://github.com/mxhofer/OrgSim-RL`, accessed 22 June 2022.

We organize the paper as follows. First, we discuss theory around the division of labor, coordination, dynamic environments, and organizational learning (Section 2.2). Second, we describe our research design and simulation implementation (Section 2.3). Third, we report and analyze the simulation results (Section 2.4). Finally, we discuss organizational implications (Section 2.5) and conclude (Section 2.6).

2.2 Theory

Our theoretical development proceeds as follows. The division of labor can increase the productivity of an organization (Smith, 1965). However, coordination costs can limit the extent of the division of labor (Becker and Murphy, 1992). We build on the basic model in Becker and Murphy (1992) (hereafter, Becker & Murphy) to evaluate the impact of automation in the context of learning and adaptation in dynamic environments.

2.2.1 Division of Labor

The division of labor is a cornerstone of economic progress (Smith, 1965).[2] Smith argues that the division of labor allows individuals to specialize in a narrower set of tasks. With specialized knowledge, the returns to time spent working increase. More specifically, specialized individuals can absorb and process new information efficiently (Bolton and Dewatripont, 1994) and increase organizational knowledge through effectively acquiring new knowledge (Garicano and Wu, 2012). As such, the division of labor can increase the productivity of individuals and, therefore, increase the benefits to the organization.

More formally, the total benefit of the division of labor per team member, B, in Becker & Murphy is a function of both general knowledge, H, and team size, n, as defined in Equation 2.1

[2]Adam Smith published the original edition of *The Wealth of Nations* in 1776.

below. The marginal benefits of adding an additional team member are positive at a decreasing rate, i.e. $\frac{\partial B(H,n)}{\partial n} > 0$ and $\frac{\partial^2 B(H,n)}{\partial n^2} < 0$.

$$B = B(H, n) \tag{2.1}$$

While (Smith, 1965) notes that the market limits the extent of the division of labor, one can find examples of limitations to the division of labor in the organizational literature. Kogut and Zander (1992) argue that organizations exist because they can share and transfer knowledge more efficiently than individual actors in a market could. The authors focus on tacit knowledge, a particular type of knowledge that can be valuable to organizations but difficult to explicitly codify (Polanyi, 1962). Tacit knowledge, however, can limit the division of labor as the noncodifiability can increase costs associated with communicating between the principal and the agent (Garicano and Wu, 2012). Taken together, the division of labor is often important for generating and maintaining knowledge but can be challenging to organize efficiently.

2.2.2 Coordination Costs

The costs of coordinating specialized knowledge can limit the extent to which organizations divide up labor (Becker and Murphy, 1992). Coordinating a group of individuals can be costly due to (at least) two aspects. First, principal-agent conflicts such as free-riding and skirting can reduce the returns to time spent working (Jensen and Meckling, 1976) and can increase costs in the form of salaried compensation to control agent behavior (Eisenhardt, 1985). Second, issues due to incomplete contracting and opportunistic behavior, for example, can increase the internal transaction costs within an organization (Williamson, 2002). The work on agency theory and transaction cost economics suggests that the cost of coordinating a group of specialized individuals increases as the number of individuals increases.

The strategy literature investigates the coordination of specialized knowledge and identifies the importance of management practices that reduce coordination costs. For example, Bolton and Dewatripont (1994) argue that centralization can economize on coordination costs as it avoids unnecessary duplicate communication processes. Garicano and Wu (2012) find that organizational codes and culture can facilitate the coordination of dispersed knowledge because they make it more efficient for individuals to identify problems and match them with an appropriate solution. Reagans et al. (2005) show that when individuals work together in a team, they learn about who knows what, which can facilitate the coordination of activities inside the organization. Investigating production in teams, Deming (2017) shows that individuals with social skills can coordinate with other individuals more efficiently because social skills reduce the costs of trading tasks in a team. Overall, the research suggests that individual members of an organization and how they coordinate matter for firm performance (Bloom and Van Reenen, 2007; Mollick, 2012).

We use the formal model in Becker & Murphy to motivate our investigation of the effect of automation on the returns to the division of labor of an organization, and later we define functional forms for many of the parameters in our simulation model to comply with and extend the Becker & Murphy model. In their model, Becker & Murphy define total coordination costs per team member, C, as a function of team size, n, as in Equation 2.2 below. The marginal costs of adding an additional team member are positive at an increasing rate, i.e. $\frac{\partial C(n)}{\partial n} > 0$ and $\frac{\partial^2 C(n)}{\partial n^2} > 0$.

$$C = C(n) \tag{2.2}$$

Given the benefits to the division of labor in Equation 2.1 and the costs to coordination in Equation 2.2, it follows that the total net reward per team member is the difference between the

benefits and the costs as defined in Equation 2.3 below.

$$\text{Net reward} = B - C \tag{2.3}$$

With both $\frac{\partial B(H,n)}{\partial n} > 0$ and $\frac{\partial C(n)}{\partial n} > 0$, an efficient team generally consists of more than one individual and less than all individuals available in the market, N (Smith, 1965). Becker & Murphy assume that $\frac{\partial B(H,n)}{\partial n} > \frac{\partial C(n)}{\partial n} > 0$ for small teams, meaning that the marginal benefits exceed the marginal costs in small teams.

To provide visual clarity of the net reward of the division of labor, we assume functional forms for the benefits and costs that conform with the properties in Becker & Murphy. Figure 2.1 shows a basic representation of the Becker & Murphy model with the benefits to the division of labor (B) and the costs of coordinating (C). N represents the limitation to the division of labor due to the market (Smith, 1965) and n^* represents the limitation due to coordination costs (Becker and Murphy, 1992). The highest net reward equals the difference between the benefits b^* and coordination costs c^* at a division of labor of n^*.

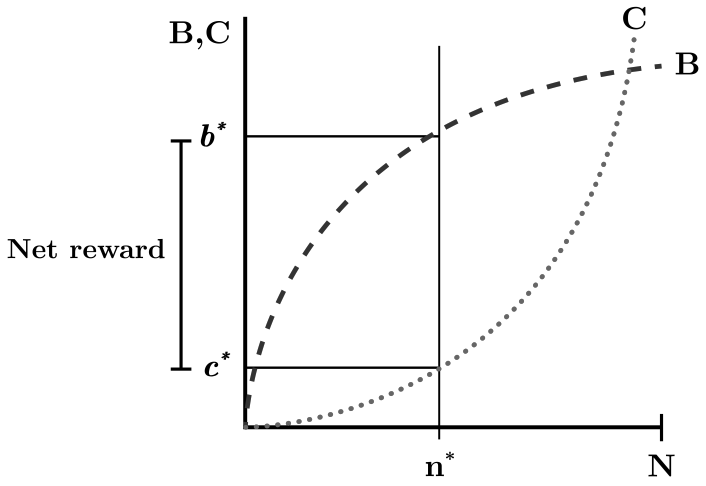

Figure 2.1: The basic model in Becker and Murphy (1992).

2.2.3 Automation

Automation can disrupt the organization of labor as evidenced by Frey and Osborne (2017), who estimate that 40% of jobs might be automated by 2030. More specifically, automation particularly impacts middle-income, routine manufacturing jobs where robots and algorithms can efficiently perform increasingly complex tasks (Charles et al., 2013). As a consequence, Autor and Dorn (2013) observe a shift in the labor market, where individuals reallocate their labor away from middle-income manufacturing jobs to lower-income service jobs. Autor et al. (2003) argue that the manual tasks of service occupations are less susceptible to automation because they require a higher degree of flexibility and adaptability. Given these findings, we conceptualize automation as a computerized, static routine that executes a fixed sequence of actions without the ability to learn and adapt (Autor, 2015). For example, automated teller machines (ATMs) automate the task of dispensing cash (Bessen, 2015). As ATMs become increasingly popular, the number of bank tellers tasked with dispensing cash to customers decreases and the productivity of dispensing cash increases because the division of labor remains unchanged while the coordination costs decrease. As such, we view automation as executing static, codified knowledge (e.g., knowing how to dispense cash securely) without the flexibility to learn and adapt. A more recent example is robotic process automation (RPA), a technology to automate business processes by following a sequence of actions without human intervention. The following two paragraphs expand on this tension that automation introduces.

On the one hand, adopting new technologies for automating tasks can increase the flexibility of allocating tasks to factors of production and reduce costs associated with coordinating tasks across individuals, which Acemoglu and Restrepo (2019) refer to as the *productivity effect* of automation. Related research has shown that knowledge is becoming increasingly specialized, suggesting that

successive generations have to absorb an increasingly large body of knowledge (Jones, 2009) and that coordinating specialized knowledge within teams is becoming increasingly important to producing knowledge (Wuchty et al., 2007). Substituting automation for human labor in specific tasks could mitigate the organizational challenges due to coordinating increasingly specialized knowledge. For example, Cockburn et al. (2019) find that artificial intelligence can increase research efficiency by automating parts of the innovation process of developing specialized drugs. Moreover, Treleaven and Batrinca (2017) describe how automated monitoring of online and social media for financial compliance can increase back-office efficiency, reducing errors at lower costs to the organization. The following proposition summarizes the insight that automation can reduce coordination costs.

Proposition 1: *Replacing human labor with an automated routine reduces coordination costs for a given division of labor.*

On the other hand, automating might have a hidden downside for several reasons. First, automating tasks that require flexibility is problematic due to the limited ability of automation to learn and adapt (Autor, 2015). Mobius and Schoenle (2006) point to the product demand mix as an example of an uncertain and dynamic factor that necessitates a flexible organization of work. While the functionalities of ATMs have generally remained unchanged, one can think of more recent automation of cognitive tasks in marketing, sales, and finance where a lack of adaptability could result in indirect opportunity costs for the organization (Autor et al., 2003). Second, Dogan and Yildirim (2021) point out that substituting automation for human labor in performing a task can incur indirect costs due to a lack of incentive mechanisms for governing principal-agent conflicts under automation. Third, Zollo and Winter (2002) argue that learning mechanisms for knowledge accumulation and articulation are important for developing and refining the operating routines of an organization. Often, automated routines cannot generate new knowledge or discuss such

knowledge with others as humans can. For these reasons, we argue that automation might incur indirect opportunity costs for the organization. The dynamics of such potential opportunity costs under different environmental conditions remain unknown.

Taken together, the two sides of automation create a tension between coordination costs and opportunity costs. To model this tension, we add automation (τ) and the frequency of environmental change (δ) to the total costs, C. Compared to the original Becker & Murphy model, we make two changes. First, we refer to C as the total costs, including opportunity costs and coordination costs. Second, we use λ instead of n for the division of labor to generally distinguish our model from Becker & Murphy. The division of labor (λ), automation (τ), and environmental change (δ) are the central elements of interest for our investigation. Therefore, the new total costs with automation are defined as follows:

$$C = C(\lambda, \tau, \delta)$$

Given the cost structure defined above, an organization can decide whether to automate or not, resulting in an *Automation Mode* and an *Adaptation Mode*. In *Automation Mode*, the organization incurs lower coordination costs but cannot learn; in *Adaptation Mode*, the organization incurs higher coordination costs but can learn and adapt to the environment. As such, we model a trade-off in the benefits and costs of automation that can unfold dynamically over time, i.e. lower coordination costs come at the expense of limited adaptation.

We assume that the general knowledge, H, in the Becker & Murphy model is constant over time for all team members. As such, we depart from Becker & Murphy in that we are not concerned with accumulating a stock of knowledge or human capital over time. Instead, our simulation study will focus on routine optimization and adaptation dynamics. Therefore, we re-conceptualize the total net reward per team member defined in Equation 2.3 in the context of organizational learning

over time. The organizational net reward is our quantity of interest, defined in Equation 2.4 below.

$$\begin{aligned} \text{Net reward} &= B - C \\ &= B(\lambda) - C(\lambda, \tau, \delta) \end{aligned} \tag{2.4}$$

Comparing the Becker & Murphy model in Equation 2.3 to our model in Equation 2.4, one can see how automating tasks might alter total costs, C, and, as a consequence, the net reward gained from the division of labor.

As shown in Figure 2.2, the overall efficiency gain or loss of automation depends on how automation changes coordination costs and opportunity costs, holding the division of labor constant. Automation could increase the net reward ($b^* - c^-$) or decrease the net reward ($b^* - c^+$).

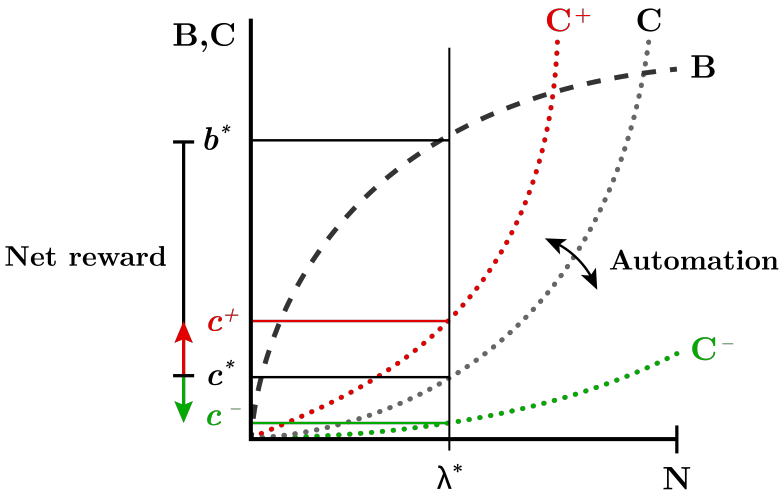

Figure 2.2: The conceptual impact of automation as an extension of Becker and Murphy (1992).

Given the structure of benefits, costs, and the net reward developed above, we specify how automation affects coordination costs and opportunity costs under different environmental conditions in the next section.

2.2.4 Dynamic Environment

Today, organizations often operate in rapidly changing environments (e.g., Brown and Eisenhardt (1997)). Adopting new technologies in such environments necessitates a balance between exploring new opportunities and exploiting existing competencies (March, 1991). March suggests that finding and maintaining the appropriate balance between exploration and exploitation can be challenging as the effectiveness of generating new knowledge in turbulent environments deteriorates with time. To cope with changing environments, Teece et al. (1997) proposes that organizations continuously reconfigure internal and external competencies. Brown and Eisenhardt (1997) argue that successful organizations adapt their managerial procedures and organizational structures in high-velocity markets. The research generally suggests that environmental change can affect how organizations operate.

The division of labor can enable the development of specialized knowledge assets (Garicano and Wu, 2012). Developing, maintaining, and efficiently coordinating such a knowledge base is central to the growth and survival of organizations (Kogut and Zander, 1992). However, environmental change can render existing capabilities no longer effective (Tushman and Romanelli, 1985). More specifically, environmental change can erode the future value of existing knowledge due to lost fitness with the environment. Change can also mitigate the value of the effort to generate new knowledge because change can erode the half-life of returns to new knowledge (Posen and Levinthal, 2012). Changing environments necessitate organizational flexibility and adaptation to ensure fitness with the environment. Given the limited ability to learn and adapt under automation, automation might incur indirect opportunity costs in a dynamic environment, as the following proposition summarizes.

Proposition 2: *Replacing human labor with automation reduces organizational capabilities to learn and adapt to environmental change for a given division of labor, resulting in an indirect opportunity cost.*

2.2.5 Learning and Adaptation

Modern business operations contain a large number of complex and interdependent elements (Sculley et al., 2015) and human problem solvers often cannot calculate optimal solutions for such complex systems because they lack complete knowledge of all the relevant aspects (Simon, 1955). Moreover, even if a problem solver would have near-complete knowledge, the combinatorial complexity of the solution space grows exponentially and quickly becomes intractable. Consequently, problem solvers often engage in adaptive learning about the business environment to find and maintain a satisfactory solution (Simon, 1955; March and Simon, 1958).

In organizational learning, agents balance between allocating resources to explore the novel or exploit the known (March, 1991). For example, the agent might choose to invent a new technology or refine an existing one, a trade-off that becomes especially challenging under environmental instability and ambiguity (Levinthal and March, 1981). Zollo and Winter (2002) investigate how different learning mechanisms for accumulating, articulating, and codifying knowledge can create and maintain dynamic capabilities. Zollo and Winter note that the relative importance of learning for accumulating knowledge depends on the rate of environmental change. Posen and Levinthal (2012) suggest that both stable environments and environments with very high rates of change can diminish the importance of learning new knowledge. To investigate the dynamic aspects of learning and adaptation, we embed our modified version of the Becker and Murphy (1992) model into an environment that we can simulate with reinforcement learning. We explain our hybrid approach in the next section on research design.

2.3 Research Design

We design a simulation to explore the effects that arise when organizations choose automation to gain rewards from the division of labor. More specifically, we develop a computational representation of a multi-domain environment with a trade-off between exploration and exploitation, and simulate learning and adaptation in that environment with reinforcement learning agents.

2.3.1 Simulation

Simulation approaches can be useful for investigating the outcomes of the interactions across multiple underlying economic and organizational elements, especially as these evolve over time (Repenning, 2002), and to develop theory around a fundamental tension (Davis et al., 2007). A simulation is appropriate for our study because we investigate the potentially non-linear interactions across organizational elements in the context of a dynamic learning problem.

Several studies in the organizational literature use a simulation approach to investigate problems of dynamic learning and adaptation. One can group the types of simulations into (1) generative algorithms, (2) NK models, and (3) reinforcement learning approaches. A generative algorithm models assumptions to understand which dimension(s) generated the data. For example, Levinthal and March (1981) and March (1991) use a generative model to investigate the exploration-exploitation trade-off under varying learning and environmental conditions. The NK model builds on the work by Kauffman et al. (1993) and allows researchers to model learning on a tunably rugged fitness landscape.[3] For example, Levinthal (1997) uses an NK model to investigate how different organizational forms can adapt to change more or less effectively. More recently, studies have turned

[3] N represents the attributes of an organization, where each attribute can take a binary value. Each of these attributes can interact with K other attributes. The overall fitness of the organization depends on the combination of attribute values and their interactions. For a more detailed introduction to the NK model in the organizational literature, see Levinthal (1997).

to reinforcement learning approaches. The multi-armed bandit problem is one representation of the exploration-exploitation problem for studying path dependence (Denrell and March, 2001) and turbulent environments (Posen and Levinthal, 2012). The bandit model represents a sequential choice model, in which the organization has to select among alternative arms with associated reward distributions. The agent reinforces its beliefs about the quality of each arm based on the realized reward. Finally, reinforcement learning approaches with a grid world environment have been used to model behavioral problems in an organizational context, such as procurement auctions (Greenwald et al., 2010), the willingness to persist at innovating in the face of failure (Rosokha and Younge, 2020), and devising optimal tax policies (Zheng et al., 2021).

2.3.2 Environment

We develop a grid world environment to represent the essential aspects of our propositions (see Figure 2.3). We design that environment to have a structure and a set of rules such that: (1) organizations with a stronger division of labor have superior knowledge, (2) learning agents trade-off exploring a risky but potentially rewarding path and exploiting a risk-free path, and (3) researchers can independently manipulate the parameters of interest. The environment is just one of many possible grid world configurations that might be appropriate for our investigation, and we open-source all source code to facilitate the study of alternative environments. As seen in Figure 2.3, there are four separate domains within the environment, wherein a single domain is analogous to the *blocking maze* and *shortcut maze* examples in Sutton and Barto (2018, p.167).

Structure

The simulation operates on a 13×13 grid that consists of four domains (denoted as 1, 2, 3, and 4) and three types of cells: walls (in light gray), doors (in dark gray), and empty cells (in blue, green,

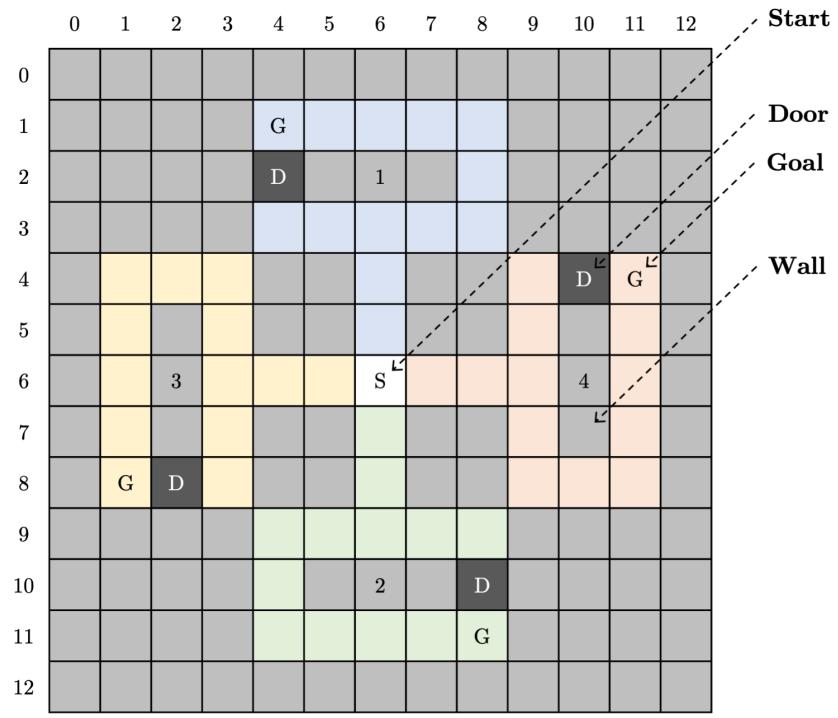

Figure 2.3: The two-dimensional grid world.

yellow, and red). The domain-specific colors have no purpose other than distinguishing domains for the reader. Each domain contains a goal state, G, that is associated with a goal reward and a door state, D. The agent reaches the goal state within seven steps if the door is open (short path) or 11 steps (long path). The short path is risky as the door might close unexpectedly, in which case the agent makes six steps and ends up at the start without obtaining the goal reward. The long path is risk-free but takes more steps to reach the goal.

Rules

Agents can step through empty cells but cannot go through walls or closed doors. Agents start a new episode at the start state, S. Agents are sent back to start when (1) reaching the goals state (G) or (2) attempting a closed door (D). Agents are restricted to their domain and cannot backtrack to the start state or within the domain. States only permit valid actions from the set of all states, $\{\text{up}, \text{down}, \text{left}, \text{right}\}$. Valid actions are actions where the agent does not hit a wall

or backtrack. We chose this channeled design with a restricted set of actions to remove learning complications not directly relevant to our investigation.

The environment might be useful beyond the current study for the following reasons. First, deciding which domain to enter at the start state, S, is essentially a multi-armed bandit problem. The environment extends the bandit model to add flexibility for operationalizing elements of interest. As such, the environment builds on existing research using a bandit model (e.g., Denrell and March (2001); Posen and Levinthal (2012)) while enabling researchers to investigate questions requiring a malleable simulation environment. Second, one can increase the number of domains beyond the current four domains by increasing the number of actions available at the start state, S, though at the expense of the ability to visualize more than four domains.

2.3.3 Learning and Adaptation

We select Q-learning (Sutton, 1990) to model how an agent learns while moving through the environment. Q-learning has previously been used to investigate related topics such as credit assignment (Denrell et al., 2004) and exploration versus exploitation in multi-stage problems (Fang and Levinthal, 2009). An alternative simulation approach to investigate learning problems in the management sciences is the NK model (Kauffman et al., 1993), a hill-climbing technique. However, The NK model has two main limitations for our investigation. First, our research design requires a simulation approach where we can inspect an agent's internal understanding of the environment to validate the speed and degree to which agents adapt to environmental changes under different parameter combinations. While the NK model does not provide access to an agent's understanding about the environment and focuses on the outcome of the learning process, Q-learning allows us to inspect the actual adaptive process. Second, the NK model only defines the statistical properties of the landscape, not its actual shape (Valente, 2008). To operationalize the division of

labor, automation, and environmental change, we require greater plasticity of the landscape. One might alternatively investigate these topics through human subject experiments or formal economic modeling, but we chose a computational approach because it requires precisely defined constructs, logic, and assumptions that can improve internal validity (Davis et al., 2007).

We extend the Dyna-Q reinforcement learning algorithm in Sutton and Barto (2018, p. 164) into a multi-agent system with division of labor, automation, and environmental change.[4] In Dyna-Q, an agent i learns by updating its beliefs based on a realized reward, R, of taking a particular action, A, to move from the current state, S, to the next state, S'. Those beliefs are stored in a Q-table, Q_i. At each step t, the agent updates its Q-table according to the Bellman equation in Equation 2.5 below, where α is the learning rate that controls the magnitude of the learning update and γ is the discount factor that discounts future expected rewards.

$$Q_i(S,A) \longleftarrow Q_i(S,A) + \alpha\big[R + \gamma \max_a Q_i(S',a) - Q_i(S,A)\big] \tag{2.5}$$

2.3.4 Organizational Episodes

We conceptualize the organization in a broad sense as a team, a business unit, or an entire firm. The organization has a given division of labor and a given tendency to automate, which it uses to execute organizational episodes sequentially. Our terminology matches that in Sutton and Barto (2018) such that a single simulation *run* consists of many *episodes*, which consist of individual *steps* through the grid world.

Figure 2.4 depicts an organizational episode. The *Organization* starts an episode at the start state, S in the grid world in Figure 2.3, and decides which of the four domains to enter – the

[4]The multi-armed bandit model is an alternative reinforcement learning model that has also been used in organizational studies (e.g., Denrell and March (2001); Posen and Levinthal (2012)). We use the Dyna-Q algorithm because it gives us more flexibility in operationalizing the elements of interest to our study than a bandit model could.

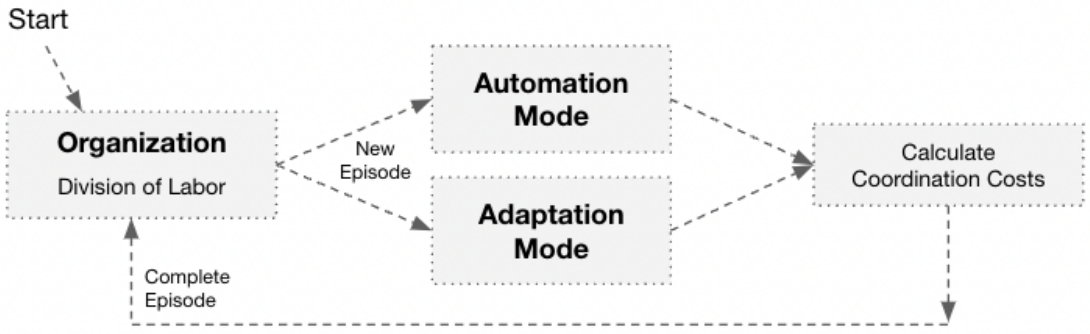

Figure 2.4: Organizational episode.

organizational decision at the start is essentially a four-armed bandit problem. The organization can leverage specialized expertise in each domain by following the agent for that domain. Next, the organization decides whether to automate the episode (*Automation Mode*) or not automate and adapt (*Adaptation Mode*). At the end of the episode, the organization observes the goal reward, the number of steps to reach the goal, and coordination costs to compute the net reward per episode, our main outcome of interest. Equation 2.6 defines the organizational net reward per simulation episode, where $\phi \in [0, 1)$ is the percentage of coordination costs incurred in the *Automation Mode* – a type of "carrying cost" for running the automation system.

$$\begin{aligned} \text{Net reward} &= \text{Benefits} \quad - \text{Costs} \\ &= \text{Goal reward} - \text{Step cost} - \begin{cases} \phi \cdot \text{ Coordination costs, if } \textit{Automation Mode} \\ 1 \cdot \text{ Coordination costs, if } \textit{Adaptation Mode} \end{cases} \end{aligned} \tag{2.6}$$

The net reward defined in Equation 2.6 models the dynamic tension that *Automation* introduces: in the short term, *Automation* might economize on coordination costs given that $\phi < 1$. In the longer term, however, *Automation* might incur higher step costs as it fails to find paths where the door is open as the environment keeps changing, resulting in a low net reward. As such, Equation 2.6 implements the extension of the basic model in Becker and Murphy (1992) in Equation 2.4,

where automation affects the total costs of gaining benefits from the division of labor. The following sections describe the conceptualization and implementation of the elements of interest, where we select functional forms and scalar values to correspond with the properties specified in the basic model in Becker and Murphy (1992). More specifically, scaling the elements of interest is a form of hyperparameter tuning to implement the baseline tension between the benefits and costs of the division of labor without automation in stable environments.

Division of Labor (λ)

The number of agents, λ, defines the extent of the division of labor (DOL) of an organization (Becker and Murphy, 1992). In the case of a single agent, the agent has to cover all four domains in the grid world. Two agents split up the domains to cover two domains each, and in the case of four agents, an agent covers a single domain. As such, our design contrasts no DOL (a single agent, $\lambda = 1$) with intermediate DOL (two agents, $\lambda = 2$) and a high DOL (four agents, $\lambda = 4$). Each agent has its own independent Q-table, which we initialize to zero (Sutton and Barto, 2018, p.164). The DOL is exogenous to the learning and adaptation process. To correspond with Becker and Murphy (1992), we have to operationalize the (1) benefits and costs to the DOL and the (2) productivity of individuals.

Benefits. Modeling the benefits as a function of the DOL (λ) comes naturally in Q-learning. A higher DOL increases the benefits (Becker and Murphy, 1992, p.1142). We operationalize the benefits as a function of how efficiently agents learn about the environment and reach the goal reward, $B(\lambda)$. The DOL impacts the output per individual worker (Becker and Murphy, 1992), represented by the goal reward amount in our simulation. We select the natural logarithm as one of many functions that model the idea that the benefits to the DOL exhibit positive marginal returns at a decreasing rate, and that corresponds with the properties specified in Becker and Murphy

(1992). Equation 2.7 defines the goal reward per episode, where π scales the goal reward to an appropriate amount and adding 1 to λ ensures a positive reward for the goal state with no division of labor (i.e., $\lambda = 1$) as $ln(1) = 0$. The realized goal rewards equal $\pi\ ln(2) \approx \pi\ 0.69$ with no division of labor, $\pi\ ln(3) \approx \pi\ 1.1$ with an intermediate division of labor, and $\pi\ ln(5) \approx \pi\ 1.61$ with a high division of labor.

$$\text{Goal reward} = \pi\ ln(\lambda + 1) \tag{2.7}$$

Costs. The *DOL* also impacts the knowledge of agents (Becker and Murphy, 1992), which they use to navigate the grid world. We assign a cost of 1 to each step that the agent makes through the grid world.

Productivity. We conceptualize the productivity of individuals as their learning ability. In other words, a higher *DOL* means that individuals are more productive in adapting their knowledge of the environment (Becker and Murphy, 1992). We operationalize the advantage of more specialized learning in terms of the learning rate, α, with decreasing returns to model the cognitive limitations of individuals and correspond with Becker and Murphy (1992). Higher learning rates are beneficial as the environment is largely deterministic and there are very little disadvantages to rapid learning given our operationalization of *Environmental Change* as described later in this section. We set the actual learning rate, α' as defined in Equation 2.8, where ν is a scalar value and the term $ln(\lambda)$ increments the base learning rate as the division of labor increases. With no division of labor, $\alpha' = \alpha$ as $ln(1) = 0$.

$$\alpha' = \alpha + \nu\ ln(\lambda) \tag{2.8}$$

Coordination Costs

The *DOL* affects the costs of coordinating specialized individuals (Becker and Murphy, 1992). We operationalize the costs associated with coordination as a polynomial function with increasing

marginal returns to model the exponential growth in coordinating activities between individuals as the number of individuals increases and to correspond with the properties specified in Becker and Murphy (1992). We assume that knowledge is not *perfectly* codifiable as perfectly codifiable knowledge would require no coordination across individuals. The coordination costs are exogenous to the learning and adaptation process. Equation 2.9 defines coordination costs per episode, where η scales the costs to an appropriate amount and an exponent on λ equal to $\frac{5}{3}$ ensures that marginal costs are positive at an increasing rate. We conduct robustness checks by varying the value of the exponent in Section 2.4.5. Moreover, subtracting 1 from $\lambda^{\frac{5}{3}}$ ensures that the organization incurs no coordination costs with no division of labor (i.e., $\lambda = 1$).

$$\text{Coordination costs} = \eta\ (\lambda^{\frac{5}{3}} - 1) \tag{2.9}$$

Automation (τ)

We operationalize *Automation* as an exogenous probability, τ, that the organization chooses a static sequence of actions in a given episode. As such, *Automation* alters the reinforcement learning policy. In the *Automation Mode*, the organization selects the *argmax* action from the start and uses the Q-table of the agent in that domain to construct the fixed sequence of *argmax* actions to reach the goal. Following the conceptualization of *Automation* as a static sequence of actions with limited adaptability (as in Autor, 2015), an automated routine does not update the Q-table ($\alpha = 0$) and does not explore ($\epsilon = 0$). It is important that the Q-table is not updated in the *Automation Mode* to represent the limited adaptability and static nature of automated routines in organizations. We implement the difficulty of automating tasks that require creativity (as discussed in Autor, 2015) as a greedy policy that cannot take random actions. In the *Automation Mode*, the organization incurs reduced coordination costs as defined in Equation 2.6 but has a higher likelihood of being

stuck in a suboptimal outcome due to the inability to learn and a greedy action policy.

Environmental Change (δ)

It can be challenging to forecast when new opportunities open up or become unavailable, especially in the context of workplace automation (Brynjolfsson and Mitchell, 2017). Consistent with studies investigating learning under environmental change (Robert Baum and Wally, 2003; Posen and Levinthal, 2012), we conceptualize *Environmental Change* as the frequency of change in the environment.

We operationalize *Environmental Change* as the exogenous frequency of doors opening and closing per simulation run, δ. We designed doors to change at punctuated points (i.e., at equally spaced intervals across all episodes, synchronized across simulation runs) to uncover the subtleties of the dynamic learning behavior. For example, for $\delta = 3$ and a simulation run with 1,000 episodes, door states change at episodes 250, 500, and 750 in all simulation runs.[5] Door states are initialized at random but require that at every point in time, two doors are open and the other two are closed. The door initialization strategy ensures that the expected reward remains constant throughout a simulation run. Given our operationalization of *Environmental Change*, higher learning rates, as described earlier, are beneficial because the agent benefits from rapidly learning the new mapping of a state-action pair to the next state.

2.3.5 Execution Flow

We now combine the environment, the learning mechanism, and the organizational elements of interest into a simulation procedure that we summarize visually in Figure 2.5 and algorithmically

[5]The organizational literature on dynamic environments often models environmental change as punctuated changes that interrupt periods of stability (e.g. Tushman and Anderson (1986); Romanelli and Tushman (1994)). We also test an alternative specification of *Environmental Change* where doors change smoothly at random episodes, which corresponds to continuous environmental change as in Brown and Eisenhardt (1997).

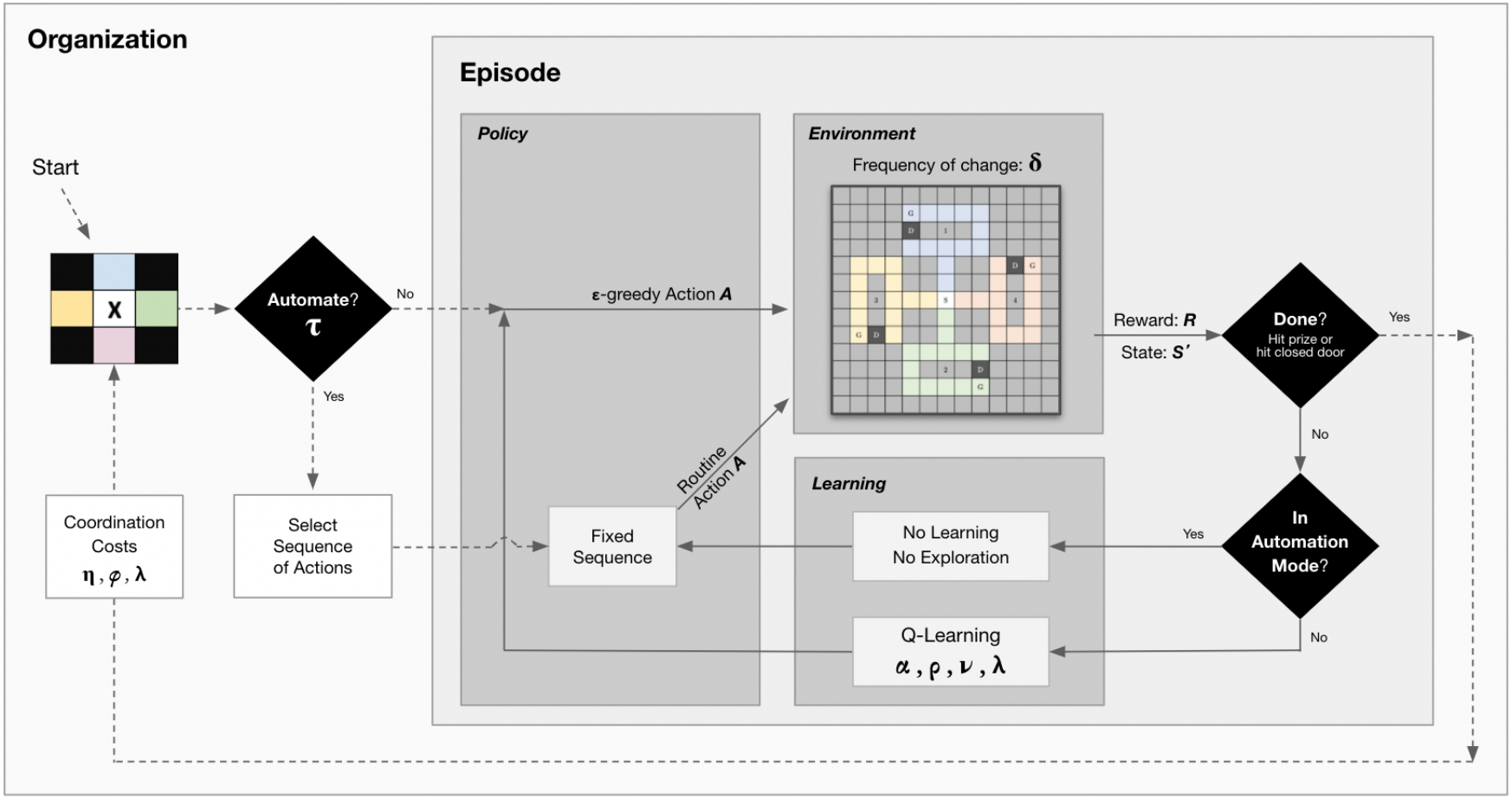

Figure 2.5: Execution flow of our simulation.

in Algorithm 1. Table 2.1 summarizes the pseudocode notation and Table 2.2 lists parameters and parameter values along with a description and operationalization. All parameter values have been tuned to match the functional forms and characteristics described earlier while scaling to appropriate magnitudes that match our research design.

Figure 2.5 shows the execution flow of our simulation. Dashed arrows represent organization-level connections. Solid arrows represent episodic connections. The starting point is denoted as *Start* in the *Organization* plate. At the start of an episode, the organization makes an ϵ-greedy decision that decides which domain it will enter. The organizational decision at the start is analogous to a four-armed bandit model, in which an agent selects the arm with the highest expected reward, $1-\epsilon$ percent of the time. We implement indirect learning from past experience as described in Sutton and Barto (2018, p.161) to backpropagate changes in the environment more quickly, such that the expected rewards per domain at the start state better reflect the value of entering a domain. For indirect learning, we let *each* agent sample ρ state-action transitions at each step of an episode in *Adaptation Mode*. The next step in the diagram shows the organizational decision to automate the episode or not, based on the value of τ. When automating, the agent's policy is to execute a fixed *argmax* sequence of actions – the *Automation Mode*.[6] When not automating, the agent takes ϵ-greedy actions – the *Adaptation Mode*. Next, the agent evaluates the selected action, A, against the environment, where doors change based on the value of δ. If the agent is not yet done with the episode, it (1) skips learning in the *Automation Mode* or (2) updates its Q-values in the *Adaptation Mode* (governed by parameters α, ρ, ν, and λ) before executing the next step. Finally, the episode ends when the agent reaches the goal state (G) or attempts a closed door (D) – the agent is sent back to the start (S) in both cases. The organization now computes coordination costs (governed

[6]We assume that organizations can discard invalid automation routines, e.g., routines that never reach the goal state. Therefore, the organization does not automate the first few episodes as it needs to make a first pass through the environment. In a similar context, MacCormack et al. (2013) suggests that organizations can distinguish valid from invalid innovations.

by parameters η, ϕ, and λ) and observes the goal reward gained and the number of steps it took the agent to reach the goal.

$\mathcal{S}$	Set of all non-terminal states	$\mathcal{A}(s)$	Set of all actions available in state s
t	Discrete time step	e	Episode
A_t	Action at step t	A_{start}	Action at start for domain selection
S_t	State at step t	R_t	Reward at step t
S_{start}	Start state	D_e	Domain in episode e (1, 2, 3, or 4)
S_{goal}	A goal state	M_e	Mode in episode e
S_{door}	A closed door state	Q_i	Q-values of agent in the current domain
Q_{start}	Q-values start	P_{routine}	Sequence of actions for automation

Table 2.1: Summary of pseudocode notation.

One might argue for empirical validation to ensure that the simulation results are generalizable as opposed to particularities of the simulation model. Davis et al. (2007) suggest two avenues for empirical validation. First, one could use large-scale empirical data to validate the predictions generated by the simulation model. Second, one might conduct case study interviews to generate qualitative data to assess the generalizability of the simulation results. Given that organization-level data on automation initiatives are scarce as reported by a recent National Academies of Science Report (NAS 2017) and case studies on automation initiatives lie outside the scope of the current study, we mitigate concerns about the generalizability of our results with sensitivity checks for alternative specifications as reported in Section 2.4.5.

To summarize, our simulation consists of a grid world environment and an organization with λ agents that step through the environment. The *DOL* (λ) impacts the goal rewards and the learning ability of the organization (Becker and Murphy, 1992). *Automation* (τ) alters the behavioral policy to a fixed *argmax* sequence of actions based on the Q-values of the agent for that domain. *Environmental Change* (δ) defines how frequently doors open or close. In the next section, we will examine how organizational net rewards change for different paramater configurations of the *DOL*, *Automation*, and *Environmental Change*.

Algorithm 1: Dyna-Q with Division of Labor, Automation, and Environmental Change

Initialize an organizational-level $Model(s, a)$ and λ independent $Q(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$.
Initialize Q-values to zero. Initialize half the environment doors as open, half as closed.
Set door switching episodes based on δ. Set automated episodes based on τ. Initialize $S_0 \longleftarrow S_{\text{start}}$
while $e <$ *episodes*:
    if $\lambda == 1$: `// Division of labor (λ)`
        $Q_{\text{start}} \longleftarrow Q_1$
    elif $\lambda == 2$:
        $Q_{\text{start}} \longleftarrow \{\ Q_1(S_{\text{start}}, \{\text{up, down}\}),\ Q_2(S_{\text{start}}, \{\text{left, right}\})\ \}$
    elif $\lambda == 4$:
        $Q_{\text{start}} \longleftarrow \{\ Q_1(S_{\text{start}}, \text{up}),\ Q_2(S_{\text{start}}, \text{down}),\ Q_3(S_{\text{start}}, \text{left}),\ Q_4(S_{\text{start}}, \text{right})\ \}$
    $A_{\text{start}} \longleftarrow$ select ϵ-greedy action across all 4 action values in Q_{start}
    $M_e \longleftarrow$ *Adaptation Mode*
    $D_e \longleftarrow$ select domain given A_{start}(domain for episode)
    if $e \in$ *list of automated episodes*: `// Automation (τ)`
        $P_{\text{routine}} \longleftarrow$ Construct routine from Q_{De} as *argmax* sequence of actions from start
        $M_e \longleftarrow$ *Automation Mode*
    else:
        $P_{\text{routine}} \longleftarrow$ None
    while $S_t \neq S_{goal}$ ***or*** $S_t \neq S_{door}$*; loop through steps for* $t = 0, 1, 2, ...$:
        $S_t \longleftarrow$ current state
        if $M_e ==$ *Automation Mode* :
            $A_t \longleftarrow$ next action from P_{routine}
        else:
            if $t == 0$:
                $A_t \longleftarrow A_{\text{start}}$
            else:
                $A_t \longleftarrow \epsilon$-greedy$(S_t,\ Q_i)$
        Take action A_t ; observe reward R_t and state S_{t+1}
        if $M_e ==$ *Automation Mode* : `// Automation vs. Adaptation Mode`
            No learning, no experience gained
        else:
            Continue with standard tabular Dyna-Q learning (Sutton and Barto, 2018, p. 164)
        Net reward $= R_t - 1 \cdot t - \begin{cases} \phi \cdot \eta\ (\lambda^{\frac{5}{3}} - 1) & \text{, if } M_e == \textit{Automation Mode} \\ 1 \cdot \eta\ (\lambda^{\frac{5}{3}} - 1) & \text{, if } M_e == \textit{Adaptation Mode} \end{cases}$ (see Equation 2.6)
        $S_{\text{t+1}} \longleftarrow S_{\text{start}}$
        if $e \in$ *list of door switching episodes*: `// Environmental change (δ)`
            Open all doors; randomly select two new doors to close
        e += 1

Fixed Setup Parameters

Name	Value	Description
Cost per step	1	Cost to make a step
Steps for risk-free path	11	Steps to execute the risk-free path
Steps for risky path	7	Steps to execute the risky path
Episodes	1,000	Number of episodes for a simulation run
Simulation runs	10,000	Number of simulation runs

Q-Learning Algorithmic Parameters

Name	Parameter	Value	Description
Learning Rate	α	0.3	Learning rate for Q-Learning when $\lambda = 1$
Indirect Learning	ρ	4	Number of indirect planning steps per agent per episode
Discount Factor	γ	0.9	Discount factor for Q-learning continuation value
Exploration Rate	ϵ	0.1	Probability of taking a random action on each step

Economic Parameters

Name	Parameter	Value	Description	Implementation
Learning Advantage	ν	0.35	Scaling of learning rate (α) as a function of the division of labor (λ)	Learning rate $= \alpha + \nu\ ln(\lambda)$
Benefits	π	10	Scaling of reward at goal state as a function of the division of labor (λ)	Goal reward $= \pi\ ln(\lambda + 1)$
Coordination Costs	η	0.9	Scaling of coordination costs as a function of the division of labor (λ)	Coordination costs $= \eta\ (\lambda^{\frac{5}{3}} - 1)$
Carrying Costs	ϕ	0.5	Percentage of coordination costs incurred under automation (τ)	Carrying costs $= \phi\ \eta\ (\lambda^{\frac{5}{3}} - 1)$
Transition Costs	ψ	0	Costs to transition in/out of automation as a function of lambda (λ)	Transition costs $= \psi\ \lambda$

Main Parameters of Interest

Name	Parameter	Values	Description
Division of labor	λ	$\{1, 2, 4\}$	Number of agents in a simulation
Automation	τ	$\{0, 0.5, 0.95\}$	Share of episodes that are automated in a simulation
Environmental Change	δ	$\{0, 10, 1000\}$	Number of times environment changes in a simulation

Table 2.2: Simulation parameters.

2.3.6 Parameter Configuration

We configure the simulation to execute 10,000 runs of 1,000 episodes for each of the different combinations of *DOL*, *Automation*, and *Environmental Change* conditions that are summarized in Table 2.3.

	None	Intermediate	High
Division of labor (λ)	1	2	4
Automation (τ)	0	0.5	0.95
Environmental change (δ)	0	10	1000

Table 2.3: Main parameter values.

2.4 Results

We organize our results as follows. First, we validate that our simulation captures the trade-off between the benefits and costs of the division of labor that constitute the core of the Becker and Murphy (1992) model. Second, we investigate how *Automation* and environmental change jointly shape net rewards. Finally, we examine density estimates of the net reward and trace the trajectory of learning over time.

2.4.1 Baseline

We first validate the benefits and coordination costs with no *Environmental Change* and no *Automation* (Becker and Murphy, 1992). Figure 2.6, Panel (A), shows functional forms for the benefits and coordination costs that is consistent with the properties in Becker and Murphy (1992). Panel (B) plots the benefits[7] and coordination costs as they arise in our simulation. The result validates that coordination costs and the division of labor are positively related and that an intermediate

[7]For a direct visual comparison of Becker and Murphy (1992) and our simulation output, the benefits in Panel (B) have to equal the goal reward net of the step costs because the step costs are a feature of our behavioral learning model that is not present in the formal model in Becker and Murphy (1992).

DOL results in the highest net reward.[8] Panel (C) plots the difference between the benefits and costs, the accumulated net reward to the organization[9], our main outcome of interest.

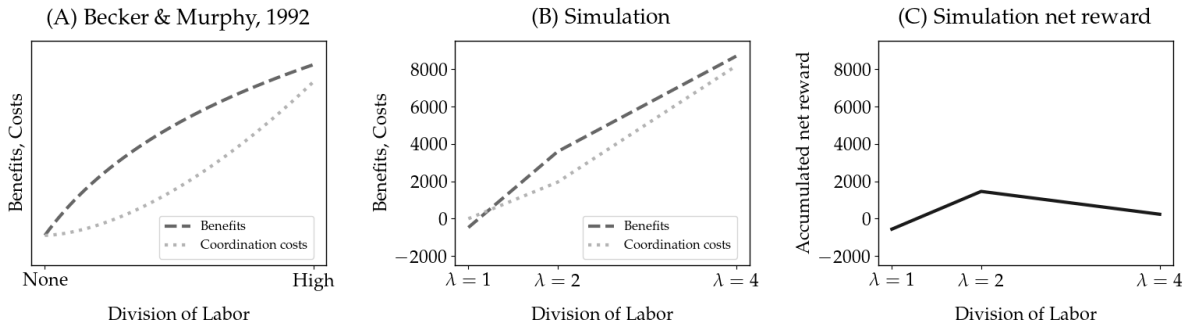

Figure 2.6: Baseline validation of *DOL*.

2.4.2 Automation

We now introduce *Automation*. Figure 2.7, Panel (A), shows the accumulated net reward in a stable environment across levels of *DOL* for different levels of *Automation*. The dark grey line (no *Automation*) corresponds to the baseline result in Figure 2.6, Panel (C). When adding *Automation* in stable environments, the net reward generally increases. *Automation* is particularly beneficial for highly specialized knowledge due to the considerable reduction in coordination costs. In short, automation routines are generally beneficial in stable environments.

Upon further examination of Figure 2.7, Panel (A), however, one can see that with no *DOL* in stable environments, intermediate *Automation* outperforms high *Automation*. We observe this result because *Automation* is greedy: it automates the first path that reaches the goal, which might not be the short path with the highest net reward. Intermediate *Automation* is more likely, on average, to find and automate the short path to the goal, resulting in higher accumulated net rewards.

[8] We validate that the functional form requirements for benefits in Becker and Murphy (1992) hold when introducing *Automation* and *Environmental Change*. The functional form for coordination costs does not depend on *Automation* or *Environmental Change*.

[9] The accumulated net reward equals the sum of episodic net rewards across all episodes, averaged across all independent simulation runs.

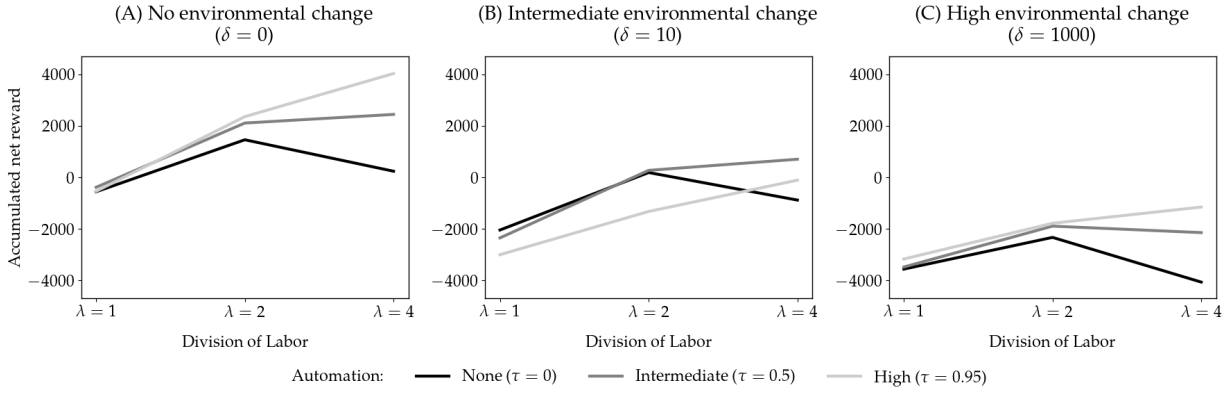

Figure 2.7: *Automation* by levels of *DOL* and *Environmental Change*.

2.4.3 Environmental Change

We add *Environmental Change* in Panel (B) and Panel (C) of Figure 2.7. In general, dynamic environments reduce net rewards as agents invest in learning and adapting to unexpected changes in the environment. In Panel (B), one can see that intermediate *Automation* strictly outperforms high *Automation*. As doors change in the environment, intermediate *Automation* is more flexible while high *Automation* gets locked into a fixed sequence of actions that can become unfit with the environment. Despite higher coordination costs, no *Automation* outperforms high *Automation* for low and intermediate *DOL*. In general, these results provide evidence that automation can incur indirect opportunity costs in dynamic environments (*Proposition 2*).

In Figure 2.7, Panel (C), *Environmental Change* is random: doors change at each episode, eliminating any benefits to learning. High *Automation* strictly outperforms less aggressive automation strategies. Agents do not benefit from learning these rapid changes, making the automation of existing knowledge a more effective strategy. These results are in line with Posen and Levinthal (2012), who find that under high environmental turbulence, the appropriate response can be a "focus on exploiting existing knowledge and opportunities."

The results for *Automation* and *Environmental Change* suggest that *Automation* can be beneficial to the organization in environments where learning and adaptation are ineffective, shown in Figure 2.7, Panel (A) and Panel (C). However, in environments where learning and adaptation are beneficial, plotted in Figure 2.7, Panel (B), automation incurs an indirect cost of lost learning and slow adaptation that results in low net rewards when aggressively automating.

2.4.4 Opportunity Costs

To further investigate *why* high *Automation* underperforms for particular levels of *Environmental Change*, we turn to the dynamic trade-off between efficiency and learning in Equation 2.6. We measure the emergent opportunity costs of lost learning as the difference between the actual net reward per episode and the best possible net reward per episode that an oracle with full information could attain. Figure 2.8 shows the coordination costs (dark grey) and opportunity costs (light grey) by *Automation* and *Environmental Change* with intermediate *DOL* ($\lambda = 2$). Coordination costs generally decrease with automation, consistent with *Proposition 1.* Opportunity costs generally rise as the environment changes more frequently. When the environment occasionally changes in Panel (B), and the organization aggressively automates, opportunity costs rise because the automation routine is unfit with the environment, consistent with *Proposition 2.* Opportunity costs have two sources: (1) the organization automates the long path and cannot find the short path with an open door because automation cannot take exploratory actions and (2) the organization automates the short path with an open door but cannot adapt away from that path when the door closes. In essence, automating when learning and adaptation are beneficial can be dangerous due to the opportunity costs of lost learning and slow adaptation.

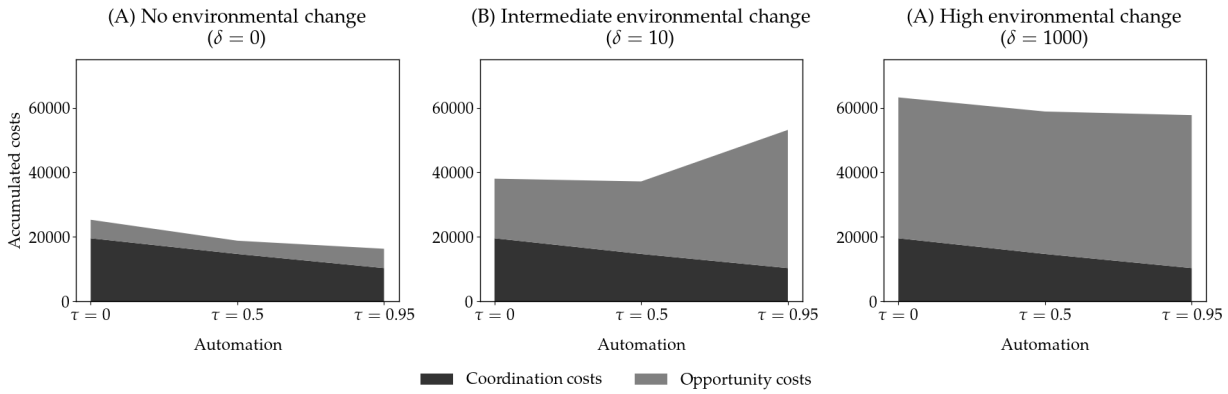

Figure 2.8: Coordination costs and opportunity costs at an intermediate *DOL*.

2.4.5 Sensitivity

We examine the sensitivity of our results to alternative specifications of (1) *Automation*, (2) *Environmental Change*, (3) learning rates, (4) levels of exploration, and (5) functional forms. The results are robust in all cases. First, we altered levels of *Automation* to include $\tau = 0.25$, $\tau = 0.75$, and $\tau = 1$. Second, we evaluate levels of *Environmental Change* when $\delta = 5$ and $\delta = 100$. We also change the operationalization of *Environmental Change* from punctuated to smooth changes. While the number of door changes per simulation run, δ, remains unchanged, the door changes occur at different episodes across simulation runs. In the aggregate across all simulation runs, the learning trajectory appears smooth rather than punctuated.[10] The results are robust to the alternative levels of *Environmental Change* and the alternative operationalization. Third, we set the learning rate (α) and the learning advantage (ν) in Equation 2.8 to pairs where ($\alpha = 0.3$, $\nu = 0.35$) and ($\alpha = 0.7$, $\nu = 0.15$). Fourth, we examine the results with lower ($\epsilon = 0.05$) and higher ($\epsilon = 0.15$) exploration rates. Our results for alternative learning rates and exploration rates remain robust. Finally, we vary the functional form specifications of the goal reward (we replaced the natural logarithm in Equation 2.7 with the logarithm with base 2 and 10, the square root, and

[10]We include the validation results of smooth *Environmental Change* in Appendix A.3.

$1 - \frac{1}{\lambda+1}$), learning rate (we replaced the natural logarithm in Equation 2.8 with the logarithm with base 2 and 10, the square root, and $1 - \frac{1}{\lambda}$), and coordination costs (we replaced the exponent equal to $\frac{5}{3}$ in Equation 2.9 with $\frac{4}{3}$ and 2) to find that the results remain robust.

2.4.6 Density Estimates

To gain a deeper understanding of the impact of automation routines on the net reward, we inspect the distributional characteristics for different levels of *Automation* with intermediate *DOL* ($\lambda = 2$). Figure 2.9 below shows the Gaussian kernel density estimates of accumulated net rewards by levels of *Environmental Change* and levels of *Automation*. While the densities for low and intermediate *Automation* are approximately normally distributed across all levels of *Environmental Change*, the density for high *Automation* resembles a bimodal distribution in stable environments (Panel A) and highly dynamic environments (Panel C). Further examination of the standard errors with Bessel's correction shows that the variability of accumulated net rewards across simulation runs increases with *Automation*, irrespective of the level of *DOL* and *Environmental Change*. In short, the density estimates suggest that automation can increase the variability of net rewards because locking into an automation routine inhibits adaptation.

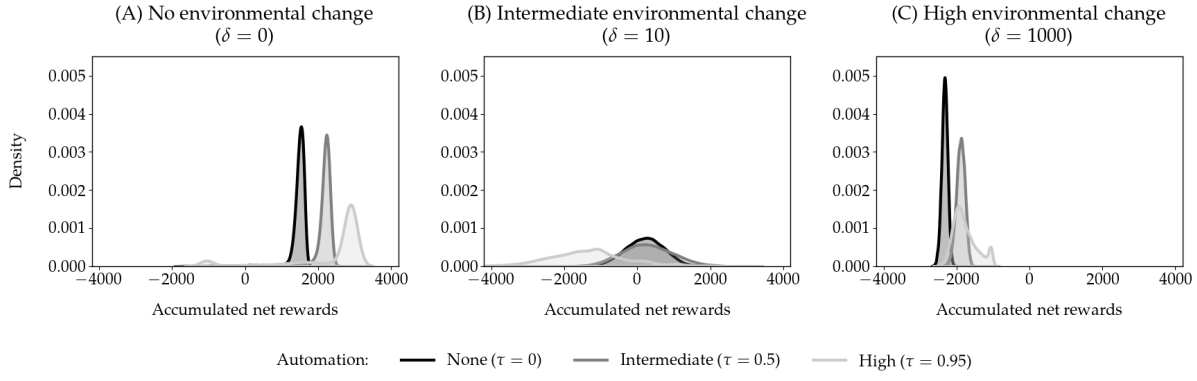

Figure 2.9: Kernel density estimate by levels of *Environmental Change* with intermediate *DOL*.

One might find it counter-intuitive to see that outcome variability increases under automation, given that automation cannot select random ϵ actions. The reason for observing this result lies in how we operationalized *Environmental Change*. There are two paths to the goal for each domain in our grid-world. When automation executes the short path and the door closes, the net reward per episode is at the lowest possible value. However, when automation executes the short path and the door is open, the net reward is the highest possible value.

Given that automation does not generate new knowledge, our results correspond to the observation by March (1991, p.83) that "increased knowledge seems often to reduce the variability of performance rather than to increase it." March argues that knowledge can reduce the variability in the time it takes to complete a task and the quality of task performance. Similarly, the density estimates show that more knowledge (i.e., no *Automation*) reduces the variability of net rewards.

2.4.7 Analysis of Learning Trajectories

In this section, we disaggregate the results above to understand the dynamic learning behavior for different levels of *Automation*. We focus on the case of intermediate *DOL* ($\lambda = 2$) and intermediate *Environmental Change* ($\delta = 10$). The results for the other combinations of values of *DOL* (λ) and *Environmental Change* (δ) are in Appendix A.1.

We first calculate the set of possible episodic net rewards based on Equation 2.6. The goal reward per episode equals $\pi \cdot ln(\lambda + 1) = 10 \cdot ln(2 + 1) \approx 11$ (Equation 2.7) and the coordination costs per episode equal $\eta \cdot (\lambda^{\frac{5}{3}} - 1) \approx 0.9 \cdot (3.17 - 1) \approx 2$ (Equation 2.9). In an episode through the grid-world (Figure 2.3), agents make six steps when attempting a closed door, seven steps for the short path to the goal, and 11 steps for the long path to the goal. The cost of a single step equals 1. Table 2.4 shows the possible episodic net rewards by *Automation Mode* and *Adaptation Mode*. One can see that the short path is the most rewarding but risky path to the goal.

	Automation Mode	*Adaptation Mode*
Short path (open door):	$11 - 7 - 0.5 \cdot 2 = 3$	$11 - 7 - 2 = 2$
Long path:	$11 - 11 - 0.5 \cdot 2 = -1$	$11 - 11 - 2 = -2$
Short path (closed door):	$0 - 6 - 0.5 \cdot 2 = -7$	$0 - 6 - 2 = -8$

Table 2.4: Possible episodic net rewards by mode.

Myopic Automation

Figure 2.10, Panel (A), shows the episodic net rewards[11] for different levels of *Automation*. One can see the dips in the $\delta = 10$ episodes in which doors can open or close. The net reward for those episodes drops as the agent hits the newly closed door and fails to reach the goal reward, resulting in the lowest possible net reward, as calculated above. High *Automation* performs well up to the first environmental change at episode 91. After the first door change, however, automation incurs low net rewards as it is unable to re-learn the changes in the environment.[12] The episodic net reward smaller than -1 suggests that automation might be locked into the short path with a closed door for some time. Automation behaves myopically as it locks into the first path it can find to reach the goal state, optimizing the short-run while overlooking the long-run consequences of such a lock-in (Levinthal and March, 1993). Panel (B) shows the cumulative net rewards, underlining the detrimental consequences of myopic automation to the organization over time. The high initial net reward can provide a misleading signal that the automated routine performs well.

While the *average* episodic net reward of high *Automation* in Figure 2.10, Panel (A), is low, the *minimum* episodic net reward is consistently higher than for lower levels of *Automation*. We observe this behavior because, without *Automation*, the agent is more likely to find the short-but-risky path through an open door that yields the worst possible outcome when the door closes unexpectedly.

[11] The episodic net rewards represent a mixture of realized net rewards across 10,000 simulation runs.
[12] Figure A.2 in Appendix A.2 validates that the opportunity costs increase after the first environmental change.

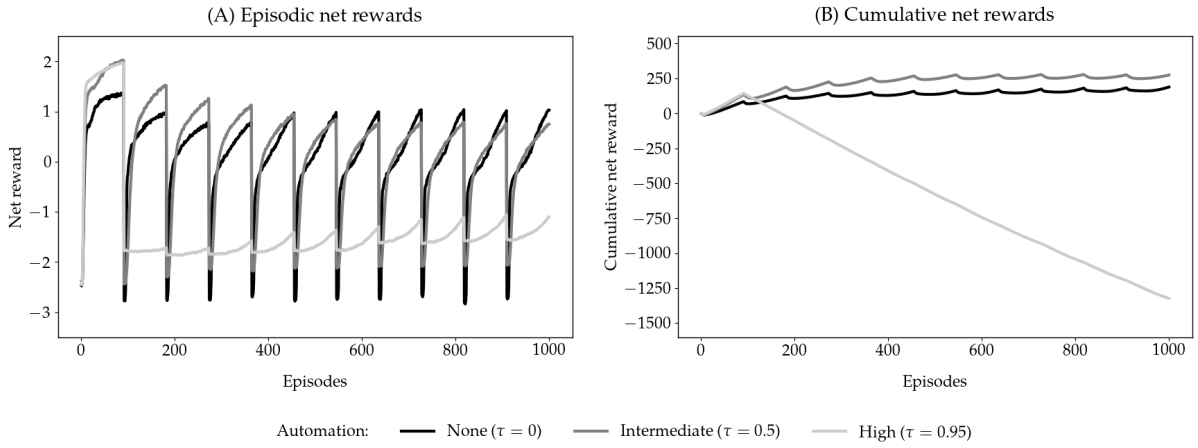

Figure 2.10: Net rewards at an intermediate *DOL* and intermediate *Environmental Change*.

As such, the inability to explore under automation can act as a hedge against the brief but very low net rewards before adapting away from the closed door.

Failure to Adapt

Myopic automation can be detrimental because the organization is unable to efficiently adapt away from the short path when the door closes. Figure 2.11 shows the average path shares across simulation runs per episode by levels of *Automation*. The dark grey area shows the share of paths where the agent attempts a closed door. These dark grey areas spike only for a few episodes under no and intermediate *Automation* in Panel (A) and Panel (B), indicating that agents can re-learn and adapt. High *Automation* in Panel (C), however, takes a long time to adapt, frequently incurring the lowest possible net rewards.

Summarizing across all results, we find that automation can partially escape the coordination costs that limit the division of labor (*Proposition 1*). However, automation can incur hidden opportunity costs due to lost learning and slow adaptation (*Proposition 2*). Moreover, automation can exhibit myopic behavior when it locks into the first viable path with negative consequences for the long-term net rewards. Our approach to studying the impact of automation differs from

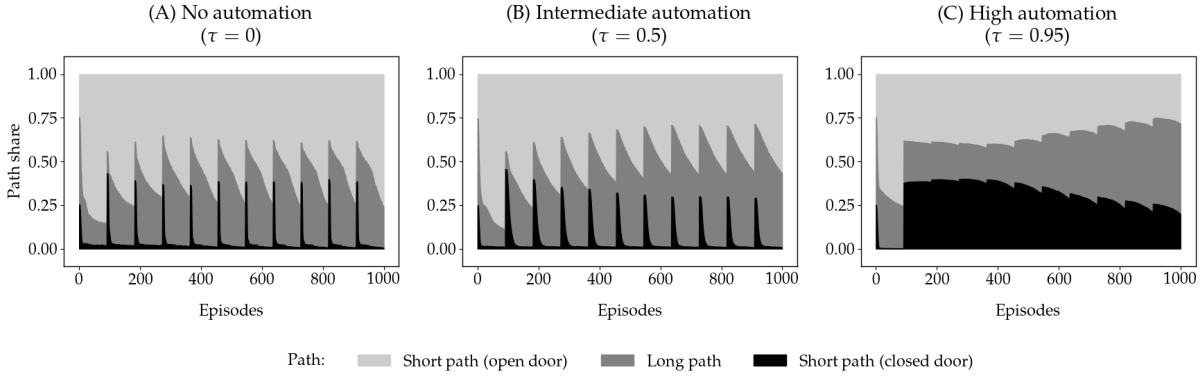

Figure 2.11: Path shares at an intermediate *DOL* and intermediate *Environmental Change*.

existing research (e.g., Dogan and Yildirim (2021)) in that our simulation accounts for the inability of automation to learn new knowledge and adapt (Autor, 2015). The results provide novel evidence that automated routines can incur indirect opportunity costs that complicate capturing the returns from the division of labor. In the next section, we discuss how organizations can manage these hidden automation costs.

2.5 Discussion

Automation can introduce a tension between economizing on coordination costs and incurring opportunity costs in the context of dynamic environments. We designed a reinforcement learning simulation to investigate how automation and environmental change interact. The results show that automation can partially escape the coordination costs that limit the division of labor. However, the results suggest that hidden opportunity costs due to lost learning and slow adaptation can reduce the returns to the division of labor and result in myopic automation behavior. Given these results, we now discuss ways organizations can manage this tension when adopting automation in an organization. We anchor our discussion in the research on dynamic capabilities and organizational ambidexterity.

2.5.1 Dynamic Capabilities

Teece et al. (1997) define the concept of dynamic capabilities as "the firm's ability to integrate, build, and reconfigure internal and external competencies to address rapidly changing environments", but leaves open how organizations can develop and refine such dynamic capabilities. In theorizing about the development and evolution of dynamic capabilities, Zollo and Winter (2002) re-define dynamic capabilities as "a learned and stable pattern of collective activity" that generates and maintains an organization's operating routines. Zollo and Winter refer to operating routines as the organization's operational activities. They argue that learning mechanisms for accumulating, articulating, and codifying knowledge shape an organization's operating routines in two ways. There exists a direct link between learning and modifying operating routines and an indirect link where learning impacts dynamic capabilities, which impact operating routines (Figure 1 in Zollo and Winter (2002, p.340)). The authors conceptualize dynamic capabilities as "the firm's systematic methods for modifying operating routines." One can regard automation as an operating routine that evolves through the co-evolution of learning mechanisms and dynamic capabilities. While the results in this study show how automation can codify and execute existing knowledge, we build on Zollo and Winter to discuss how managers might develop dynamic capabilities to manage automation routines systematically.

For example, consider a commercial bank with an organizational routine for ensuring the security and compliance of transactions. Today, banks like HSBC use machine learning fraud detection systems to identify fraudulent transactions (Wilson and Daugherty, 2018). Zollo and Winter (2002) separate routines depending on the rate of environmental change. In stable environments, it might be enough for bank employees to complete a single learning episode, automate the knowledge they have accumulated about detecting fraudulent transactions, and improve the routine incrementally. In environments of rapid change, however, using the same sequence of actions can be hazardous

because the value of existing value can deteriorate (Posen and Levinthal, 2012) as our results on myopic automation show. A dynamic capability can modify the automated operating routine through accumulating new knowledge (e.g., on how the characteristics of fraudulent transactions change), articulating knowledge (e.g., coordinating with team members to identify appropriate modifications), and codifying knowledge (e.g., re-training the predictive model to identify fraudulent transactions accurately) (Zollo and Winter, 2002). One can view a dynamic capability for automation as a systematic pattern of collective activity that integrates learning with automation, ensuring the continuous fitness of automation routines. As such, the managerial challenge might lie in handling the tension between flexibility through learning and efficiency through automation.

2.5.2 Organizational Ambidexterity

One view of how managers can resolve the tension between flexibility and efficiency emphasizes ambidexterity. The organizational ambidexterity literature studies the balancing and synchronization between exploring new opportunities and exploiting existing capabilities (Tushman and O'Reilly III, 1996). Organizations can use various mechanisms to promote ambidexterity: Structural mechanisms include semistructures that balance order with a lack thereof (Brown and Eisenhardt, 1997), separating exploratory and exploitative activities into different integrated business units (Levinthal, 1997), and outsourcing a type of activities and establishing partnerships (Holmqvist, 2004). Furthermore, a particular emphasis lies on the role of managers to shape internal processes (Tushman and O'Reilly III, 1996) and contextual factors such as support and trust that can create ambidextrous capabilities (Gibson and Birkinshaw, 2004). Future research might further investigate how different ambidexterity mechanisms can impact the efficiency of automation.

Ambidexterity only becomes a dynamic capability to the organization when exploratory and exploitative activities are strategically integrated (O'Reilly III and Tushman, 2008). Our results

show that successful automation requires integrating up-to-date knowledge about the environment. Organizations that exclusively focus on learning and accumulating new knowledge miss out on the productivity benefits of automation. Organizations that exclusively focus on automating existing knowledge incur opportunity costs due to incongruity with the environment. Under this view, a dynamic capability for automation consists of integrated learning and automation activities that systematically refine and leverage organizational knowledge. As such, our discussion emphasizes the complementarity of humans and machines in capturing value from the division of labor (Autor et al., 2003; Wilson and Daugherty, 2018).

2.6 Conclusion

In this study, we investigate how automation can change the returns from the division of labor. Given the context of changing environments and behavioral learning, we design a reinforcement learning simulation to find that automation can be beneficial when efficient execution is more important than learning and adaptation. However, when adapting to environmental change is beneficial, automation can incur considerable opportunity costs due to lost learning and slow adaptation as well as increased return variability. In such environments, superstitious automation can lead to myopic behavior with detrimental consequences for an organization that aggressively automates.

Our insights have implications for how organizations manage automation. Based on our results, the discussion suggests that organizations can develop a dynamic capability for automation that combines learning mechanisms with automation to ensure that automation routines are congruent with the environmental requirements (Teece et al., 1997; Zollo and Winter, 2002). While the literature on organizational ambidexterity (Tushman and O'Reilly III, 1996) suggests a balance between exploitation of the known and exploration of the new, our results shift the focus toward balancing human learning with machine automation. These results are in line with recent studies

on the effect of automation on labor, which suggest that successful organizations leverage the complementary strengths of humans and machines in creating and capturing value from specialized knowledge (Choudhury et al., 2020; Agrawal et al., 2021).

Our study has several limitations that open up potential avenues for future research. First, the focus of this study is on the impact that automation can have on the returns to the division of labor in uncertain environments. Future research could investigate how endogenous automation strategies can balance the tension between economizing on coordination costs and incurring opportunity costs. Second, the current implementations of the benefits to the DOL and coordination costs are determined algebraically and are exogenous to the learning problem; however, firms might adjust the DOL and coordination mechanisms during learning. For example, coordination costs might relate to how frequently the organization decides to shift between knowledge domains and, therefore, between different agents. Future work might endogenize the DOL and coordination costs. The *OrgSim-RL* platform can serve as a starting point for implementing endogenous conceptualizations of automation, the DOL, and coordination costs. Second, the organization in our simulation currently pays no transition costs when starting or ending automation, and agents pay no "messaging costs" to communicate or internalize feedback from the environment. However, changing procedures (Feldman and Pentland, 2003) and sense-making about the environment (Weick, 1995) in the real world involve their own costs and dynamics. Future work could examine such factors.

Chapter 3

Data Clans: An Exploration of the Organization of Artificial Intelligence*

3.1 Introduction

"For more than 250 years the fundamental drivers of economic growth have been technological innovations . . . The most important general-purpose technology of our era is artificial intelligence, particularly machine learning."
— Brynjolfsson and McAfee (2017)

General-purpose technologies (GPTs) such as the steam engine, the personal computer, and the internet often require significant investments into complementary intangible assets such as business processes, knowledge, and software to create value for organizations (Brynjolfsson et al., 2021). Scholars have suggested that artificial intelligence (AI) exhibits characteristics of a GPT (Cockburn et al., 2018; Brynjolfsson et al., 2021) as it becomes increasingly pervasive, improves over time, and can spur complementary innovations (Bresnahan and Trajtenberg, 1995). While management scholars have investigated the role of AI for innovation (Cockburn et al., 2018), human capital (Choudhury et al., 2020), and the tension between human labor automation and augmentation (Raisch and Krakowski, 2020), for example, foundational strategic management theory emphasizes the importance of organizing strategic asset stocks (Barney, 1986b; Dierickx and Cool, 1989).

The resource-based view of the firm suggests that a sustainable competitive advantage requires the possession of strategic assets that are rare, difficult to trade (i.e., immobile), difficult to imitate, and difficult to substitute (Wernerfelt, 1984; Barney, 1991). While some strategic assets can be

*The content of this chapter is based on: Hofer, M. W. (2022). Data Clans: An Exploration of the Development of Artificial Intelligence Assets in Incumbent Organizations. Under review at the *Journal of Engineering and Technology Management*.

bought and sold on strategic factor markets (Barney, 1986b), other assets cannot be traded and must be accumulated and developed inside a firm (Dierickx and Cool, 1989). In particular, intangible assets such as knowledge, organizational culture, and custom software systems are often developed internally as intangible assets can be co-specialized and interdependent, which makes them difficult to decompose and trade on strategic factor markets (Teece, 1998). So far, the asset stocks necessary for developing and organizing AI capabilities have received little attention in the management literature.

AI is a new technology of economic and organizational significance (McAfee et al., 2012; Von Krogh, 2018) as the following industry observations suggest. The 2021 McKinsey Global Survey on AI, based on responses from 1,843 participants, finds that 56% of the organizations surveyed have adopted AI in at least one business function (Chui et al., 2021). Furthermore, the 2022 AI Index Report curated by Stanford's Institute for Human-Centered AI (HAI) reports that total private corporate investment in AI was $93.54 billion in 2021, more than double the total private corporate investment in 2020 (Zhang et al., 2022). Specific examples of recent real-world AI use cases in incumbent organizations include improving the drug discovery and development process at Pfizer (Fleming, 2018), streamlining the talent acquisition process at Unilever (Marr, 2018), tracking chicken inventory at Tyson Foods (Castellanos, 2020), detecting machine maintenance needs at E.ON (Evgeniou and Boza, 2020), and detecting fraudulent transactions at UBS (Walsh, 2020). At the time of this writing, the list of AI use cases is expanding rapidly.

One factor underpinning the organizational adoption of AI is recent technological innovations in computer vision and natural language processing. The following technical advances described in The AI Index Annual Report by Stanford University (Zhang et al., 2022) exemplify the recent technological progress in AI. In computer vision, a team from Microsoft Research, for example, surpassed human-level performance in 2015 on the task of classifying images into one of the nearly

20'000 classes of the ImageNet dataset (He et al., 2015). In natural language processing (NLP), the SuperGLUE benchmark contains tasks such as question answering, reading comprehension, and common sense reasoning to evaluate natural language understanding (Wang et al., 2019). He et al. (2021) trained a deep learning model with 1.5 billion parameters to surpass human performance on SuperGLUE in 2021. Overall, the AI Index Reports show that the past decade has witnessed several technological innovations that organizations aim to exploit as the growing amount of corporate investments into AI underlines (Zhang et al., 2022).

Real-world machine learning (ML)[1] systems in an organization contain numerous components that are interdependent, which can complicate capturing value from a technological innovation as Teece (1986) points out. Sculley et al. (2015) draw attention to such complex interdependencies between components, conceptually shown in Figure 3.1[2]. For example, the authors describe *data dependency*, in which case there might be one engineering team responsible for the input data and another engineering team responsible for maintaining the statistical learning model. In such organizational settings, updating the input data by re-calibrating particular features might improve data quality but can negatively affect the system consuming these new data. Put differently, the data asset and the modeling asset are codependent. Another example is a *code dependency*, which broadly refers to situations in which software code depends on other code (Sculley et al., 2015). The authors describe how general-purpose ML software packages might require engineers to write "glue code" to format the data going into the package and coming out of it. As package specifications can change, maintaining such glue code can be expensive. In short, technology-related asset stocks for real-world ML systems can contain complex interdependencies between individual assets.

[1] Consistent with Raisch and Krakowski (2020), we view AI as comprising rule-based systems and machine learning (ML) approaches. Rule-based systems rely upon explicitly defined, discrete rules to make predictions. In contrast, ML algorithms can recognize patterns in large amounts of data. In this study, we use a broad definition of AI, containing related areas of ML, rule-based systems, business analytics, and big data.

[2] Reproduced with the written permission of the authors.

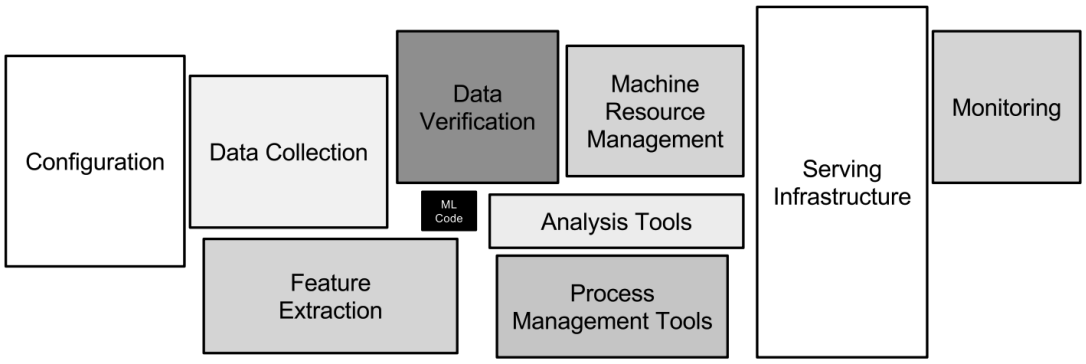

Figure 3.1: Components of real-world ML systems (Sculley et al., 2015).

Interdependencies, however, do not only exist in the technical realm of AI. The strategic management literature has found interdependencies between information technology (IT) and organizational assets. For example, Powell and Dent-Micallef (1997) investigate the retail industry to discover that ITs alone do not generate sustainable performance advantages but that firms use IT to leverage existing assets to capture value. In particular, the authors find that using IT to leverage human assets, such as a flexible culture, is particularly predictive of overall firm performance. Similarly, Tambe (2014) show that investments in Hadoop, a technology to manage large amounts of data, are associated with 3% faster productivity growth, but only for firms with significant existing data assets and employees with complementary technical skills. Finally, Rock (2019) finds that the surprise launch of Google's TensorFlow, a machine learning software package, is associated with a 4-7% increase in market value, but only for firms with existing AI skills. In brief, the management literature shows interdependencies between technology-related and organizational assets such as human capital and large-scale proprietary data.

The discussion up to this point leads to the two main starting points for our investigation. First, we are witnessing an AI revolution of economic and organizational significance (Von Krogh, 2018). Second, existing literature examining IT and AI in organizations suggests that organizations coordinate a stock of interdependent assets to create and capture value from the new technology

(Tambe, 2014; Rock, 2019). These starting points lead to the research question, which is purposefully general to guide data collection and analyses without leading to "prior hypothesis bias" (i.e., confirmation bias) (Gioia et al., 2012).

Research Question: *What asset stocks do managers at incumbent organizations focus on to develop artificial intelligence capabilities?*

Given the novelty and complexity of AI solutions (Von Krogh, 2018), we argue that managers are important informants for understanding AI assets because (1) they focus their attention and (2) select initiatives to pursue. First, the attention-based view of the firm suggests that managerial attention is central to strategic choice, particularly when adapting to changing environments (Ocasio, 1997). For example, Eggers and Kaplan (2009) investigate attention in the context of incumbent firms to find that managerial attention toward new technologies is associated with faster entry while attention to existing technologies is associated with slower entry. Ocasio and Joseph (2005) suggest that attention structures can impact an organization's strategic plan, which guides the allocation of resources in organizations. Second, managers are agents of selection that allocate resources to build and organize strategic asset stocks for the organization (Burgelman, 1991; Mollick, 2012). Taken together, managerial attention and selection suggest that managers' perspectives are particularly salient for understanding AI assets, in line with recent calls for a renewed focus on managers in management scholarship (Aguinis et al., 2022).

Data measuring the uses of AI in organizations are scarce (Seamans and Raj, 2018). Therefore, we conduct an inductive study and analyze semi-structured interviews with 19 managers exposed to AI initiatives at nine incumbent organizations with an office presence in Switzerland. Based on managers' exposure to and interpretation of the subject area, we first computationally explore the topics that managers mention. The topic model analysis shows that managers talk about

human-related and technology-related aspects of applying AI and use cases in areas such as asset management, trading, and supply chain. Second, we manually code AI assets that managers describe during the interviews to find that managers talk most frequently about investments into building an organizational culture and developing data assets. Finally, in discussing the inductive implications for theory, we develop theory to suggest that a strong, data-driven organizational culture – what we define as a *Data Clan* – might be a mechanism to facilitate intraorganizational coordination of interdependencies between AI assets in unpredictable environments.

We aim to contribute to the management literature investigating AI in organizations (Von Krogh, 2018; Raj and Seamans, 2019; Raisch and Krakowski, 2020). First, the qualitative evidence shows that managers frequently talk about allocating resources to building a *Data Clan* to support AI initiatives. Second, we develop theory to suggest that a *Data Clan* might be a mechanism for intraorganizational coordination in the face of complex asset interdependence and unpredictable environments. As such, we argue that investigating human coordination mechanisms such as a *Data Clan* can be a useful approach to developing theory on organizing interdependent assets in dynamic environments. While industry evidence suggests the importance of culture for AI applications (e.g., Bean (2022)), we are the first to our knowledge to investigate AI asset stocks in the management literature. Finally, we aim to help practitioners adopt an informed and realistic approach to governing AI initiatives.

We organize the paper as follows. Section 3.2 reviews the theoretical literature on the resource-based view, asset accumulation, and interdependence. Section 3.3 describes the research design and data collection. Section 3.4 describes the sample. Section 3.5 analyses the interview data with an automated hierarchical topic model and manual coding to extract several inductively derived AI asset stocks from the data. Finally, we revisit the theoretical literature on intraorganizational coordination to discuss the results in Section 3.6. We conclude in Section 3.7.

3.2 Theory

Given our focus on the asset stock for AI in organizations and the potential interdependence between AI assets, we briefly review the theory on the resource-based view, asset accumulation, and asset interdependence in the context of new technologies like AI. While we describe the relevant literature prior to conducting interviews, we were careful to only start consulting existing literature in tandem with the interview data to avoid a potential confirmation bias (Gioia et al., 2012).

3.2.1 Resource-Based View

The resource-based view provides a framework for how firms with strategic asset stocks[3] can gain a competitive advantage (Barney, 1986b; Dierickx and Cool, 1989). Starting with Penrose (1959), researchers have proposed a resource-based view of the firm as an alternative to viewing strategic management as primarily focused on industry structure and market position (Porter, 1980). Barney (1986b, 1991) and Dierickx and Cool (1989), among others, have further developed the resource-based view, arguing that ownership of strategic asset stocks can lead to sustained competitive advantage and superior firm performance when assets are valuable, rare, inimitable, non-substitutable, and immobile. Barney (1986b) theorizes that assets can be bought and sold in "strategic factor markets", a conceptual market for trading factors necessary to implement a strategy. According to Barney (1986b), when strategic factor markets are perfectly competitive, the price of these assets will reflect the economic value they generate once they are applied to implement a strategy. However, strategic factor markets may be imperfectly competitive when firms have different expectations about the future value of a strategic asset. In other words, firms can generate value by

[3]We follow Dierickx and Cool (1989) to distinguish between "resources" and "assets" conceptually. In the words of Dierickx and Cool, firms allocate "resource flows to accumulate a desired change in strategic asset stocks." The asset stocks refer to all capabilities, organizational processes, (tacit) knowledge, and information firms can use to implement their strategies.

having better expectations about the future value of strategic assets (i.e., information asymmetry) or luck (Barney, 1986b).

3.2.2 Asset Accumulation

Accumulating strategic asset stocks, however, can be challenging. Dierickx and Cool (1989) point out that some assets (e.g., a reputation for quality) are not readily tradeable on a strategic factor market, requiring organizations to develop such assets internally (e.g., by following a rigorous set of quality controls). Consequently, Dierickx and Cool (1989) view organizations as containing a certain level of strategic asset stocks, which are accumulated by adjusting appropriate time paths of resource flows over a period of time. Competitors will try to follow the accumulation paths to imitate an organization's privileged asset position. Dierickx and Cool (1989) identify several process factors that affect the relative difficulty through which assets are accumulated: time compression diseconomies, asset mass efficiencies, the interconnectedness of asset stocks, asset erosion, and causal ambiguity.

Issues around information, markets for ideas, timing, and measurement can further complicate asset accumulation. Arrow (1972) argues that generating and purchasing information, an intangible asset to the firm, is important, but contracting for information is difficult in the context of uncertainty. Moreover, markets for exchanging ideas and technologies are fundamentally incomplete (Gans and Stern, 2010), making effective trading of intangible assets difficult. Finally, as organizations adopt a new GPT like AI, they commit *measurable* resources to build largely *unmeasured* intangible asset stocks over prolonged periods of time. Brynjolfsson et al. (2021) describe the measurement aspect of this phenomenon, the *Productivity J-curve*, which posits that productivity growth will initially be underestimated because organizations invest measurable capital and labor to build intangible assets. Later, measured productivity growth overestimates true productivity

growth because of the benefits flowing from these hidden, largely unmeasured intangible assets (Brynjolfsson et al., 2021). In brief, the literature suggests that numerous factors can impede the development of valuable intangible asset stocks.

3.2.3 Asset Interdependence

Asset interdependence can be of particular importance in the context of a new GPT like AI because firms often fail to obtain significant economic returns from the technology alone, but rather, firms require suitable complementary assets to profit from the innovation (Teece, 1986; Brynjolfsson et al., 2021). Similarly, Clemons and Row (1991) argue that complementary strategic asset stocks are central to explaining differences in competitive advantage derived from advances in information technology (IT). Finally, Thomke and Kuemmerle (2002) also find in the context of drug discovery that the inimitability of a particular asset is often determined by its interdependence with other assets. In this study, we extend these insights to investigate further the role of interdependence between the assets required to develop and organize AI capabilities.

The resource-based view suggests that socially complex assets might be important complementary assets to technology for building competitive advantage. For example, an organizational culture, which Barney (1986a) defines as a "complex set of values, beliefs, assumptions, and symbols that define the way in which a firm conducts its business", can be a source of competitive advantage if it is valuable, rare, and difficult to imitate. Moreover, the complexity of social interactions can make it difficult to articulate why an organizational culture has value, generating causal ambiguity about exactly *which* asset stocks to accumulate (Reeds and De Filippi, 1990). As Barney (1991) points out, firms might possess the same technology, but only firms with the right culture and social relations can fully exploit a particular technology for implementing strategies.

Taken together, we argue that there is a need to investigate asset interdependence in the context

of AI for three main reasons. First, intangible assets are often interdependent and increasingly important for strategic planning in environments of rapid technological change (Teece, 2007). Second, for GPTs like AI to generate value for a firm requires significant investments into complementary assets that are largely unmeasured and, therefore, difficult to identify (Brynjolfsson et al., 2021). Third, there exists limited research investigating asset interdependence in the context of new technologies such as AI (for a notable exception in the context of pharmaceutical drug discovery, see Thomke and Kuemmerle (2002)).

3.3 Research Design

This section describes the research design and data collection for the study. We conducted semi-structured interviews to understand what assets organizations accumulate for developing AI capabilities. We designed the interview questions to allow for computational analysis of the interview transcripts in addition to qualitative analyses. Our inductive research design is in line with other inductive, theory-building studies related to resource-based strategies and technology-related assets (e.g., Brown and Eisenhardt (1997); Maritan (2001); Thomke and Kuemmerle (2002)).

Inductive studies are particularly useful in the early stages of research on new technologies for the following reasons. First, inductive theory-building is suitable when extant theory is of limited use, and novel insights are less likely to emerge from existing research or laboratory experiments (Glaser and Strauss, 1967). Rather, Gioia et al. (2012) emphasize the importance of structuring qualitative data as the basis for modeling the phenomenon of interest. Second, Eisenhardt (1989a) argues that inductive field research is particularly likely to generate novel theory to explain a new phenomenon or topic better. In sum, we argue that inductive theory-building from interview data can be a useful approach to exploring the emerging topic of AI asset interdependence.

Mixing quantitative and qualitative methods can strengthen conclusions compared to mono-

method approaches (Greene et al., 1989). In the field of strategic management, Molina-Azorin (2012) found that articles using mixed-methods approaches tend to receive more citations compared to monomethod papers and, therefore, have a larger impact. A possible explanation is that mixing methods may enable a better understanding of the research problems and complex phenomena (Creswell and Clark, 2017). We first conduct a computational analysis to understand the context in which managers in our sample operate. Second, we manually code the asset stocks that managers mention during the interviews and count codes as an indication of the asset stock's prominence across the sample (Hannah and Lautsch, 2011). We choose to run a computational analysis using a topic model *before* manually coding asset stocks for the following reasons (Hannigan et al., 2019). First, probabilistic topic models require minimal researcher input and interpretive rules on the data, mitigating concerns around researcher bias. Second, a topic model can identify topics that a human investigator might miss. In short, we mix computational and qualitative analyses to generate deeper insights and stronger conclusions from qualitative data.

3.3.1 Interpretive Assumptions

The interpretive investigation in this study builds on the following assumptions. First, organizational members actively create, shape, and enact the professional reality they inhabit (Webb and Weick, 1979; Ghoshal and Bartlett, 1994). In other words, managers make sense of the world to create their version of history, "symbolic records of actions" (Smircich and Stubbart, 1985, p.726), which they then use to predict and make sense of the future. Second, organizational members make interpretations *a posteriori*, meaning that the elapsed actions may be altered through what has happened since the action took place (Weick and Daft, 1983). Individuals in organizations can impact each other's interpretations of reality through social interactions (Daft and Weick, 1984). Third, the interpretations of managers are particularly salient, partly because managers are at the

center of organizing asset stocks through their actions, as discussed in Section 3.2. We put particular weight on what the managers say, treating them as "knowledgeable agents" that can explain their thoughts, intentions, and actions (Gioia et al., 2012). Within the literature on how organizations operate in an environment of rapid change, Keisler and Sproull (1982) found that managerial views are indeed critical to understanding the process of change and Aguinis et al. (2022) argue for a renewed focus on integrating managerial perspectives in management scholarship.

3.3.2 Theoretical Sampling of Managers

In theoretical sampling, one selects subsequent interviews for theoretical reasons – not statistical reasons – to replicate and extend emerging concepts (Glaser and Strauss, 1967), which can be preferable to random sampling when limitations to the sample size exist (Eisenhardt, 1989a). We argue that theoretical sampling is appropriate to investigate AI assets because we aim to build new theory based on qualitative data – rather than validate existing hypotheses – and develop insights into how asset stocks can help establish AI capabilities, which necessitates a deep understanding of the emerging concepts.

The starting point for data collection was the "Data Science for Managers" (DSFM) executive education course at EPFL. DSFM was a one-week course for managers interested in data science. The process of gaining a deeper understanding of AI in organizations began with informal conversations with managers, in which we learned that strategic assets and developing AI capabilities were topics of academic and practical relevance.

The unit of analysis is the manager and their understanding of their organizational priorities. For the first few interviews, we randomly sampled managers from DSFM, a selected group of managers particularly interested in AI. We then used theoretical sampling to gain access to managers outside of DSFM. For example, as particular strategic asset stocks emerged during the first few

interviews, we focused on sampling managers with profiles that could shed more light on allocating resources toward developing AI asset stocks. At the same time, we used triangulation of managers in different business functions (e.g., supply chain, sales, and product management) and at different hierarchical levels (e.g., vice president, director, project manager) to provide stronger substantiation of the emerging asset stocks.

3.3.3 Interview Format

During semi-structured, one-on-one intensive interviews between February and October 2020, each manager was asked about their perspective on the role of AI inside the organization, how they allocate resources to AI initiatives, and what they view to be the important factors for developing AI capabilities. Due to the Covid-19 pandemic, we conducted some interviews via videoconferencing software. While our theoretical interest lies in asset *interdependence* in the context of AI, the interview questions focus on resource *allocation* to adopt the managers' terminology in how they think about building and organizing internal asset stocks and to correspond to our purposefully general research question. We return to discussing the inductive implications for theory in Section 3.6.

To improve study validity, we also asked managers to define AI in their own words and then compared their definitions of AI to the conceptualization of AI in organizations by Von Krogh (2018) in order to clearly distinguish AI from other digital technologies (see Section 3.5; interview questions are in Appendix C). Even though some managers were not directly involved in the processes around organizing assets, we assume that the description they provided represents a dominant reality, which they would have learned from others within their organization (Gephart, 1984).

We kept an interview diary to note themes that emerged as we conducted interviews throughout

the interview process. We took field notes by hand during each interview to keep track of the main concepts, relevant comments raised off the record, and non-verbal observations (e.g., tone, body language, and gestures). In addition, we slightly adjusted the interview questions based on managers' responses to improve the reliability of the results (Charmaz, 2014). For example, the manager in our fifth interview noted that "when you say *investment*, you automatically assume a return on investment", which was something we also observed with earlier managers. Therefore, we changed the wording from "investment" to "allocation" to mitigate a possible instrumentation bias, which might skew managers to only talk about assets with a directly measurable financial return. Initial analyses and the detailed interview diary helped judge the point of saturation when information is repeated, and existing conceptual categories solidify such that "fresh data no longer sparks new theoretical insights, nor reveals new properties of your theoretical categories" (Charmaz, 2006, p.113).

3.3.4 Topic Modeling

Semi-structured interviews ensure that researchers ask the same questions in each interview while keeping some flexibility to ask follow-up questions on particularly relevant topics during the interview. As such, the interview transcripts can be used for computational analyses with a topic model, similar to Huang et al. (2018), who apply a topic model to transcripts of conference calls. Topic models are a computational tool useful for inductively discovering constructs in textual data, such as interview responses (Roberts et al., 2014). One can also view topic models as a more recent technique for visualizing and presenting qualitative data without inhibiting the meaning of the data through intensive coding (Miles and Huberman, 1984). In essence, topic models cluster word co-occurrences in text data to suggest topics. We select a topic modeling approach to set the context of what interviewees discuss. Moreover, a topic model can mitigate the potential issue

of investigator bias in qualitative analyses and triangulate emerging concepts vis-à-vis qualitative analyses to converge to stronger conclusions (Greene et al., 1989).

A potential challenge when using a topic model is that managers use different words to refer to the same concept. 90% of managers in the sample were non-native English speakers. While humans can judge relatively well when managers refer to the same concept using different words, the topic model algorithm might not. We mitigate this issue by manually coding the asset stocks that managers mention. Furthermore, we use a topic model that incorporates polysemy by allowing individual words to appear across different topics with different likelihoods (DiMaggio et al., 2013), e.g., when managers use the same word to refer to different things. Mixing methods and using a topic model that can model polysemy mitigate issues from managers using language differently.

Several different topic models exist, each with its advantages and disadvantages. The most well-known type of topic model is likely the latent Dirichlet allocation (LDA) model (Blei et al., 2003), which models documents as random mixtures of latent topics, where the topics themselves are distributions over words. Since the introduction of LDA, other variants of probabilistic topic models have been developed for different applications (Blei, 2012). A more recent variant is the hierarchical stochastic block model (hSBM) (Gerlach et al., 2018). The main innovation of the hSBM relative to the LDA is that the former is not limited to extracting a fixed number of topics determined by the researcher. Instead, the hSBM automatically detects a hierarchical structure of topics and subtopics to represent text data. More specifically, the hSBM first represents the text corpus as a bipartite graph of documents and words. Next, a generative, mixed-membership stochastic block model with non-parametric priors detects hierarchical clusters of documents and words. Finally, we interpret the hierarchical clusters of words as topics. We select the hSBM because it requires minimal researcher input and visualizes hierarchical topic relationships, which might be relevant given our interest in asset interdependence. We use the Python implementation

of hSBM that is publicly available on Martin Gerlach's GitHub profile[4] in combination with Tim Hannigan's pull request on GitHub from 29 November 2019 that prints the topic order on the visualization.

3.3.5 Manual Coding

We code the particular AI-related asset stocks that managers mention. To start the initial hand coding, we identify all assets that managers mention and summarize those text fragments in the transcripts. Then, we continue with line-by-line coding to identify granular leads that emerge across managers' responses, resulting in a list of first-level codes. Finally, we associate first-level codes with managers, who can be associated with at most one first-level code, even when they repeatedly mention the same code, to ensure that the analysis is not biased toward managers that talk more. Next, to triangulate the conclusions reached by purely qualitative hand-coding and to strengthen the conclusion, we also count first-level codes as a form of corroborative counting (Hannah and Lautsch, 2011). While we do not claim that higher frequency counts automatically imply higher importance than lower frequency counts, managers tend to talk more about topics they find more important than topics they find less important. As such, frequency counts might be useful to triangulate between insights generated by the topic model, our field notes, and verbatim quotes. Finally, we follow the procedures of Corley and Gioia (2004) to sum up the frequencies of initial codes and iteratively move to second-level, aggregate codes, and create an overall "data structure".

3.3.6 Reliability and Validity

To improve reliability in the analysis stage, we iteratively code the transcripts, link emerging concepts to quotes, and use topic modeling, as described in Section 3.5. In addition, investigating

[4]See `https://github.com/martingerlach/hSBM_Topicmodel`, accessed 22 June 2022.

the interview data with mixed methods and actively comparing concepts across interviews mitigates concerns about information-processing bias, which can lead to immature or false conclusions (Eisenhardt, 1989a).

To improve the construct validity of the interview questions, we conducted and analyzed two pilot interviews. We used the feedback to refine the interview questions and mitigate concerns related to, e.g., leading questions (Chenail, 2011), removing references to *specific* assets in the questions. In the data-gathering stage, we kept a detailed interview diary, took detailed field notes during the interview, and compared these notes with the transcripts for analysis. Taken together, we applied precautionary steps toward more reliable and valid conclusions.

3.4 Sample

Table 3.1 shows summary statistics of the anonymized sample. The sample consists of 19 managers that work for nine different organizations (2.11 managers per organization, on average). The managers' job titles suggest direct exposure to the organizations' technology initiatives and include Head of AI, Group IT Manager, Director of Research, Head of Advanced Analytics, Head of Global Product Management, and Senior Data Scientist. To characterize the sample in greater detail, we include information about managers' work experience, tenure at the current employer, whether they work in a team dedicated to analytics, and whether they have a PhD degree. Managers have an average of 19.61 years of work *Experience* and an average of 11.53 years of *Tenure* with their current employer. Eleven managers (57.89%) work in teams *Dedicated* to big data and AI initiatives. 10 managers (52.63%) have completed a *PhD* degree. Seven managers (36.84%) graduated from DSFM. On a firm level, the average founding year of the nine organizations is 1939 (the median is 1971), with the oldest organization founded in 1836. According to publicly available records, the nine organizations had an average revenue of more than $26bn and a median revenue

of more than $11bn, both in 2019. The nine organizations had an average of 28'000 employees and a median of 7'000 employees. All organizations have an office presence in Switzerland. At the end of the interviewing process, we transcribed all 19 interviews, covering 18 hours and 52 minutes of dialogue, into plain text documents.

Organization	Industry	Manager	Job title	Experience	Tenure	Dedicated	PhD
Verbier	Consumer electronics	V1	Senior Director; Head of AI	23	23	1	0
		V2	Senior Data Scientist	15.5	2.5	1	1
		V3	Senior Manager for Data Insights	20	6.5	0	1
		V4	Senior Product Data Analyst	10	1	0	1
Engelberg	Utilities	E1	Head of Business Management	25.5	11	0	0
		E2	Head of Merchant Trading	20	20	0	1
		E3	Head of Advanced Analytics	14	8	1	1
		E4	Head of Strategy	14	10	1	1
Chamonix	Telecommunications	C1	Director of Research	9.5	4	1	1
		C2	Open Innovation Director	15	7	1	1
		C3	Principal Product Manager	6	5	1	1
Zermatt	Food production	Z1	Senior IT Manager	29	19	1	0
		Z2	IT Business Partner	42	9	0	0
Davos	Investment management	D1	Partner Private Equity	17	14	0	0
		D2	Head of Quantitative Research	18	3	1	1
Andermatt	Information Technology & Services	A1	Sales & Channel Lead	32	30	1	0
Saas Fee	Pharmaceuticals	S1	Head of Site Quality	13	6	0	0
Wengen	Electrical & Electronic Manufacturing	W1	Group IT Manager & Vice President	25	25	1	0
Flumserberg	Textiles	F1	Head of Global Product Management	24	15	0	0

Notes: *Experience* represents the years of work experience after the last degree. *Tenure* represents the years of tenure at the current organization. The *Dedicated* variable equals 1 if a manager works in a team dedicated to big data and AI, 0 otherwise. The *PhD* variable equals 1 if a manager has completed a PhD degree, 0 otherwise. Organization names are pseudonyms for preserving confidentiality.
Industry categorization are retrieved from company profiles on LinkedIn.

Table 3.1: Descriptive statistics ($n = 19$).

3.4.1 Internal Validation

As the first step in describing the interview data, we aim to understand how managers define AI in their own words. Managers generally defined big data and AI as making sense of vast amounts of data, primarily through machine learning (ML) methods. Managers also mentioned several AI-related projects at their organization to contextualize what AI means to them. Examples of such projects include automating repetitive administrative tasks, recommendation systems for trading,

and face detection for a better product experience. Five managers view big data as data that require new tools and methods for processing, and four managers mention the four V's of big data (volume, variety, velocity, and veracity), a characterization of data that originated in the three V's (volume, variety, and velocity) described by Laney (2001). Taken together, managers view AI in organizations as a set of technologies combining computational and statistical methods with large amounts of data to generate business value.

16 managers (84.21%) note that AI is important or very important to their organizations. The six organizations that these 16 managers work for have made concrete investments into AI, including hiring engineering talent, establishing a central data management system, and setting up a dedicated function to centralize AI efforts. Furthermore, all 19 managers note that AI has become more important to their organization over the past two or more years. One might argue that these managers have a vested interest in the success of AI because their careers likely benefit from AI becoming more important. While managers might overestimate the true importance of AI to the organization, their vested interests should not be an issue for investigating AI asset stocks. On the contrary, having a personal stake in the success of AI initiatives might make managers more engaged with building strategic AI asset stocks, which could benefit our analysis.

3.4.2 Text Preprocessing

Given the raw interview transcripts, we separated interviewee text from interviewer text. We then trimmed the beginning and end of the interview because these sections did not contain data relevant to answering the research question, such as small talk. Next, we split all interview transcripts into individual responses to increase the number of documents, similar to Hannigan et al. (2019) and Mohr and Bogdanov (2013), who both split articles into paragraphs for topic modeling. We only keep responses longer than 140 characters to avoid known issues when using topic models with

short texts (Yan et al., 2013) and to automatically filter out short, non-informative responses, like "Okay.", "Yes.", and "Let's see.".

Next, we read through all transcripts to ensure that domain-specific language and abbreviations (e.g., AI, ML) were correctly transcribed. We filter out any hypothetical or theoretical statements that managers make about, for example, how allocation processes work "in theory". Through this preprocessing step, we aim to look beyond the current excitement around AI (Raisch and Krakowski, 2020) that might blur this investigation. Next, we lowercase all words, remove all "stop words" (e.g. "the", "a"), and stem all words to transform them into their roots (Kobayashi et al., 2018). As we go through these preprocessing steps, the number of words changes as summarized in Table 3.2. Starting from raw text, the total number of words equals 97,698, and the unique number of words equals 8,417. As one goes right, each column shows how the number of words changes for each preprocessing step. Eventually, 2,963 unique word stems are fed into the topic model.

	Raw text	Stop-words	Stemming
Total words	97,698	42,838	42,838
Unique words	8,417	4,683	2,963

Table 3.2: Dimensionality reduction through text preprocessing.

3.5 Results

This section reports results from two inductive analyses of the qualitative interview data. First, we use an automated hierarchical topic model to show that interview participants talk about technology-related and human-related topics as well as a range of AI applications, including asset management, trading, and customer support. Second, we manually code the transcripts to investigate what particular asset stocks managers develop.

3.5.1 Computational Analysis

The hSBM provides two major outputs: the network visualization in Figure 3.2 and the most likely word stems per topic (Table D.1 in Appendix D). To render the topic model outputs, we start with the raw subtopics from the hSBM algorithm to stay close to the data and label these subtopics as first-level codes based on the subtopics' most likely word stems and the interviewee responses with the largest share of a topic or subtopic. Second, we move up one level in the hierarchy and label the topics as second-level codes, identifying more abstract concepts and themes. Third, we analyze the topic prevalence in all documents to assess the semantic validity of the discovered topics and refine the topic and subtopic labels accordingly. In other words, we query the topic model for documents with a high prevalence for that particular topic or subtopic. A known behavior of the hSBM model is that it can cluster words frequently occurring across all documents (Gerlach et al., 2018). In our application, the hSBM detects three subtopics with uninformative words common in human conversations, such as "say", "let", and "right", which we do not show in Figure 3.2 for clarity.

In Figure 3.2, one sees that the topic model identified five high-level topics: *Computation*, *Project management*, *Business impact*, *Business needs*, and *Governance*. Each topic is further split into subtopics, resulting in 43 unique subtopics. Often, these subtopics are more specific, which facilitates labeling and refining the topic labels. We now briefly describe each topic and its subtopics in turn.

T1: Computation. Within the *Computation* topic, managers discuss the use of *Data* and algorithms *Algorithms* within the organization. Managers note that data are stored and configured in some internal data platform to make "data accessible" (manager C2[5]). Managers also mention two concrete AI use cases: *Asset management* and *Quantitative finance*.

T2: Project management. The *Project management* topic includes different aspects of

[5]The manager labels correspond to the "Manager" column in Table 3.1.

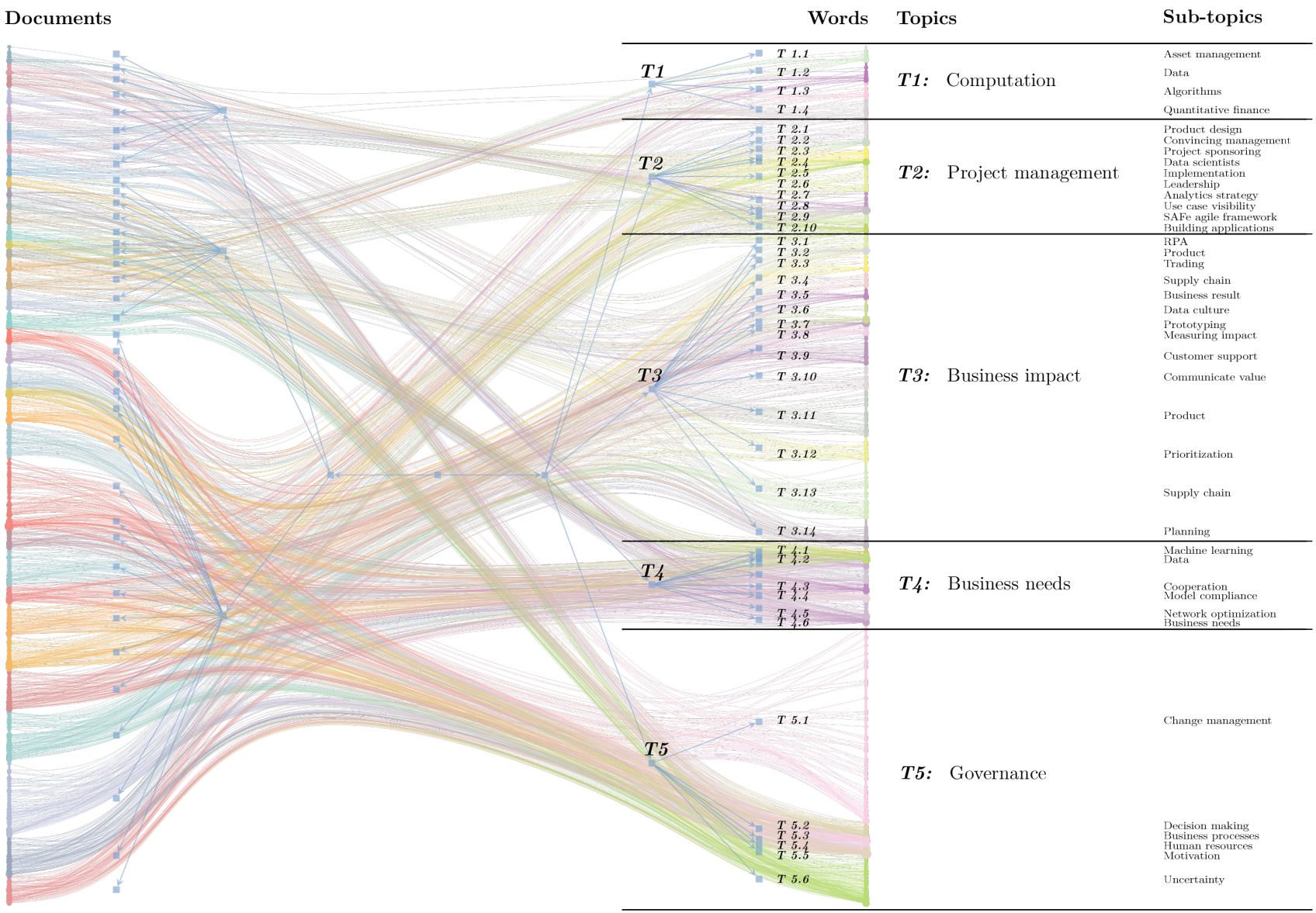

Figure 3.2: hSBM inference from all interview responses.

Notes: Inference from all 769 interview responses using the fitted hSBM topic model. The bipartite graph contains 769 document nodes (on the left), 1,877 word nodes (on the right), and 29,818 edges. For visual clarity, we only show 2,000 randomly sub-sampled edges. Topics are indexed from T1 to T5. Subtopics are indexed sequentially with reference to the topic they belong to, e.g., T 1.3 is the third subtopic in topic T1. The blue squares in the right half represent topics and subtopics. We align all subtopics with their labels to the right of the bipartite graph.

managing AI projects, including *Project sponsoring*, *Implementation*, *Leadership*, *Analytics strategy*, and *Building applications*. Managers further talk about *Convincing management* and *Use case visibility* as related aspects to managing AI projects. Within the *Leadership* subtopic, managers describe the important role of "technical leadership in the business" (manager V2) and how efforts go into "convincing our manager" (manager E1).

T3: Business impact. Within the *Business impact* topic, managers talk about how AI can impact their organization. Subtopics include *Prototyping*, building a *Product*, and *Measuring impact*. Another aspect of creating business impact is concerned with *Data culture*, *Communicating value*, and *Planning*. Managers also mention concrete AI use cases in *Trading* and *Supply chain*. Taken together, managers talk about building AI solutions and communicating the value of AI to generate business impact.

T4: Business needs. The *Business needs* topic splits into subtopics that discuss internal *Cooperation* between business representatives and engineers, mentioned explicitly by eight managers. Managers across different organizations also note specific *Business needs* as a common starting point for allocating resources to AI initiatives, mentioning that, "it starts with a business need" (manager F1), "above all its the business need" (manager D1), and "we really tried to have the business needs first" (manager C3). In other words, the topics and subtopics suggest that business needs can be a starting point for AI use cases, which demands *Cooperation* between the "business side and the technology side" (manager W1) around *Business needs*.

T5: Governance. The *Governance* topic describes change in organizations in the larger context of a "digital transformation" (managers E3, E4, Z2, S1, and A1). The subtopic of *Change management* spans many words, covering areas such as digitization efforts and employing external advisors to support the change efforts. Other subtopics include *Decision making*, *Business processes*, *Human resources*, *Motivation*, and *Uncertainty*.

Overall, the hSBM identifies various topics and subtopics related to humans and the organization of AI such as *Project management*, *Leadership*, *Data culture* to generate impact, and *Cooperation* around business needs. The hSBM also clusters topics and subtopics related to technology and engineering aspects such as *Data* (twice), *Algorithms*, and *Machine learning*. Finally, the hSBM uncovers particular use cases, such as *Asset management*, *Quantitative finance*, *Robotic process automation (RPA)*, *Trading*, *Supply chain*, *Customer support*, and *Network optimization*. These use cases provide context around potential practical AI applications that managers in our sample are considering. One can summarize the topic model analysis as managers talking about human-related aspects of organizing AI, technology-related aspects of applying AI, and concrete AI use cases that provide additional context.

3.5.2 Qualitative Analysis

With the insights and the contextual understanding from the computational analysis in mind, we use manual coding to investigate the specific asset stocks that managers mention for developing AI capabilities. Table 3.3 presents the data structure, the analytical framework that illustrates how we moved from raw data to codes to aggregate codes (Gioia et al., 2012). *Human Assets* and *Technology Assets* emerged as the two major aggregate codes of AI asset stocks. We will examine the four second-level codes, *Culture*, *Knowledge*, *Data*, and *Computation* in greater detail in the following sections.

Human Assets

Culture. The most frequently mentioned second-level code concerns *Culture*, which includes aspects of the top management team, a willingness to take risks, a culture of experimentation, a data-driven mindset, change management, and trust in data. The 31 counts of *Culture*-related

First-level code	Counts	Second-level code	Counts	Aggregate code
Top management support	7	Culture	31	Human Assets
Willingness to take risks	6			
Culture of experimentation & innovation	5			
Data-driven mindset	4			
Stakeholder engagement	4			
Change management	3			
Users' trust in data	2			
Domain knowledge	8	Knowledge	19	
Technical and communication skills	8			
Temporary employees (e.g. interns)	3			
High-quality data	16	Data	28	Technology Assets
Data engineering	4			
Data augmentation	3			
Data visualization	3			
Data generation	2			
Robust models	7	Computation	12	
Cloud computing resources	4			
In-house algorithms	1			

Notes: The table shows the result of iterative coding of interview transcripts to identify asset stocks at different levels of granularity, based on the method described in Corley and Gioia (2004).

Table 3.3: Data structure of AI asset stocks with counts.

codes correspond to 16 managers in our sample. Managers mention *Top management support* seven times. Toward the end of the interview with manager E3, we asked whether there were any particularly important aspects that we had not yet discussed. He responded,

> *"I would say what is very important is that you have trust and curiosity from the board, . . . particularly if you're in an early phase like us. And trust also means patience, right? You need to get . . . time to do this."*

Managers noted employees' *Willingness to take risks* six times. A closely related first-level code discusses a *Culture of experimentation & innovation* that encourages employees to be courageous and curious in trying out new things. In response to hiring versus re-training employees, Manager W1 pointed out that changing organizational culture is important. He noted,

"I think in certain areas we must look into changing minds, changing cultural mindsets by bringing in new people into the organization. We don't like to say that, but what I have seen so far with really, really young people with two, three years of development experience, so not too much, those great ideas we would not have got from people who have been with this company in a, let's say, different cultural setup."

One might argue that an "appropriate" culture is especially valuable for implementing AI strategies. Consider organization Engelberg, for example. Engelberg's board signed off a large advanced analytics initiative in summer 2019, employed external consultants, and hired an initial team of eight data scientists. Manager E3 led the implementation of this large initiative. The historical conditions at Engelberg are such that traders are key to generating revenue. Working effectively with traders requires a particular personality trait for new hires on top of engineering skills. When asked about hiring, manager E3 emphasized the importance of cultural fit in developing a valuable organizational culture. He noted,

"I also decided not to go to the super lead data scientist because you have to be very humble. The team now comes in and goes to a trader and says, "Look, I can show you how you can make more money." And this doesn't work if you think you're the king . . . You have to be very humble. So I think there were various elements let's say in those profiles that I thought were important."

Regarding the coordination aspect across teams that manager E3 alludes to, manager C2 mentioned that a common difficulty in the resource allocation process is the "interdependence between different teams" that can be "complicated to solve". On a related note, managers characterize the organizational environment under which they allocate resources as complex. For example, in his reply to the question about his challenges in allocating resources to AI projects, manager Z2 noted the "complexity of actors" and "complexity of the organization" as the two main factors. Similarly, manager A1 noted that complexity might be higher for incumbent organizations that are "big and have a big legacy". Finally, three managers explicitly mention *Change management capability*. Manager V3 noted that what makes an AI-related investment successful is "the change

management" and according to manager A1, "change management is very, very important".

To summarize, the *Culture* asset comes up frequently when managers talk about allocating resources to building AI capabilities. The interview data suggest that the *Culture* asset is concerned with topics around aligning people, creating trust, learning, and bringing about change in complex environments. In addition, the manual coding broadly corroborates the results from the computational analysis, which positioned the *Data culture* subtopic within the broader topic of *Business impact*, for example.

Knowledge. The other second-level code within *Human Assets* is related to *Knowledge* assets. The 19 counts of *Knowledge*-related codes correspond to 13 managers in our sample. Managers mention *Domain knowledge* and *Technical and communication skills* an equal number of eight times. Domain knowledge refers to understanding the business processes that can be "grown over time and is not always applicable in the same way across products" (manager Z1). The quote below suggests that AI requires considerable complementary, business-specific knowledge. When asked about the details of a recent project on using AI for automatic network optimization at organization Chamonix, manager C2 noted,

> *"This project is 85% domain knowledge, so trying to understand the antennas, and then 15% machine learning. Let me put it differently. It's 85% data wrangling, so understanding the data, being able to know the data. For that, you need domain knowledge, because otherwise, if you take random parameters, they're not going to work, and 15% is about machine learning . . ."*

Regarding *Technical and communication skills*, manager E3 explains how he allocated extra time to develop appropriate engineering skills that fit into the organization. In response to the question about which parts of the resource allocation process went well, he noted,

> *"What went well is the hiring. Well, we made an extra investment there in terms of that . . . I went through 250 CVs and . . . managed to hire six people out of them . . . I've never done so much effort in hiring, . . . but I think it was really worthwhile."*

Finally, managers find it challenging to judge whether an AI initiative was successful or not. More specifically, eight managers said that assessing the direct and indirect returns on investment was challenging to measure. For example, one can measure direct AI project returns in terms of cost savings. Indirect returns, however, are particularly difficult to measure, as manager V1 pointed out, saying that it was essentially impossible to tie the CTO team's work on AI over the last six months to future revenue. In short, managers talk about the challenges of assessing AI progress and performance within the organization.

To summarize the section on *Human Assets*, we find that managers talk about *Culture* and *Knowledge* assets. More specifically, managers explain how they allocate resources to building a culture that supports AI initiatives, developing domain knowledge, and developing technical skills and knowledge.

Technology Assets

Data. The most commonly mentioned *Technology Asset* concerns *Data*, which was mentioned 28 times, making it the most frequently mentioned first-level code. The 28 counts of *Data*-related codes correspond to 17 managers in our sample. Managers talk about allocating resources to develop *High-quality data*. Manager W1 describes allocating resources to develop and maintain reliable, accessible, and configured data. He said,

> *"If you want to have artificial intelligence, you need to have reliable data first. And in many cases this is exactly the issue. So, when we want to build an artificial intelligence to predict, for example, customer behavior in the market, it's only worth to build it if we have reliable data. Otherwise, it's going to be a kind of shit-in, shit-out topic."*

Data engineering aspects, such as Extract, Transform, and Load (ETL) pipelines, are technical processes that help managers and engineers access data. When asked about recent AI-related projects that she has been involved with, manager V4 described how she developed data pipelines

to enable others to leverage data. She noted,

> *"When I joined there was a need to start from ground zero, meaning we had . . . three year's data . . . that . . . was not leveraged at all, and this data was captured in a raw format, and my role actually, or the first thing I had to dig in was to prepare this data and build all those data pipelines so that the data can be consumed and leveraged. "*

Other data-related assets include *Data augmentation*, *Data visualization* for continuous monitoring, and exploiting existing products for *Data generation*, i.e., leveraging sensor data generated as a by-product by an organization's hardware. Concrete examples of how organizations use data generated by the organization's hardware are to "classify the defects" (manager F1) in the textiles industry and to "do face detection so that you can automatically pan, tilt, crop, and zoom on the number of people in the room" (manager V1) in video conferencing systems.

Computation. Another commonly mentioned type of *Technology Assets* is related to *Computation*. The 12 counts of *Computation*-related codes correspond to nine managers in our sample. Seven managers mention *Robust models*, including "neural networks" (managers C2, D2, and A1), "random forest" (managers V3 and A1), "XGBoost" (managers D2 and E3), "decision trees" (managers D2 and A1), and "linear regression" (manager C2). Given that these statistical models are often available as open-source packages maintained by universities, for-profit and not-for-profit organizations, and individuals (Thompson et al., 2020), managers allocate resources to understanding and applying these models. Four managers talk about *Access to computing resources*, especially computing resources hosted in the cloud. Particular third-party providers of such cloud services include "Amazon Web Services" (managers V1 and A1) and "Microsoft Azure" (manager W1). Managers talk about allocating resources to set up "a stable cloud set up" (manager E3) and "understanding the do's and don'ts of a cloud environment" (manager E3). Finally, one manager mentions that his team has developed *In-house algorithms* for a particular supply chain use case.

To summarize the section on *Technology Assets*, we find that managers talk about allocating

resources primarily to develop and maintain *Data* assets. For example, manager Z1 calls resource allocations to create a platform with high-quality, accessible data a "foundational investment" for using AI. The prominence of data is not surprising given the topic of the interview and the practical challenges to ensuring high-quality data infrastructure (Sambasivan et al., 2021). In addition, the manual coding broadly corroborates the results from the computational analysis, which identified the *Computation* topic and further technology-related subtopics such as *Data*, *Algorithms*, and *Machine learning*.

3.5.3 Summary of Results

To summarize the computational and qualitative analyses on developing AI capabilities under a resource-based view, we find that managers frequently talk about allocating resources to building an organizational *Culture* and developing high-quality *Data* assets. The results suggest that a strong, data-driven culture is central to building AI capabilities in an organization. On a high level, one can summarize the results as proposing that humans matter in AI projects, as the words of manager Z2 pointedly state. He concluded our interview by re-emphasizing that,

> *"Any of what we're doing in here is a huge change. And we are totally underestimating the change of any of those AI or Big Data projects. Because everyone thinks it's a technology project. And yet it's a human project."*

3.6 Discussion

> *"The more stable and predictable the situation, the greater the reliance on coordination by plan; the more variable and unpredictable the situation, the greater the reliance on coordination by feedback."* — March and Simon (1958, p.182)

We began by observing that AI is a new technology of economic and organizational relevance (McAfee et al., 2012; Von Krogh, 2018) that necessitates investments into interdependent assets

(Brynjolfsson et al., 2021). Some assets required to build and organize AI capabilities might not be readily available on strategic factor markets (Dierickx and Cool, 1989). Therefore, we explored the organizational asset stocks managers exposed to AI initiatives in incumbent organizations develop and accumulate. On a very high level, the interview results suggest that managers often talk about AI initiatives as a "human project", frequently describing the allocation of resources toward building an organizational culture that supports AI initiatives.

To examine the inductive implications for theory, we adopt a transaction cost perspective (Williamson, 1973). The results suggest that managers view the organization of AI primarily as a problem of building a strong, data-driven organizational culture. Given the complex and interdependent AI components discussed in Section 3.1 and visualized in Figure 3.1, managers might focus on a data-driven organizational culture as a way to govern intraorganizational contracting issues when building AI capabilities. This section aims to discuss a speculative explanation as to *why* managers emphasize solutions based on organizational culture to organize AI assets. We theorize that asset interdependence is a problem of coordination in the face of complexity and uncertainty, a context in which strong cultures might be particularly effective at reducing coordination costs.

3.6.1 Intraorganizational Coordination

In the context of interconnected assets, the terms "communication", "coordination", and "cooperation" have slightly different meanings. Communication is concerned with the exchange of information (Shannon, 1948). March and Simon propose that the greater the efficiency of communication within an organization, the greater its tolerance for interdependence among its component parts. Coordination is concerned with aligning interdependent organizational activities to complete collective organizational tasks (March and Simon, 1958) with shared and limited resources (Malone and Crowston, 1994). Kogut and Zander (1996) conceptualize organizations as systems

of coordination and learning that face a fundamental dilemma: productivity increases with the division of labor, but specialization increases coordination costs. March and Simon (1958) note that organizations can impact the volume of communication required within the organization by coordination by feedback instead of coordination by plan. The authors argue that communication can become more efficient through changing how to coordinate, and organizations can better manage highly complex interrelations. Cooperation between individuals occurs when a task is too large for a single person (Ouchi, 1980). Ouchi introduces transaction costs as a solution to the problem of internal cooperation, which necessarily requires mediating transactions through coordination between individuals. According to Ouchi, different governance mechanisms are more or less efficient at mediating transactions and bringing about cooperation.

Coordination is central to allocating shared resources (e.g., money, an employee's time) to develop capabilities for several reasons. First, coordination theory proposes that coordination means managing dependencies between activities, such as the activities that comprise the resource allocation process (Malone and Crowston, 1994). Second, Crowston (1997) finds that different coordination mechanisms for managing interdependent activities may also require or create resources that can improve organizational processes. In the context of software bug fixing, Crowston (1997) notes examples of such interdependent activities (e.g., writing code, integrating code with the rest of the system), coordination mechanisms (e.g., plans, schedules), and intangible assets (e.g., knowledge about problems, patch software). Third, Okhuysen and Bechky (2009) argue that coordination mechanisms can impact organizations by creating conditions of accountability, predictability, and common understanding, by which people accomplish their interdependent tasks. The coordination literature argues that it is not only assets that can be interdependent but also that activities to allocate resources to develop such assets can be interdependent.

There exists a tension between coordination by feedback and coordination by plan. On the one

hand, organizations focus more on reactive coordination based on feedback and new information under more variable and unpredictable conditions (March and Simon, 1958). For example, Adler (1995) argues that when applying novel technologies, a high degree of informal interaction and mutual adjustment between individuals can lower production costs. Similarly, Orlikowski (1996) finds that in situations of organizational transformation and change, actors coordinate through mutual adaptation, improvisation, and reaction – they are "going back and forth" – to manage interdependencies. On the other hand, organizations focus more on coordination based on pre-established plans under more stable and predictable conditions (March and Simon, 1958). For example, Argote (1982) finds that programmed means of coordination have more impact on organizational effectiveness under conditions of low uncertainty than under conditions of high uncertainty. Crowston (1997) argues that coordination based on plans and rules can mitigate dependency constraints on how tasks can be performed. In short, each of the two coordination mechanisms is more or less appropriate under different conditions.

3.6.2 Clan Culture

Strong organizational cultures can mitigate coordination problems under conditions of high uncertainty and complexity. Ouchi (1980) describes "clans" as particularly strong cultures that can reduce transaction costs compared to other governance mechanisms – the market and bureaucracy – under conditions of high performance ambiguity and high goal congruence. The shared understanding among employees in a clan can make governing particularly efficient under conditions of high uncertainty and high complexity (Wilkins and Ouchi, 1983). The inductive results on the challenges of assessing and measuring AI project success suggest that high performance ambiguity, high uncertainty, and high complexity are common in AI initiatives. Therefore, managers might focus on a strong culture to reduce coordination frictions through mutual adjustment and reaction.

The clan governance mechanism might also be effective in situations of organizational change. Under the market and bureaucratic mechanisms for mediating transactions, reciprocity and equity have to be satisfied by incurring economic costs, e.g., compensation for labor (Ouchi, 1980). However, Ouchi proposes that reciprocity and equity can also be met by strong socialization, which is the key governance mechanism of a clan. Topics such as a *Culture of experimentation & innovation*, *Data culture*, and *Change management* in Figure 3.2 and Table 3.3 suggest that human aspects related to socialization play a role in AI projects. Wilkins and Ouchi (1983) describes that clans are especially efficient in mediating transactions under conditions of high uncertainty and complexity through increased tolerance for internal contract misspecification. *Uncertainty* is a subtopic of the *Governance* topic (Figure 3.2). Taken together, building a strong group of supporters with congruent (not mutually exclusive) goals that establish a common understanding of the value of data can be an antecedent to enabling organizational change. Under this view, managers might focus on building a *Data Clan* as a mechanism to facilitate mutual adjustment and reaction through feedback to bring about organizational change.

Effective clans, however, generally require a long history and stable membership (Wilkins and Ouchi, 1983). Schein (1981), for example, investigates Japanese management styles and finds that a long history and relatively stable membership in a group or team are required to develop a complex social understanding. However, a long history is not typically found in situations of complex technological change. Instead, rapid technological change characterizes the environment of organizations that adopt AI technologies (Varian, 2018). Therefore, it is unclear how organizations can develop a *Data Clan* quickly and effectively. What might be an effective process to build a *Data Clan*? How do organizations maintain a *Data Clan*?

The present study establishes new insights for future research investigating how organizations can develop AI capabilities with *Data Clans*. One possible avenue is to run a survey that investigates

two aspects. First, the survey could examine *how* managers can go about developing a *Data Clan* in the absence of a long history and stable membership. Second, the survey could test the extent to which a *Data Clan* can support AI initiatives. Together, the survey could contribute to building a process model that describes how organizations build a *Data Clan* and subsequently utilize it to support AI initiatives.

3.6.3 Synthesis of Coordination, Clans, and Asset Stock Accumulation

To connect the implications with the initial discussion on the resource-based view, we propose that a *Data Clan* is a strong, data-driven organizational culture that can be a valuable asset stock for an organization to capture value by using AI for the following reasons. First, it takes time and effort for GPTs like AI to impact organizations. Brynjolfsson et al. (2021) find empirical evidence for what they call the *Productivity J-curve*, which describes how productivity growth is underestimated in the early years of intangible asset investments and overestimated at a later stage when harvesting the benefits from the investment. Second, socially complex resources can be imperfectly imitable because they are beyond the ability of organizations to manage and shape systematically (Barney, 1991). Third, the asset stock accumulation process can have causal ambiguity as to what discrete factors play a role in the accumulation process, even for organizations that already possess that asset stock (Dierickx and Cool, 1989). In short, a *Data Clan* might be a strategically valuable asset stock for an organization in the age of AI.

An alternative explanation for the prominence of a strong culture might be survivorship bias. More specifically, one might argue that the managers in our sample have exposure to AI initiatives because they use organizational culture as a ready scapegoat if projects do not progress as planned. So, managers focusing on culture have "survived" within their organization and might bias the results toward cultural aspects of AI. While we cannot exclude survivorship bias, it seems unlikely

to have a notable effect because it is unclear why culture would be a more effective scapegoat than knowledge or data, for example. In addition, participants' *Experience* and a count of *Culture* codes mentioned show a correlation of 0.13. One might also argue that the prominence of *Human Assets* compared to *Technology Assets* is due to the sample consisting of managers whose job is to manage people. While we cannot eliminate the possibility that our sample composition might bias the results toward human-related aspects, our research question necessitates interviewing managers, as argued in Section 3.1. Moreover, a considerable share of managers (52.63%) has completed a *PhD* degree, as Table 3.4 shows, suggesting that managers have some level of technical training and awareness. Finally, an alternative explanation for the prominence of *Data* might be selection bias, as all managers in our sample are exposed to AI projects that require data. While we cannot exclude selection bias, it is not clear that *Data* would be mentioned more frequently than *Computation* assets, which are also required for AI projects.

The insights in this study contribute to the management literature on the organization of strategic asset stocks in the age of AI. First, we extend our understanding of strategic assets and their interdependence in the context of AI by re-interpreting interdependence as a problem of informal human coordination. Second, a *Data Clan* might be a valuable strategic asset stock to the firm, particularly for economizing on intraorganizational coordination (Ouchi, 1980) and for adjusting flow variables to develop and maintain valuable asset stocks (Dierickx and Cool, 1989). Finally, the results inform practitioners working on AI initiatives about what assets might be relevant to developing AI capabilities. The insight of leveraging a strong culture for AI is in line with popular science accounts of the importance of building a "data-driven culture" (Davenport and Mittal, 2020), creating a "fertile environment" for implementing AI solutions (Ransbotham et al., 2019), and overcoming "cultural obstacles as the greatest barrier to becoming data driven" (Bean, 2022), for example.

3.7 Conclusion

To conclude, this study uses a hierarchical topic model and manual coding of semi-structured interviews to find evidence that managers who build AI capabilities focus on accumulating human assets and technology assets. Managers put particular emphasis on establishing a strong, data-driven organizational culture – something we call a *Data Clan* – to support AI initiatives. For the theoretical interpretation of results, we adopt a transactional perspective. The interpretation focuses on re-interpreting the interdependence between AI assets as a problem of informal human coordination. We propose that a *Data Clan* can be a governing mechanism under conditions of high uncertainty and complexity (Ouchi, 1980) by focusing on mutual adjustment and reaction through feedback (March and Simon, 1958). Under an asset accumulation perspective, a *Data Clan* shares characteristics with difficult-to-imitate strategic asset stocks that take time and effort to build and accumulate (Dierickx and Cool, 1989). Therefore, a *Data Clan* might be a valuable asset for building competitive advantage and, consequently, enable superior financial performance in the age of AI (Barney, 1986a).

The study has limitations. First, the sample size of 19 managers makes it difficult to generalize beyond the immediate research setting. Organizations of a particular size, industry, or country and their managers might have different perspectives on allocating resources to build AI capabilities. For example, firm size and internal structure can impact coordination costs (Malone and Crowston, 1994). Second, we do not account for the initial asset positions and conditions of organizations and managers (Cockburn et al., 2000). While some managers mention the role of existing assets, no robust patterns emerged.

The insights presented in this study open several avenues for future research. First, future research could investigate how organizations can go about building a *Data Clan* in the rapidly

changing environment of AI and assess the extent to which employees (as opposed to managers) reflect a *Data Clan* culture. In particular, future work could study whether there exists a difference between the managerial role and the execution role in leveraging a *Data Clan* to triangulate the managerial perspective developed in our study. Second, we argue that coordination might be an important strategic consideration when managing complex interdependencies arising from new GPTs such as AI. Future research might design a survey based on the results in this study to examine the mechanisms with which a *Data Clan* can facilitate the coordination of asset interdependencies to support AI initiatives. Third, one might focus on a particular industry or firm size to extract a richer understanding of the microfoundations for building a *Data Clan*. Finally, as the applications of AI in incumbent organizations become increasingly pervasive, further research opportunities will emerge for management scholars interested in AI in organizations.

Chapter 4

Risky Data:
The Disclosure of Technology Risk at IPO*

4.1 Introduction

Economists have documented how investments in information and communication technology (ICT) are becoming an increasingly large part of the economy (Corrado and Hulten, 2010). However, the intangible nature of a firm's technology-related assets, such as proprietary software and databases, can make them difficult to measure and evaluate (Heeley et al., 2007; Brynjolfsson et al., 2021) due to, in part, outdated accounting standards that do not capitalize many intangible investments in an increasingly intangible economy (Haskel and Westlake, 2018; Lev and Sougiannis, 1996). While scholars in finance and strategic management have investigated the role of technological innovations in the context of initial public offerings (IPOs) (Heeley et al., 2007; Morricone et al., 2017) and the effects of text-based information disclosure at IPO (Hanley and Hoberg, 2010; Loughran and McDonald, 2013), one aspect in the discussion on technology assets has received relatively little attention: technology risk disclosure.

Firms preparing for an initial public offering (IPO) are often conducting business in high-tech areas such as internet products, biotechnology, and science-based offerings that rely on intangible assets (Morricone et al., 2017; Lev, 2018). Given that the economic characteristics of intangible assets differ from those of tangible assets (Haskel and Westlake, 2018), Lev and Sougiannis (1996) argue that the quality and relevance of earnings reports might decrease, and investors might shift

*The content of this chapter is based on: Hofer, M. W. and Younge, K. A. (2022). Risky Data: The Disclosure of Technology Risk at IPO. Under review at *Research Policy*.

their attention to alternative sources of information. The IPO literature has investigated the role of alternative disclosure content (i.e., non-accounting information), particularly textual information disclosure. For example, Hanley and Hoberg (2010) found a positive effect of unique and informative textual disclosure in the IPO prospectus on underpricing, the difference between the issue price and the closing price at the end of the first trading day, and, therefore, the first-day return to an IPO investor. A particular type of textual information disclosure concerns potential risk factors for the IPO firm. Loughran and McDonald (2013) finds that the amount of risk-related words in the prospectus is positively associated with underpricing. Given that IPO firms often rely on technology assets for their offerings and text-based risk disclosures matter for underpricing, what is the role of technology risk disclosure in evaluating IPO firms?

This study aims to develop a new approach to measuring text-based technology risk disclosure at IPO and validate the measure with the *return-for-risk* association that finance theory suggests. It is not the intention of this study to identify a causal effect of the around the complex choice around disclosure and underpricing. The IPO context is particularly suited for a cross-sectional examination of technology evaluation because firms are legally required to disclose relevant risk factors before going public, IPO firms are at a similar stage of their life-cycle (Jain and Kini, 1994), and many IPO firms directly rely on technology to generate revenue (Lev, 2018). Given the choice to disclose or withhold risk factors, we argue that IPO firms aim to keep their technological know-how secret to mitigate the risk of imitation (Barney, 1991). Consequently, IPO firms might only disclose the technology risks that are necessary to avoid a post-IPO class action lawsuit (Hanley and Hoberg, 2012). We theorize that technology risk disclosure and underpricing are positively associated in a *return-for-risk* association. Moreover, patents might attenuate the *return-for-risk* association of technology risk disclosure as disclosing technology information through patents grants the IPO firm the exclusive right to exploit that technology, can facilitate the evaluation of technology assets, and

might signal technological capabilities.

Although risk is important for firm evaluation, the underpricing association of text-based risk factors has been complicated by (at least) three problems: First, text data cannot be readily transformed into observations for later use with econometric methods (Grimmer et al., 2022). Second, off-the-shelf topic models compute a pre-defined number of *risk topics*, not *risk magnitude* that can be compared to other firms in the sample. Third, managers decide which risk factors to disclose or withhold. As such, it is challenging to empirically distinguish between managers actively withholding risk disclosures and managers being unaware of risk in the first place. We aim to alleviate these complications by computing a new measure of text-based risk that we call *aggregate risk disclosure* to gauge the risk disclosed by the firm in the IPO prospectus. The new measure of text-based risk normalizes risk relative to a firm's year and industry peer group to enable a direct comparison of *risk magnitudes* (i.e., how much risk is disclosed) and *risk topics* (i.e., what risk factors are disclosed) across IPO firms. Our measure allows scholars to investigate particular risk topics such as technology-related risks. We describe the measure of risk and its advantages in more detail in Section 4.3.

Our results show that the new measure of risk exhibits a *return-for-risk* association, suggesting that investors require higher first-day returns for taking on more technology risk (Validation). Moreover, we discover that patents can act as an effective information disclosure mechanism to attenuate the *return-for-risk* association of technology risk disclosure (Main Hypothesis). The results are robust to changes in the measurement of technology risk disclosure and a firm's patent stock. Overall, the results suggest that formal intellectual property might allow the disclosure of technology risks without losing the competitive advantage.

The paper contributes to the literature by showing how text-based risk disclosures apply to evaluating a firm's internal assets generated through technological innovation (e.g., Chondrakis

et al. (2021)). Our insights also deepen the understanding of the role of text-based risk disclosure in the context of IPOs (e.g., Hanley and Hoberg (2010)). Finally, our application of probabilistic topic modeling contributes to the literature on using text data for economic and organizational research (e.g., Hannigan et al. (2019); Bybee et al. (2020)) by open-sourcing the *RiskyData-LDA* project[1]. To the best of our knowledge, we are the first to explicitly measure and investigate text-based technology risk disclosure in the IPO context.

We organize the paper as follows. Section 4.2 describes how the increasing importance of new technologies and intangible assets for strategic planning can shift attention from accounting disclosure to textual disclosure content, why risk matters for evaluating firms at IPO, the validation, and the hypothesis. Section 4.3 reviews challenges with measuring text-based risk disclosure and summarizes how we compute aggregate risk disclosure. Section 4.4 contains an overview of the data used, variable construction, and the regression model specification. Section 4.5 describes the empirical results, which provide evidence for the *return-for-risk* association of technology risk disclosure and the moderating effect of patents. Section 4.6 concludes by outlining our main findings, limitations, and contributions.

4.2 Theory

4.2.1 Technology Assets

Technology assets such as patents (e.g., Heeley et al. (2007)) and proprietary data (e.g., Brynjolfsson and McElheran (2016)) are increasingly important for strategic planning (Teece, 2007; Furr, 2021). Such assets are often intangible, which makes them difficult to quantify and evaluate (Lev, 2018; Haskel and Westlake, 2018). In the context of research and development (R&D) and firm

[1] *RiskyData-LDA* on GitHub: `https://github.com/mxhofer/RiskyData-LDA`, accessed 22 June 2022.

value, Griliches (1981) finds a positive relationship between the market value of the firm and its investments in innovation, which the author proxies by using the number of patents applications and past R&D expenditures. What is particularly relevant for the context of young firms preparing for an IPO is Griliches' comment that the valuation of R&D investments need not occur only after converting inventions into product sales but that the valuation reflects the "current present value of expected returns from the invention". As such, inventions can create intellectual capital that IPO firms commercialize after raising capital at the IPO, a common strategy in the biotechnology industry (Hermans and Kauranen, 2005). Technology assets[2] can hold important strategic value but can be inherently ambiguous and risky to evaluate due to their intangible nature.

4.2.2 Evaluation of Technology Assets

As the strategic importance of technology assets is growing, several characteristics contribute to making these assets inherently risky (Haskel and Westlake, 2018). First, Gans and Stern (2010) argue that designing and operating a market mechanism to trade ideas or technologies can be challenging, limiting price revelation of input factors required for developing technology assets. Second, many technology-related investments represent sunk costs, meaning that when their development stops, the entire investment amount will be lost (e.g., a drug under development that fails clinical tests). Third, the non-rivalrous nature and limited excludability of technology mean that non-owners can often benefit from someone else's technological innovation through reverse engineering products, for example. Fourth, assets created through technological innovation often have complementarities that can make the value of a combination of assets very unpredictable (Teece, 1986). For example, Haskel and Westlake (2018) describe how the MP3 protocol combined with miniaturized hardware and Apple's licensing agreements with record labels make up a very

[2]We use the term "technology asset" broadly to include patents, trademarks, licenses, proprietary databases and libraries, software, pharmaceutical formulations, and algorithms, among others.

valuable innovation: the iPod. In short, technology assets have economic attributes that can make them inherently risky.

Evaluating technology assets, particularly *internally-generated* ones, does not come without problems for analysts and potential investors (Heeley et al., 2007) as most measurement conventions ignore intangible assets (Haskel and Westlake, 2018). In the accounting literature, Lev (2018) summarizes how the US accounting standard (GAAP) and the international accounting standard (IFRS) require firms to expense most internally-generated technology assets such as software and business designs. Lev argues that the standard-setters exhibit a "resistance to change" the accounting practices to an increasingly technology-based business world with potentially harmful consequences for investors. Srivastava (2014) empirically shows that the earnings quality of successive cohorts of newly listed firms has been continuously decreasing since 1970, mainly due to the increasing intensity of technology assets. One possibility to provide investors with relevant information about technology assets is to use channels other than traditional accounting disclosures, such as textual disclosures in 10-K annual report filings or 424(b) filings when going public.

4.2.3 Information Disclosure at IPO

Many young and successful IPO firms depend on technological innovations to create and capture value in areas such as internet products and biotechnology (Heeley et al., 2007). Analyzing newly listed firms in the US, Doidge et al. (2018) observe that from the listing peak in 1997 to 2015, the likelihood for firms to list on a major stock exchange declined by 54%. On average, firms with fewer employees have seen the steepest decline in the propensity to list. While many factors are responsible for this decline, the authors point to the inability of accounting methods to reflect the value of technology assets. The authors suggest that investors might be more skeptical about the value of a firm in light of less-informative accounting information.

One possible consequence of the decreasing relevance of accounting data is that young technology firms do not go public at all, knowing that investor skepticism likely raises the costs of capital, making other capital-raising options such as debt offerings more attractive. Another possible consequence is that disclosure of accounting information becomes less relevant for investors than open-ended textual information disclosures. Indeed, research has found that textual disclosure can impact IPO underpricing. For example, Hanley and Hoberg (2010) distinguish between textual information that is re-used across different IPOs in the same industry or year (the "standard" component of disclosed risk) and text that is unique to a particular firm (the "informative" component of disclosed risk). The authors find that informative textual disclosure is associated with lower underpricing and thus higher proceeds. Along a similar line of investigation, Agarwal et al. (2017) found that only textual information in the form of the "Risk Factors" section of the IPO prospectus filed with the US Securities and Exchange Commission (SEC) directly affects underpricing. We conceptually follow a similar approach to measure risk disclosed in the prospectus' "Risk Factors" section relative to a focal IPO firm's year and industry group.

The decreasing relevance of accounting information for evaluating technology assets raises a broader issue around whether to disclose or withhold text-based risk factors. On the one hand, the SEC legally requires IPO firms to disclose all relevant risk factors. The omission of material risk factors results in a post-IPO class-action lawsuit if the share price falls substantially lower than the offer price due to the omission. Approximately 10% of IPO firms face such class-action lawsuits up to three years after the issue date as Hanley and Hoberg (2012) report in a sample of US IPOs issued between 1997 and 2005. On the other hand, we know that issuers conduct strategic disclosure as a hedge against litigation risk – enhancing disclosure reduces the probability of a material omission (Hanley and Hoberg, 2012).

The decision to disclose or withhold risk is directly related to strategic planning. While there are

many definitions of what strategic planning entails, we follow Leiblein et al. (2018) to understand what makes a strategic decision. First, disclosure integrates with other decisions the firm makes internally regarding what assets and capabilities it develops. Second, the disclosure also affects the firm's relationship with other industry actors, e.g., its competitors and suppliers. Third, deciding what to disclose is integrated across time, meaning that what a firm discloses now can affect what the firm can, or cannot, do in the future. Considering these complexities, we view the IPO firm's decision to disclose or withhold risk as given.

Given our focus on text-based risk disclosure, it is important to describe what we mean by risk. One can unpack risk by distinguishing between idiosyncratic risk (i.e., more firm-specific risks that can be diversified away) and systematic risk (i.e., economy-wide market risks that cannot be diversified away). Examples of idiosyncratic risk include unexpected poor earnings, an employee strike, and the geographical location of production facilities, while systematic risk generally affects all firms in a market. As Lopez-Lira (2020) notes, classifying text-based risk in 10-K annual reports into either one type can be challenging. For example, many companies discussing a particular supply chain risk does not automatically imply that investors can diversify away from that risk. In practice, an economic war might trigger that supply chain risk, which can be a systematic risk. For this study, we assume that disclosed risks are generally systematic.[3]

Our first-order concern in this study is developing a measure of text-based technology risk disclosure and validating it in the context of IPO underpricing. We conceptualize text-based technology risk disclosure as the component of a firm's text-based risk disclosures that discusses technology-related risks such as risks related to updating core technologies, the risk that technology does not work reliably as the company scales, patent infringement litigation, and the risk related to licensing a third-party technology. From existing research on corporate textual risk disclosures, we know

[3]For a more formal treatment of systematic and idiosyncratic text-based risk factors, see Lopez-Lira (2020) and Hanley and Hoberg (2019).

that risk factors identified and disclosed by the IPO firm "meaningfully reflect the risks they face" (Campbell et al., 2014), on average. Therefore, we can take the disclosed risk factors at face value without a more formal empirical distinction between different types of risk as the capital asset pricing literature would suggest.

The firm's decision to disclose or withhold risk might relate to the first-day return that IPO investors receive. In the IPO literature, the term *underpricing* describes the commonly observed phenomenon that IPO shares rise during their first trading day, making the issued shares underpriced. More specifically, underpricing measures the difference between the issue price and the closing price after the first trading day (Beatty and Welch, 1996). The IPO literature has investigated how and why underpricing occurs (e.g., Lowry and Shu (2002); Bartov et al. (2002); Loughran and Ritter (2004)). For example, the asymmetric information hypothesis explains underpricing as a result of the asymmetry of information that often exists between new and existing, more informed investors or between the issuing firm and the investment bank (Rock, 1986). The implicit insurance hypothesis views underpricing as insurance against post-IPO litigation when the new issue performs below expectations (Lowry and Shu, 2002). For an overview of further theoretical explanations of underpricing see Certo et al. (2001).

One theoretical perspective views underpricing as a type of compensation or return to the initial IPO investors (Hanley, 1993; Bruton and Prasad, 1997). Under this view, the book-building process cannot resolve all risks and IPO investors bear some quantity of residual risk when the IPO firm goes public. Logue (1973) argue that investment banks set the offer prize below the expected market value to compensate IPO investors, the investment bank's clients, for taking on the residual risk. At the end of the first day on the public stock market, the assumption is that the closing price fully reflects all available information, resolving all residual risks (Fama, 1970). Given this logic, we consider underpricing as the return to the IPO investor for taking on the risk associated

with investing in a particular IPO firm. We do not examine the information disclosure during the book-building process or the channels through which the IPO firm, the underwriter, or the legal team might disclose information a priori. Furthermore, our objective is not to explain underpricing. Rather, we aim to investigate the role of technology risk disclosure at IPO and use underpricing to validate our new measure of text-based risk.

4.2.4 Validation and Hypothesis Development

With our new measure of risk disclosure, we will investigate the *return-for-risk* association of technology risk disclosure and underpricing as a validation of the new measure and test whether patents moderate the risk-return relationship.

Risk Disclosure and Underpricing

One might not find any systematic association between risk disclosure and underpricing. Issuing firms primarily consult with their underwriting investment bank(s) and legal counsel when preparing the filing documents for the SEC. After filing the S-1 document with the SEC, the IPO firm receives comments from the SEC and feedback from investors during the book-building process. The comments and feedback contribute to the updated, final prospectus called 424(b) prospectus. Throughout this process, the underwriters and the legal counsel might re-use parts of past filings for broadly applicable risks (e.g., loss of key personnel, demand uncertainty, macroeconomic effects). Hanley and Hoberg (2010) find that the IPO "Risk Factors" section contains relatively less informative content compared to the prospectus summary, use of proceeds, and MD&A sections in the same prospectus. Bao and Datta (2014) examine the "Risk Factors" section of 10-K annual reports to find that disclosed risk can decrease and increase risk perception, depending on the type of risk disclosed. In short, disclosed risk includes substantial amounts of re-used boilerplate risks, and

individual risk factors can increase or decrease perceived risk, suggesting little systematic impact to disclosing risk.

In contrast, disclosed risks might reveal unexpected or previously unknown risks, increasing investors' risk perception. Research on self-disclosed text-based risks in 10-K annual reports finds that increased risk disclosure is associated with higher proxies for risk (Kravet and Muslu, 2013; Campbell et al., 2014). For example, Campbell et al. (2014) find that longer Risk Factors sections are associated with higher stock return volatility in the following year, a proxy for market participants' perceived fundamental risk. The authors further find evidence that firms exposed to more risk disclose more risk factors and that those risk factors are not boilerplate, but that "managers provide risk factor disclosures that meaningfully reflect the risks they face" (Campbell et al., 2014). Similarly, Kravet and Muslu (2013) find that annual increases in textual risk disclosures are associated with high stock return volatility, suggesting higher perceived risk. In short, disclosing risk can be positively associated with proxies for perceived risk. Cross-sectional studies have found empirical support for the risk-return relationship in IPOs. For example, Loughran and McDonald (2013) count word frequencies of risk-related words (e.g., assume, risk, risky, and believe) in the Form S-1 document of 1,887 US IPOs as a proxy for risk. They find that word frequencies are significantly positively related to underpricing, controlling for a range of valuation-relevant variables. In other words, as investors perceive IPO firms as increasingly risky, investors demand higher returns in the form of underpricing. While studies have investigated individual risk factors (Bao and Datta, 2014; Agarwal et al., 2017), the role of text-based technology risk disclosure at IPO remains unknown.

Technology Risk Disclosure

Today, the concerns around managing and evaluating technology assets are central to how firms operate (e.g., Furr (2021); Bailey et al. (2022)). In the IPO context, Aboody and Lev (2000) find

that R&D is a major contributor to information asymmetry, suggesting that it is difficult for an outsider to learn about the productivity potential and value of a firm's R&D initiatives. Similarly, the accounting literature describes that technology assets are difficult to value due to outdated, insufficient, and inconsistent accounting standards (e.g., Lev (2000)). Loughran and McDonald (2013) report that their measure of IPO firm uncertainty and post-IPO volatility was highest during the Dot-com bubble (1999-2000), a time when many technology-intensive firms went public.

One implication of the difficulty to evaluate technology assets and the underpricing association of *aggregate* text-based risk measures is that text-based *technology* risk disclosure might matter for underpricing. For firms to capture value from their technology assets, protecting their unique intellectual asset stock is essential. For example, Cohen et al. (2000) survey managers in US-based R&D labs and found that secrecy is a central mechanism to capture value, especially from product innovations. While firms can protect their technology with, e.g., nondisclosure agreements (NDAs) during private financing rounds, they cannot use NDAs when going public. Firms preparing to go public might be incentivized to restrict technology risk disclosure for the following reasons. For example, the resource-based view suggests that disclosure might lower the barriers to imitation for competitors (Barney, 1991). Moreover, extensive disclosure can increase the costs of market participants in evaluating the firm, potentially discouraging the collection and analysis of information about the firm (Litov et al., 2012). Overall, we argue that secrecy to protect technological innovations can incentivize IPO firms to limit the amount of technology risk disclosure.

In the context of IPOs, we can distinguish between firm-level and cross-sectional arguments. We argue that firms trade off the risk of being sued post-IPO for not disclosing enough and the strategic consequences of revealing too much about their technology. If a firm underdiscloses, Lowry and Shu (2002) argue that the firm increases underpricing by lowering the issue price per share as a form of insurance against post-IPO lawsuits. If a firm overdiscloses, doing so might lower the issue price

per share and increase underpricing as a way of compensating investors for reducing competitive advantage due to revealing sensitive information. As a result, we argue that the relation between technology risk disclosure and underpricing on the firm level is U-shaped; IPO firms attempt to attain an optimal level of disclosure that maximizes proceeds and minimizes the risk of being sued (due to not disclosing enough) and the risk of revealing sensitive information (due to disclosing too much).

To investigate the cross-sectional argument, we return to core finance theory. Starting at least with the work of Markowitz (1952) on portfolio selection and Sharpe (1964) on capital asset pricing, finance scholars have argued that expected returns of an asset increase by taking on more risk; what is often referred to as the "risk-return tradeoff". In other words, investors demand compensation in the form of financial returns for investing in risky assets. In the context of text-based risk disclosures, Campbell et al. (2014) find that the disclosed risk factors meaningfully reflect the actual risks the firm faces, suggesting that risk disclosures affect the expected return of investors. Furthermore, Epstein and Schneider (2008) describe that when the information quality of an asset is challenging to evaluate, investors require additional compensation for holding the asset beyond the risk premium. Given that the economic attributes and outdated accounting measures make evaluating technology assets challenging, as discussed earlier, we reason that more technology risk disclosure suggests a greater reliance on difficult-to-evaluate assets, increasing the necessary return to investors. Taken together, we argue that the magnitude of technology risk disclosure and underpricing in the cross-section follows the typical *return-for-risk* association in financial markets. Figure 4.1 summarizes the theorized firm-level dynamics and cross-sectional association.

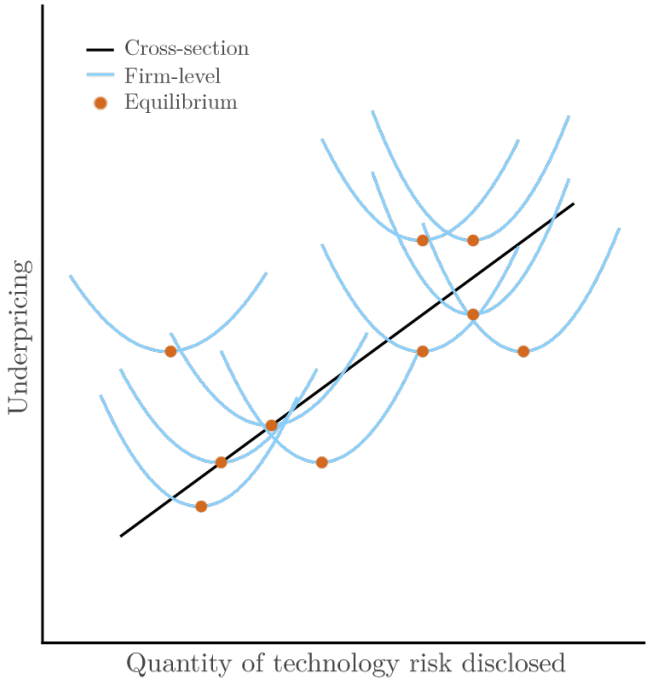

Figure 4.1: Firm-level and cross-sectional theorized associations of disclosed risk and underpricing.

To summarize the theorized association of technology risk disclosure and underpricing, our aim is to validate the *return-for-risk* association between technology risk disclosure and underpricing:

***Validation**: Technology risk disclosure is positively associated with underpricing.*

One might argue that the increasing use of machines to parse and evaluate textual risk disclosures, as evidenced by Cao et al. (2020) in the context of 10-K and 10-Q filings, might give rise to a selection effect in the words that IPO firms use in the IPO risk disclosures. The association between text-based technology risk disclosure and underpricing might be weakened because it is in the interest of the IPO firm to minimize underpricing.

Patent Moderation

Given that firms cannot use secrecy-preserving mechanisms such as NDAs when issuing shares to the public, patents might take on an important role in protecting technology information. Patents

might be a type of technology asset that is relatively easier to evaluate for potential investors than other technology assets such as proprietary databases or path software, given that Hegde et al. (2018) find that patent disclosures can reduce the perceived riskiness of a firm's R&D investments. We hypothesize that the association of technology risk disclosure and underpricing depends on whether the IPO firm owns patents[4] for the following reasons. First, firms receive the exclusive right to exploit the patented innovation in return for disclosing technology information through patents. Consequently, the incentives for IPO firms might be higher to disclose technology information through patents than technology disclosures in SEC-mandated IPO filings. Second, patents contain information about a technology, which can facilitate assessing technology assets (Heeley et al., 2007). More specifically, Hsu and Ziedonis (2013) and Morricone et al. (2017) investigate IPO firms in the semiconductor industry and find a negative association between patent stock at the time of IPO and underpricing, suggesting that patents can reduce the information asymmetry between the firm and potential investors. Third, patents can also signal the IPO firm's ability to generate technological innovations (Hsu and Ziedonis, 2013). Taken together, patents might be associated with a decrease in the perceived risk and attenuate the positive *return-for-risk* association of technology risk disclosure and underpricing. Therefore, we hypothesize:

***Main Hypothesis**: The positive association of underpricing and technology risk disclosure will be attenuated for firms with granted patents at the time of the IPO.*

Figure 4.2 below summarizes how technology risk disclosure (*TechRisk*), patents (*Patent*), and underpricing (*Underpricing*) are related. Solid arrows represent the theorized associations. Dashed arrows represent the associations of control variables and fixed effects.

[4]Given our focus on evaluating a firm's current technology assets, we follow Morricone et al. (2017) in considering granted patents rather than patent applications at the time of IPO.

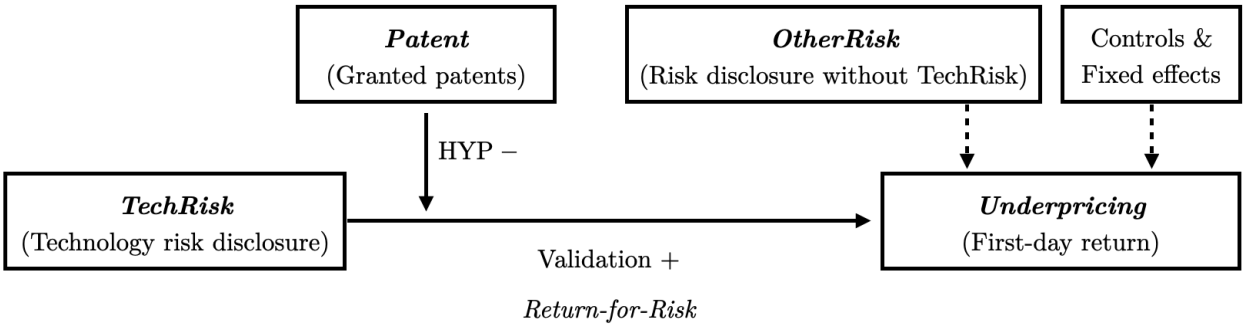

Figure 4.2: Summary of how the variables and theorized associations are related.

4.3 Research Design

4.3.1 Text-Based Measures of Risk

Our primary empirical challenge is to develop measures of text-based technology risk. In this section, we describe a series of methodological steps to compute tabulated risk topics for use as explanatory variables. Figure 4.3 summarizes the conceptual model of how a measurement model transforms text-based risk factors ($\tilde{X}$) into a technology risk disclosure measure (X) and how the technology risk disclosure measure relates to underpricing (Y) through an empirical model, estimated by using, e.g., ordinary least squares (OLS). The theories about how risk might be related to underpricing described in Section 4.2 connect $\tilde{X}$ and Y.

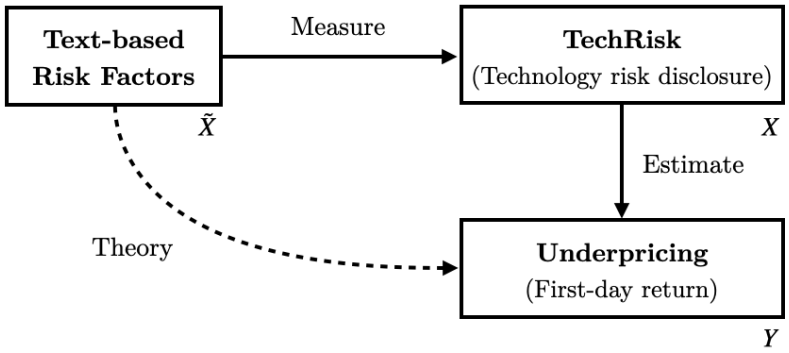

Figure 4.3: The conceptual framework.

Our measure of risk combines the idea of *counting* risk factors (Arnold et al., 2010) for computing risk magnitude and *topic modeling* of risk factor text (Bao and Datta, 2014; Lopez-Lira, 2020) for discovering risk topics semantically. In short, we first compute "aggregate risk disclosure" through z-score normalization of the output of a standard topic model and, second, extract "technology risk disclosure" from the individual risk topics in aggregate risk disclosure. Before describing the details of aggregate risk disclosure in Section 4.3.4, we briefly review how existing literature measures text-based risk, how we preprocess text, and why we use a topic model.

How can investors gauge the level of risk of an IPO firm *before* the issue? We note that one does not need to measure IPO firm risk using text but can use other proxies for risk, such as the reciprocal of the offer price (Beatty and Welch, 1996). However, text data enables a more fine-grained investigation, particularly of the meaning of risk topics. We briefly discuss two types of text-based methods previously used in the IPO literature: dictionary-based methods and unsupervised learning methods.

Dictionary Methods

A common approach to quantifying text-based information is to use a dictionary of keywords. Given a pre-defined list of K risk topics, one could define a list of keywords for each risk topic and look for keyword occurrences. For example, to measure regulatory risk exposure, one might include the keyword *law*. We would then count its occurrences across all documents. However, *law* has different meanings in different contexts. For example, *law* is a relevant keyword for describing pharmaceutical product approval, but also for tax issues and supply chain partnerships - three rather distinct contexts. Dictionary methods are not well-suited to modeling polysemy.

Li (2010) surveyed research on textual analysis of corporate risk disclosures, showing the importance of understanding the types of risks that firms disclose. Dictionary methods for automated

text extraction in the IPO context have been used to proxy text-based risk with strategic tone (Loughran and McDonald, 2013; Brau et al., 2016) or prospectus informativeness (Hanley and Hoberg, 2010), for example. Related measures extracted from prospectus texts with dictionary and counting methods include uncertainty (Beatty and Welch, 1996), ambiguity (Arnold et al., 2010; Park and Patel, 2015), and conservatism (Ferris et al., 2012), among others. As an example of measuring a particular risk topic, Hassan et al. (2019) introduce a measure of firm-level political risk. The authors set up two corpora, one concerned with political text (e.g., a political science textbook, speeches by politicians) and another one concerned with nonpolitical topics, and use these corpora as dictionaries. Finally, the measure of political risk considers the surrounding text around words, including "risk" and "uncertainty", and compares those words to the dictionaries. Two limitations of this approach include the researcher-driven selection of appropriate corpora and the limited ability to model polysemy.

Unsupervised Learning Methods

To mitigate concerns introduced by dictionary methods, we empirically compare different unsupervised machine learning methods for dimensionality reduction, including non-negative matrix factorization (NMF) (Lee and Seung, 2001), latent semantic indexing (LSI), latent Dirichlet allocation (LDA) (Blei et al., 2003), and hierarchical Dirichlet processes (HDPs) (Teh et al., 2005). Machine learning methods enable a probabilistic approach to model discrete, high-dimensional data such as text. Using variation across all available documents, we task the model with finding the optimal discrimination heuristics. We used the topic coherence measure (Röder et al., 2015) to compare the above models and find LDA to yield the highest cross-validated coherence with low variability across independent training processes relative to the other models. LDAs have been adopted in the field of strategy (e.g., Bao and Datta (2014); Agarwal et al. (2017); Choudhury et al.

(2019)), finance (e.g., Israelsen (2014); Hanley and Hoberg (2019); Lowry et al. (2020); Lopez-Lira (2020)), economics (e.g., Hansen et al. (2017)), politics (e.g., Mueller and Rauh (2018)) and the study of science (e.g., Griffiths and Steyvers (2004)).

Endogeneity, Measurement, and Identifiability

As pointed out by Healy and Palepu (2001), disclosure endogeneity might be a serious problem in empirical studies of risk disclosure. For example, firms with high levels of disclosure were found to be associated with other economic, governance, and financial variables such as earning performance (Lang and Lundholm, 1993). Similarly, firms disclosing less risk might have more underpricing as a form of insurance against litigation (Lowry and Shu, 2002), which makes the risk-litigation relation suffer from substantial endogeneity concerns. In our context, disclosed risk and underpricing are partially endogenous: the firm has considerable flexibility in deciding what to disclose, and the firm decides on the offer price, which directly affects underpricing. We mitigate potential endogeneity concerns with the following two steps, described in greater detail in Section 4.4. First, we control for the observable factors potentially affecting risk disclosure and underpricing with a range of environmental and financial variables that have been previously shown to be relevant (e.g., Heeley et al. (2007), Loughran and Ritter (2004)). Second, we also control for the industry composition and the time variation by including industry and IPO issue year fixed effects.

Another potential issue with studies using self-constructed measures of risk is researcher-induced bias and measurement error (Healy and Palepu, 2001). Unless the paper describes the entire measurement procedure in great detail, researcher-driven studies of risk factors are difficult to replicate. We, therefore, minimize researcher influence by using unsupervised machine learning methods, describing the measurement procedure in detail, and open-sourcing the code as *RiskyData-LDA* on GitHub to enable replication and extension (we describe the method in detail in Section

4.3 and Appendix E). We aim to reduce the measurement error by using unsupervised machine learning methods to capture the underlying risk signal in the text-based risk disclosures. While we cannot eliminate measurement error, we aim to minimize the potential attenuation bias due to a noisy measure of risk by increasing the number of passes through the text corpus during the training phase. Finally, we further alleviate measurement concerns by retraining alternative implementations of our measure of risk to show that our results are robust (Section 4.5.3).

One more empirical challenge when using topic models is identifiability and stability. It has been shown that topic model outputs can depend on the initialization specification (Belford et al., 2018) and the random seed (Yang et al., 2016), even when using the same input corpus. We take steps to increase the robustness of the topic model outputs by using cross-validation for finding an appropriate number of topics (Appendix E.2) and independently retraining the topic model with varying random seeds (Section 4.5.3) to show that our results are robust.

4.3.2 Text Preprocessing for Dimensionality Reduction

Text data consists of words[5], a form of high-dimensional data. Most text analysis methods represent documents as a bag of words, ignoring word order. In its most basic form, a document d can be represented as the presence or absence of a unique word v in a vocabulary containing all corpus words. Such a text representation is known as a document-term matrix of dimensionality (D, V). The corresponding document-term matrix is very large and sparse in settings with thousands of documents and a vocabulary of tens of thousands of unique words (i.e., filled mostly with zeros). At the core, text analysis methods are dimensionality reduction methods.

In the first step of reducing dimensionality, we focus on the words that contain most of the

[5]Note that the natural language processing field often uses the more general expression *term* to include words, symbols, and other non-word sequences of characters. For our purposes, the distinction between word and term is not important, so we use *word* and *term* interchangeably.

information. Risk disclosures necessarily contain words like *the, for*, and *but* to construct grammatically correct sentences. However, these so-called stop-words do not contain information that might be useful for discriminating different risk topics. Hence, we remove all stop-words using the Python Natural Language Toolkit (NLTK) package (Bird et al., 2009) list of 179 stop-words[6]. We also lowercase all terms and remove punctuation, digits, and single-character terms as these aspects are unlikely to contain relevant signals for measuring risk and further reduce dimensionality.

Next, we use point-of-speech (POS) tagging to determine the grammatical type of each term as described in Toutanova et al. (2003) and implemented by Explosion AI's spaCy package (we used the source code from `https://github.com/explosion/spaCy`). We only keep proper nouns (the PROPN tag, e.g., apple), nouns (the NOUN tag, e.g., startup), and verbs (the VERB tag, e.g., buy). We remove all terms with other POS tags.

Finally, we use the NLTK Snowball stemmer, an updated version of the well-established Porter stemmer (Porter et al., 1980), to stem all remaining terms. Stemming transforms terms into their linguistic roots, such that, for example, the words *technologies, technology* and *technological* all become *technolog*. Note that stemmed terms need not be words that appear in the English dictionary.[7]

Table 4.1 summarizes the effect of preprocessing on data dimensionality. Starting from raw text, the total number of words is nearly 35 million, and the unique number of words is nearly 60,000. Each column shows how the number of words changes for each preprocessing step as one moves right. While reductions in the total and the unique number of words are substantial, the vocabulary after preprocessing still contains 40,190 unique words and, therefore, a 40,190-dimensional computational problem.

[6]List of stop-words: `http://snowball.tartarus.org/algorithms/english/stop.txt`, accessed 22 June 2022.

[7]We have experimented with two more specific preprocessing steps used in Hansen et al. (2017): identifying collocations (i.e., N-grams) and ranking words by their term frequency-inverse document frequency (tf-idf) score. Both steps were judged not necessary given the characteristics of our text data.

	Raw text	Stop-words	POS filtering	Stemming
Total words	34,820,313	19,422,424	15,005,162	15,005,162
Unique words	57,323	57,193	53,229	40,190

Table 4.1: Data dimensionality reduction through text preprocessing.

4.3.3 Topic Modeling

Topic models are an unsupervised learning method that reduces dimensionality by discovering latent topics or themes in a set of documents. Topic models can be generally distinguished between deterministic and probabilistic models. Using keyword lists and dictionaries is the *de-facto* standard deterministic approach, where a list of keywords defines the latent theme of interest. Deterministic methods have been used in the IPO literature (e.g., Loughran and McDonald (2011, 2013); Brau et al. (2016)). These methods are easy to interpret and work well when there are few latent topics, few documents, and well-defined topics - characteristics that are not satisfied in our research setting. Probabilistic approaches, however, do not require researcher-specified word lists for each latent topic. Rather, they require selecting an appropriate model, validating hyperparameters, conducting robustness checks, and interpreting model outputs. Given the lack of an agreed-upon set of risk topics and thousands of multi-page SEC filing documents, we will use a probabilistic modeling approach to discover risk topics. In line with Hannigan et al. (2019), topic models are an increasingly popular computational tool to explore conceptual relationships from textual data to advance management scholarship.

LDA Statistical Model

The broader class of latent factor models includes negative matrix factorization (NMF), principal component analysis (PCA), and latent semantic indexing (LSI). NMF, PCA, and LSI are all key predecessors of the LDA model, which is a *probabilistic* latent factor model. In essence, an LSI

model is a mixed membership model that applies PCA with singular value decomposition (SVD) on the term-document matrix of dimensionality (V, D). NMF, PCA, and LSI are linear algebra approaches that share the main issue with dictionary methods: the inability to model synonyms and polysemy. LDA, however, is a generative model that treats each word in a document as a finite mixture over an underlying set of topics. In other words, the same word can occur in different topics. In turn, LDA treats each topic as an infinite mixture over an underlying set of topic probabilities, which measure the extent to which a given document discusses that topic. Furthermore, LDA has been used in the IPO literature (e.g., Israelsen (2014); Agarwal et al. (2017); Lowry et al. (2020)). We will use the LDA topic model and describe it more formally in Appendix E.1.

Selecting the Number of Topics

As with other unsupervised algorithms, researchers have to select the number of topics, K, a priori. There is no one correct value for K. In the literature of modeling text-based risk disclosures using topic models, Lopez-Lira (2020) and Huang and Li (2011) both use 25 topics, Lowry et al. (2020) use 8 topics, and Bao and Datta (2014), Agarwal et al. (2017), and Israelsen (2014) all use 30 topics. If one picks too many topics, topics become overly specific to particular risks disclosed, while picking too few topics will result in generic, potentially overlapping topics (Hansen et al., 2017). To avoid an arbitrary selection of the number of topics, K, we use cross-validation and topic model coherence (Röder et al., 2015) as the quantitative evaluation metric to compare different values for K, similar to Lopez-Lira (2020). We find the most coherent topic model at $K = 20$ topics as described in more detail in Appendix E.2.

4.3.4 New Measures of Risk

The following example shows why one cannot use the standard (i.e., unchanged) topic model output for comparing risk disclosures across IPO firms. Imagine a simple scenario with two documents, D_A and D_B. D_A has 15 paragraphs, five of which discuss topic T_1. D_B has three paragraphs, one of which discusses topic T_1. Which document has more exposure to risk topic T_1? Standard LDA output assigns a topic loading of approximately 0.33 to both documents because a third of each document discusses T_1. However, D_A has arguably more exposure to risk topic T_1 as five paragraphs discuss that topic. To remedy this inaccuracy, we combine topic modeling with a counting approach: we assign a dominant topic to each paragraph and count the dominant topics for each document. Following the insight that risk disclosure changes over time and depends on the industry (Kravet and Muslu, 2013), we then normalize risk exposure relative to a focal firm's year and industry group. The number of years and the number of industries are tunable hyperparameters, which we examine in Section 4.5.3. Using our measure risk in the toy scenario above, the *risk magnitude* will be larger for the firm with document D_A than for the firm with document D_B, ceteris paribus. Firm-level aggregate risk disclosure is the sum of the normalized, individual *risk topics*. We describe a step-by-step illustrative example in Appendix E.3.

Aggregate Risk Disclosure

We compute aggregate risk disclosure in the following steps. We first extract the "Risk Factors" section of each IPO firm's latest 424(b) prospectus, typically spanning many pages. We then split each extracted section into its paragraphs, assuming that, on average, a single paragraph describes a single risk topic. The assumption that one paragraph describes one topic, on average, is similar to Bao and Datta (2014), who assume that a single sentence represents a single risk topic, and Hanley and Hoberg (2019), who apply a topic model to paragraphs discussing risk factors. We

find empirical support for this assumption using the Lorenz curve in Appendix E.4. The idea of extracting the individual risk factors listed in the IPO prospectus was also used previously as a count-based proxy for overall firm risk (e.g., Beatty and Welch (1996); Leone et al. (2007)).

Next, we apply a topic model to identify the dominant topic for each paragraph, i.e., the topic with the highest probability. While this step might seem like a crude simplification, we found that it makes the resulting measure more robust to noise. Dominant topics per paragraph do not directly depend on how the researcher specifies the topic model (i.e., the number of topics, hyperparameter values). Each IPO firm is represented as a count of its dominant paragraphs – firms disclosing more risk paragraphs will, therefore, contain more dominant paragraphs and are more likely to score high on aggregate risk disclosure.[8]

We then compute the mean dominant paragraph counts for each topic k in the same year and industry group, $\mu_{k,year,ind}$. Similarly, we compute the standard deviation of dominant paragraph counts for each topic k by year and industry, $\sigma_{k,year,ind}$. Finally, we subtract the dominant paragraph mean from the dominant paragraph counts for each topic in each document and divide the result by the dominant paragraph standard deviation of the same year and industry group. Let X' represent the vector of counts of dominant topics for an IPO firm's focal prospectus document i and X the vector of individual risk disclosure topics for that focal document. Each dominant topic count $x'_k \in X'$ for topic k is normalized to $x_k \in X$ as shown in Equation 4.1 below.

$$x_k = \frac{x'_k - \mu_{k,year,ind}}{\sigma_{k,year,ind}} \tag{4.1}$$

The aggregate risk disclosure for a focal IPO firm, $AggregateRisk_i$, is the sum of all K normalized risk topics, shown in Equation 4.2 below.

[8]The correlation coefficient for aggregate risk disclosure and the number of paragraphs in the "Risk Factors" section equals 0.64.

$$AggregateRisk_i = X = \sum_{k=0}^{K} x_k \tag{4.2}$$

For simplicity, we will refer to aggregate risk disclosure as *AggregateRisk* from here onward. We continue our investigation of firm-level disclosed *AggregateRisk*, normalized with respect to the focal firm's 4-year and Fama and French (1993) 12-industry group.[9]

Technology Risk Disclosure

We now investigate the individual risk topics to find a topic that discusses technology-related risks, including patents (Heeley et al., 2007; Hsu and Ziedonis, 2013) and licensing (Morricone et al., 2017). In related work on 10-K annual reports, Bao and Datta (2014) provide mixed evidence for the effect of individual risk topics on perceived risk. The authors find that eight out of their 30 risk topics are significantly associated with post-disclosure risk perceptions of investors; three are positively associated, and five are negatively associated. We argue that firms disclosing more *TechRisk* contribute to investors' risk perception and that those firms have higher *Underpricing*, on average.

Equation 4.2 defines *AggregateRisk* as the sum of all K normalized risk topics. Each risk topic computes how much of that particular risk a firm discloses relative to its year and industry normalization group. Table 4.2 reports all risk topics, or "RT" for short, including a researcher-given title and the most likely word stems per RT. We combine four approaches to finding human-readable topic titles: the most likely words for each topic, the most likely paragraph for each topic, a two-dimensional principal component (PCA) representation of all topics, and the distribution of firm-level control variables for each topic. Given the unsupervised nature of our method and

[9]Robustness checks with different year and industry groups in Section 4.5.3 show that our main results are robust to different normalization groups and remain qualitatively stable.

the sample consisting of technology-based firms, we expect to find an RT discussing technology risk.

Topic	Topic title	Most likely word stems
0	Growth strategy	compani, busi, acquisit, oper, assur, acquir, result, effect
1	Demand	revenu, custom, sale, result, year, period, end, quarter
2	Tax	tax, invest, incom, distribut, dividend, asset, proceed, valu
3	Competition	product, develop, market, competitor, technolog, compet
4	Share price fluctuation	market, price, stock, offer, trade, factor, secur, fluctuat
5	Share sales	share, stock, offer, sale, secur, purchas, price, option
6	Top management	stockhold, director, stock, control, provis, board, vote
7	Reporting requirements	requir, control, report, account, act, compani, growth
8	Service interruption	servic, system, custom, provid, softwar, internet, inform
9	Economic conditions I	oper, affect, condit, result, risk, increas, busi, locat, impact
10	Operating results	cost, oper, loss, increas, result, expens, incur, expect, risk
11	Debt financing	capit, fund, loan, credit, interest, financ, debt, rate, abil
12	Economic conditions II	result, busi, oper, affect, condit, chang, time, effect
13	Loss of key personnel	manag, personnel, growth, employe, busi, abil, retain
14	Technology	patent, properti, licens, protect, parti, claim, technolog
15	Regulatory compliance	regul, law, state, requir, govern, subject, includ, compli
16	Healthcare regulation	insur, liabil, claim, program, coverag, provid, reimburs
17	Supply chain	product, manufactur, delay, requir, supplier, suppli, parti
18	Drug development	product, approv, candid, trial, fda, obtain, drug, requir
19	Partnerships	agreement, term, enter, termin, partner, contract

Table 4.2: Risk topics (n = 2,532 "Risk Factors" sections of the IPO prospectus).

We conduct two steps of analysis to select the appropriate RT for technology risk disclosure. First, we assume that firms with a stronger dependence on technology disclose a larger amount of technology-related risks. If this were not the case, investors would likely sue the firm for not disclosing relevant risks (Loughran and McDonald, 2011). We, therefore, correlate all RFs with the *HighTech* indicator variable in the SDC Global New Issues (GNI) database to find that RT 14 (0.1063) and RT 18 (0.127) have the highest positive correlations. Second, we examine the most likely word stems in Table 4.2 shows that RT 14 discusses patents, intellectual property, licensing, protection mechanisms, and technology. As a consequence, we select RT 14 as a proxy for normalized disclosure of *TechRisk* because it corresponds with our broad conceptualization of

technology, while RT 18 focuses on drug development risks only.

We also conduct an ex-post analysis of RT 14 by investigating the ten firms with the most extensive technology risk disclosure. These firms include three technology firms (iGo Corp, Plumtree Software Inc, Palm Inc) and seven healthcare/biotech firms (Memory Pharmaceuticals Corp, Quanterix Corp, Idenix Pharmaceuticals Inc, Roka Bioscience Inc, Selecta Biosciences Inc, Pacific Biosciences of CA Inc, Bionano Genomics Inc). For example, Quanterix Corp is a biotechnology firm that went public on 6 December 2017. Quanterix provides a technology platform for biomarker detection and testing called Simoa. The following excerpt from their IPO prospectus (424(b) filing), filed with the SEC on 7 December 2017, indicates the central role of technology in generating revenues:

> *"We are an early, commercial-stage company and have a limited commercial history. Our revenues are derived from sales of our instruments, consumables and services, which are all based on our Simoa technology, . . ."*

As an additional qualitative interpretative approach, we can plot the distribution of $K = 20$ RTs for a firm. For example, Figure 4.4 shows the risk profile for Quanterix Corp. The risk profile shows that extensive disclosure of competitor risk (RT 3) and technology risk (RT 14) drive their firm-level *AggregateRisk*. Positive (negative) deviation from the baseline means the firm discloses more (less) of a particular risk topic relative to its year and industry group.

We open-source the code that implements our new measure of risk as the *RiskyData-LDA* on GitHub. The repository also includes the textual risk disclosure data from the SEC 424(b) filings. The tool aims to facilitate replication of our results and accelerate future work on textual data of economic interest.

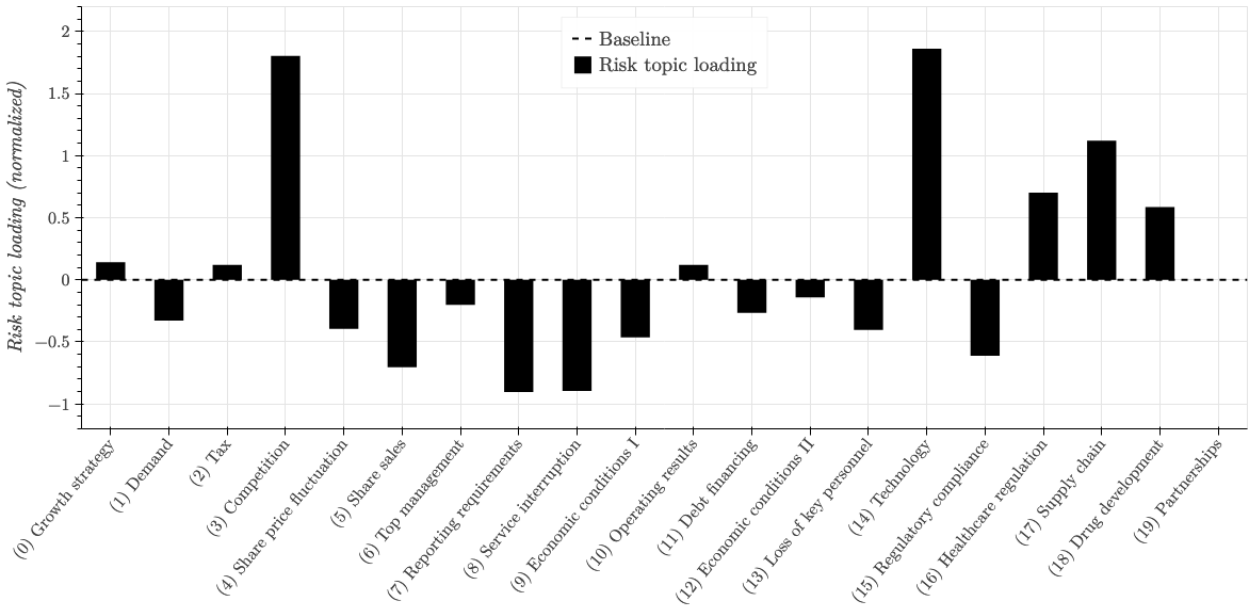

Figure 4.4: Example risk profile for Quanterix Corp.

4.4 Data and Model

4.4.1 Sources and Pipeline

We summarize the main steps of generating our cross-sectional data set in this section. We describe the data ingestion and processing pipeline in more detail in Appendix F. A summary of variables is in Table F.1 in Appendix F. We use text-based risk from SEC 424(b) filings and augment those filings with the following data sources: Center for Research in Security Prices (CRSP), Securities Data Company (SDC) Platinum VentureXpert, SDC Platinum Global New Issues (GNI), Compustat, United States Patent and Trademark Office (USPTO), Carter and Manaster (1990) underwriter reputation data updated by Loughran and Ritter (2004) and the Fama and French (1997) industry classification.

In line with previous literature (Hanley and Hoberg, 2012), our initial list of IPOs comes from the SDC GNI database. We consider issues between 1 January 1996 and 31 December 2018 on the three major US stock exchanges (NYSE, NASDAQ, and AMEX), excluding follow-on offerings,

American Depository Receipts (ADRs), and Real Estate Investment Trusts (REITs). The initial sample from SDC GNI contains 3,700 observations.

We acquired all 424(b) filing documents from the SEC Edgar database for issues from 1996 to 2018. A firm planning to go public first files an S-1 IPO prospectus. In the weeks and months preceding the offer date, the underwriter(s) and issuer promote the offering on a roadshow, meeting with potential investors. During the roadshow, the IPO firm receives feedback from investors. The S-1 is revised during this book-building process and re-submitted to the SEC as a 424(b) filing shortly before the filing date. Knowing that the risk disclosure section often changes during the book-building process (Lowry et al., 2020), we select the final 424(b) prospectus filed with the SEC and extract its "Risk Factors" section for quantifying risk. We discard the rest of the prospectus. Each IPO firm in our sample is associated with exactly one "Risk Factors" section.

The prospectus comes in either plain text (txt) or HTML format, requiring custom parsing algorithms. For a firm to remain in our sample, its file has to include a table of contents (TOC) to identify the start and end of the "Risk Factors" section (Hanley and Hoberg, 2010). We split each "Risk Factors" section into paragraphs. We use a full stop for plain text documents, followed by a new line, three empty spaces, and a capital letter as the primary paragraph divider. For HTML documents, we use the bold-italics HTML tag sequence (i.e. `<b><i>`) as the primary paragraph divider. We use a range of additional splitting rules for both plain text and HTML documents to detect paragraphs. Each firm in our SEC filing sample has to match SDC GNI on one of the following criteria: a fuzzy firm name match, the SEC filing number, or the ticker symbol and issue year. For the 3,700 observations in our initial sample, we match 3,396 machine-readable documents, equivalent to 126,343 paragraphs, to use for computing text-based aggregate risk disclosure. We fit the topic model on paragraphs because we require paragraph-level topics.

Finally, we drop all IPOs with missing values in any variable and drop three observations after

manual inspection. Due to many different data sources and the difficulty of finding reliable fields for matching, the full sample used in the regression analysis consists of 2,532 observations. A two-sample t-test indicates that firms that we dropped due to missing data were not significantly different from those in the final sample on any dependent or explanatory variable. Therefore, dropping observations does not change the underlying population for which we estimate the associations.

4.4.2 Variables

Dependent Variable

Underpricing. We compute underpricing, our outcome of interest, as the difference between the offer price of a stock and its price at the end of the first day of trading, in percent of the original offer price (e.g., Beatty and Welch (1996); Morricone et al. (2017)).

Explanatory Variable

TechRisk. We disaggregate aggregate risk disclosure (*AggregateRisk*) to extract the risk topic that describes risks concerning patents, intellectual property, licensing, and technology as described in Section 4.3.4. The estimated association of *TechRisk* tests the Validation. An unpublished histogram of *TechRisk* shows an approximately normal distribution, so we do not further transform the variable.

Moderator Variable

Patent. We queried the USPTO and matched firms to patents through disambiguated firm names to create an indicator variable that equals 1 if the firm had at least one patent granted at the time of IPO, 0 otherwise. We follow previous studies (Morricone et al., 2017) in querying patents by grant date rather than the application date. We will estimate the effect of the moderator variable

*TechRisk * Patent* to test the Main Hypothesis 2.

Control Variables

OtherRisk. We compute risk disclosure excluding technology risk disclosure by subtracting *TechRisk* from *AggregateRisk* to control for variation in risk disclosure excluding the technology risk disclosure topic. Inspecting an unpublished histogram of *OtherRisk* shows an approximately normal distribution, which suggests no need to transform the variable further.

We also require firm-level covariates to estimate the association of risk while controlling for common IPO variables affecting risk disclosure and underpricing. Financial control variables include total proceeds from SDC GNI (*Proceeds*). We also include an indicator for venture capital backing (*VC*) as in Barry et al. (1990) and Megginson and Weiss (1991), an indicator for PE backing (*PE*) as found in SDC VentureXpert, and an indicator for the dot.com boom (*Boom*) equal to 1 if the firm issued between 1 January 1997 and 1 April 2000 (Aggarwal et al., 2009), 0 otherwise, to absorb variation of IPOs during the dot.com boom. We proxy for high-tech IPO firms with the SDC GNI indicator (*HighTech*). We further compute the count of IPOs in the same 4-digit SIC code in the previous year as an indication of hot markets (*Hot*) (Ritter, 1984). Underwriter reputation is coded as the Carter and Manaster (1990) and Loughran and Ritter (2004) tombstone rank of the leading investment bank(s) (we take the maximum if more than one lead underwriter) (*Reputation*)[10], VC prominence follows Gulati and Higgins (2003), who consider all VCs with a minimum 5% stake in the IPO firm and code VC prominence as 1 if the VC was listed among the top 30 on the list of total dollar amount invested in the year prior to the firm's IPO date and 0 otherwise (*Prominence*). From Compustat, we also include total assets (*Assets*), total revenue

[10]The lead underwriters bear the responsibility for determining the appropriate marketing method and pricing the issue. The offer price range depends on the firm valuation estimated by the underwriting banks' capital markets groups, which can be more art than science. For their core role in the issuing process, lead underwriters receive the largest share of fees of all underwriters.

(*Revenue*), book value (*Book*), return on assets (*ROA*) and net income (*Ni*) to control for firm size and other financial characteristics. We extract the EGC status from the IPO prospectus text, where firms are required to declare it (*EGC*). Location indicates that the firm headquarters are in the US, controlling for spatial conditions (*US*). Firm age in years (*Age*) is computed as the difference between founding year data from Loughran and Ritter (2004) and the SDC GNI issue year. We partially control for large issues through an indicator variable equal to 1 if the firm issued on the NYSE (*NYSE*), 0 otherwise. Finally, we use industry fixed effects based on Fama and French (1997) 12 industries, similar to Hanley and Hoberg (2012). We winsorize the dependent variable and explanatory variables at the 1% level to remove outliers, similar to (Loughran and McDonald, 2011). A two-sample t-test indicates that winsorizing does not significantly change the underlying population distribution.

Other Variables

We compute post-IPO stock return volatility for robustness checks. Volatility is computed as the standard deviation of daily returns over the 15 (*Vola15*), 30 (*Vola30*), and 90 (*Vola90*) days following the issue, in line with Lowry et al. (2020). We limit the time frame to 90 days to stay well within the share lockup period of typically 180 days post IPO (Field and Hanka, 2001) to ensure clean measurement.

4.4.3 Econometric Model

For our full empirical model specification in Equation 4.3, we estimate the model via ordinary least squares (OLS) with Huber-White robust standard errors to account for potential heteroscedasticity (Huber, 1967; White, 1980). In Equation 4.3, $\boldsymbol{C_i}$ is a vector of control variables, $\boldsymbol{F_i}$ is a vector of year and Fama and French (1997) 12 industry fixed effects, and ϵ_i is the error term. The estimated

coefficient β_1 tests the Validation and β_4 tests the Main Hypothesis.

$$Underpricing_i = \beta_0 + \beta_1\, TechRisk_i + \beta_2\, OtherRisk_i + \beta_3\, Patent_i + \beta_4\, (TechRisk_i * Patent_i) + \boldsymbol{\beta_5}\, \boldsymbol{C_i} + \boldsymbol{\beta_6}\, \boldsymbol{F_i} + \epsilon_i \tag{4.3}$$

Following the advice in Egami et al. (2018), it is important to keep the steps of measuring *TechRisk* and estimating the association of *TechRisk* with *Underpricing* separate. Measuring builds on an unsupervised model (such as LDA) to extract the quantified risk topics. Using the model output for estimating associations means that we have to tune the LDA model *independently* of the estimation results. Otherwise, one might tune the model until the estimation results are desirable. Once the LDA is tuned, we can fit the LDA to all of the data without the risk of overfitting as we are interested in understanding existing relations in the data, not in making accurate out-of-sample predictions. Therefore, we do not claim that the topic model used in the estimation step is generalizable to analyze the risk disclosures of firms currently preparing for an IPO.

4.5 Results

We begin our analysis by reviewing the sample of 2,532 IPO firms. Table 4.3 shows the sample industry composition and Table 4.4 contains descriptive statistics for all variables. The sample consists mostly of firms in the business equipment (33.45%) and healthcare (22.43%) industries that are relatively young (median of 9 years), unprofitable (median net loss of USD 3.37 million), substantially underpriced (mean of 27.07%), and primarily headquartered in the US (94%). Average *Underpricing* during the dot.com boom reached levels of $60 - 70\%$, in line with Aggarwal et al. (2009). According to our *AggregateRisk* measure, no obvious high-risk trait stands out among the riskiest firms: the firms belong to a range of industries, vary in offer sizes, and issue years.

29% of firms own at least one granted patent at the time of IPO. Firms with a *Patent* disclose more *TechRisk* (mean of 0.13) compared to firms without (mean of -0.04), a difference that is statistically significant at the 0.1% significance level. We detect no pattern for the mean or the median *AggregateRisk* and *TechRisk* over time.

The bi-variate correlation coefficients in Table 4.5 are reasonably small for most variables. As a direct consequence of constructing *OtherRisk* by subtracting *TechRisk* from *AggregateRisk*, *AggregateRisk* is highly correlated with *OtherRisk* (0.98). *TechRisk* and *OtherRisk* are moderately correlated (0.21). The issue *Year* is highly negatively correlated with *Boom* (-0.66) and positively correlated with *EGC* (0.76) because all three variables are time-dependent. The mean variance inflation factors of the regression models (2.42 when estimating the model in Equation 4.3) suggest no problems of multicollinearity.

Industry	**No. of firms**
Business Equipment – Computers, Software, and Electronic Equipment	847
Healthcare, Medical Equipment, and Drugs	568
Other	365
Wholesale, Retail, and Some Services (Laundries, Repair Shops)	225
Manufacturing – Machinery, Trucks, Planes, Off Furn, Paper, Com Printing	125
Telephone and Television Transmission	104
Oil, Gas, and Coal Extraction and Products	78
Consumer NonDurables – Food, Tobacco, Textiles, Apparel, Leather, Toys	73
Finance	61
Chemicals and Allied Products	34
Consumer Durables – Cars, TV's, Furniture, Household Appliances	29
Utilities	23
n	2,532

Table 4.3: Fama and French (1997) 12 industry distribution of the sample.

Variable	Observations	Mean	SD	Min	Median	Max
Age	2,532	16.22	21.47	0.00	9.00	165.00
Assets	2,532	582.90	1,866.30	0.00	131.19	34,362.00
Book	2,532	192.21	591.27	-4,480.00	77.84	12,474.03
Boom	2,532	0.30	0.46	0.00	0.00	1.00
EGC	2,532	0.23	0.42	0.00	0.00	1.00
HighTech	2,532	0.63	0.48	0.00	1.00	1.00
Hot	2,532	9.16	17.21	0.00	2.00	90.00
Ni	2,532	-8.87	122.31	-3,445.07	-3.37	1,177.00
NYSE	2,532	0.23	0.42	0.00	0.00	1.00
Patent	2,532	0.29	0.45	0.00	0.00	1.00
PE	2,532	0.21	0.41	0.00	0.00	1.00
Proceeds	2,532	153.93	412.80	3.00	76.00	16,006.88
Prominence	2,532	0.37	0.48	0.00	0.00	1.00
Reputation	2,532	7.59	2.44	0.00	8.88	9.00
Revenue	2,532	448.83	1,642.12	-18.83	68.40	40,376.80
AggregateRisk	2,532	0.00	3.25	-6.75	-0.17	8.84
OtherRisk	2,532	0.00	3.08	-7.20	-0.18	9.66
TechRisk	2,532	0.00	0.58	-1.24	-0.11	1.85
ROA	2,532	-0.14	0.46	-12.46	-0.03	0.73
Underpricing	2,532	27.07	47.55	-21.41	12.50	256.25
US	2,532	0.94	0.24	0.00	1.00	1.00
VC	2,532	0.69	0.46	0.00	1.00	1.00
Year	2,532	2005.33	7.11	1996	2004	2018

Notes: The sample includes US IPOs between 1996 and 2018. An IPO firm *Age* equal to zero can result from e.g. a merger in the same year as the IPO.

Table 4.4: Descriptive statistics ($n = 2,532$).

	Variables	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
0	Age																							
1	Assets	0.27*																						
2	Book	0.17*	0.65*																					
3	Boom	-0.08*	-0.08*	-0.04*																				
4	EGC	-0.1*	-0.07*	-0.04	-0.37*																			
5	HighTech	-0.34*	-0.17*	-0.1*	0.04*	0.06*																		
6	Hot	-0.18*	-0.1*	-0.07*	0.14*	0.03	0.34*																	
7	Ni	0.1*	-0.01	0.09*	0.01	-0.1*	-0.12*	-0.07*																
8	NYSE	0.32*	0.3*	0.24*	-0.18*	0.02	-0.33*	-0.21*	0.11*															
9	Patent	-0.09*	-0.0	-0.01	-0.19*	0.24*	0.24*	-0.01	-0.05*	-0.08*														
10	PE	0.32*	0.25*	0.1*	-0.2*	-0.06*	-0.32*	-0.21*	0.04*	0.34*	-0.13*													
11	Proceeds	0.16*	0.55*	0.7*	-0.1*	-0.02	-0.08*	-0.07*	-0.02	0.25*	0.03	0.13*												
12	Prominence	-0.12*	-0.04*	-0.03	-0.07*	0.07*	0.26*	0.15*	-0.12*	-0.08*	0.13*	-0.07*	-0.01											
13	Reputation	0.06*	0.12*	0.1*	-0.03	-0.09*	0.0	-0.01	0.01	0.17*	-0.04	0.15*	0.12*	0.12*										
14	Revenue	0.27*	0.68*	0.48*	-0.08*	-0.09*	-0.19*	-0.11*	0.2*	0.31*	-0.02	0.23*	0.37*	-0.07*	0.11*									
15	AggregateRisk	-0.04*	-0.0	0.01	-0.04	0.03	0.01	-0.03	-0.02	-0.01	0.03	-0.05*	-0.0	0.02	0.04*	-0.02								
16	OtherRisk	-0.03	0.01	0.02	-0.04	0.02	-0.01	-0.04	-0.02	0.02	0.01	-0.02	0.01	0.0	0.05*	-0.0	0.98*							
17	TechRisk	-0.08*	-0.07*	-0.05*	-0.0	0.05*	0.11*	0.04*	-0.03	-0.13*	0.13*	-0.14*	-0.04*	0.09*	0.0	-0.09*	0.38*	0.21*						
18	ROA	0.15*	0.08*	0.08*	0.0	-0.16*	-0.19*	-0.14*	0.18*	0.17*	-0.08*	0.16*	0.07*	-0.08*	0.11*	0.09*	-0.01	0.01	-0.07*					
19	Underpricing	-0.14*	-0.06*	-0.01	0.29*	-0.08*	0.19*	0.2*	-0.05*	-0.14*	-0.02	-0.15*	-0.03	0.15*	0.1*	-0.07*	0.05*	0.03	0.12*	-0.01				
20	US	-0.02	-0.04	-0.01	0.1*	-0.08*	0.03	0.01	-0.04	-0.06*	0.03	-0.01	-0.05*	0.07*	0.01	-0.02	-0.04	-0.05*	0.04*	-0.05*	0.02			
21	VCc	-0.12*	-0.05*	-0.06*	-0.18*	0.07*	0.26*	0.13*	-0.12*	-0.09*	0.18*	0.19*	-0.03	0.51*	0.16*	-0.06*	0.01	-0.01	0.1*	-0.08*	0.11*	0.1*		
22	Year	0.03	0.13*	0.07*	-0.66*	0.76*	-0.04	-0.11*	-0.07*	0.21*	0.26*	0.16*	0.12*	0.06*	-0.03	0.08*	0.02	0.01	0.02	-0.1*	-0.18*	-0.14*	0.17*	

Note: Correlation coefficients significant at the 5% probability threshold marked with an asterix.

Table 4.5: Correlation table for all pairs of variables ($n = 2,532$).

Table 4.6 contains the main results. All models include the full set of controls described in Section 4.4.2, year fixed effects, 12-industry (Fama and French, 1993) fixed effects, and robust standard errors.[11] We use *Underpricing* as the dependent variable across all models. Model (1) tests the association of *TechRisk*. Model (2) tests the association of *OtherRisk*. Model (3) tests the association of *TechRisk* when controlling for *OtherRisk*. Model (4) tests the association of *Patent*. Model (5) and Model (6) test the association of the moderator *TechRisk * Patent* without and with controlling for *OtherRisk*, respectively. Model (6) corresponds to Equation 4.3.

4.5.1 Technology Risk Disclosure

We first examine the association of *TechRisk* and *Underpricing* (Validation) in Model (1) and Model (3) of Table 4.6. In the Validation, we posit a *return-for-risk* relation between *TechRisk* and *Underpricing*, suggesting that investors demand a higher initial return for taking on more technology risk. Table 4.6 provides evidence in support of the Validation, showing a positive and significant association of *TechRisk* and *Underpricing* when controlling for *OtherRisk* ($p < 0.000001$ in Model (3)). Model (2) tests the association of *OtherRisk* and *Underpricing* to understand how the variation in risk disclosure quantity without technology-related risk relates to *Underpricing*. Model (2) shows no evidence for a *return-for-risk* association for risks excluding technology risk. Comparing Model (1) and Model (2) also shows that *TechRisk* explains 5% more variability in *Underpricing* than *OtherRisk*. Together, Model (1), Model (2), and Model (3) suggest that technology risk disclosure is central to the market evaluation of risk. To interpret the magnitude of the coefficient in Model (3), we first note that the *TechRisk* variable is approximately normally distributed in the

[11] One might argue for additionally clustering standard errors to account for within-group correlations across observations. We follow Abadie et al. (2017), who argue that clustering in a fixed-effects model with effect heterogeneity is appropriate only in the case of clustered sampling or clustered assignment, which is not the case in our setting as both the sampling and the assignment occur at the firm level. Therefore, while our reported results do not cluster standard errors, we find that the main results are robust to clustering on 12-industry-level or year-level, both at the 5% significance level.

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	12.9655*	9.1761	12.6523*	9.5507+	13.8038*	13.5016*
	(5.747)	(5.757)	(5.753)	(5.738)	(5.755)	(5.756)
VC	4.1626+	5.0823*	4.2154+	5.0577*	4.1594+	4.2062+
	(2.190)	(2.191)	(2.190)	(2.209)	(2.198)	(2.198)
PE	-6.9318***	-8.1705***	-6.9308***	-8.3030***	-6.6579***	-6.6536***
	(1.937)	(1.914)	(1.938)	(1.966)	(1.997)	(1.998)
Boom	68.3166***	68.4335***	68.4918***	68.0442***	68.1121***	68.2781***
	(10.450)	(10.526)	(10.459)	(10.519)	(10.427)	(10.436)
Hot	0.0109	0.0164	0.0128	0.0122	0.0127	0.0146
	(0.077)	(0.078)	(0.077)	(0.078)	(0.077)	(0.077)
Reputation	1.2598***	1.2656***	1.2477***	1.2935***	1.2738***	1.2624***
	(0.308)	(0.309)	(0.307)	(0.310)	(0.308)	(0.307)
Prominence	8.1651***	8.4206***	8.1704***	8.4326***	7.9295***	7.9353***
	(2.170)	(2.181)	(2.170)	(2.182)	(2.170)	(2.170)
Assets	-0.0001	-0.0000	-0.0001	-0.0000	-0.0001	-0.0001
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Book	0.0023+	0.0022+	0.0023+	0.0023+	0.0024+	0.0023+
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
US	-2.8338	-2.0187	-2.6691	-2.3108	-2.8379	-2.6849
	(3.466)	(3.448)	(3.463)	(3.451)	(3.471)	(3.469)
Revenue	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ROA	6.0895**	5.8021**	6.0719**	5.8173**	6.3361**	6.3181**
	(1.988)	(1.959)	(2.010)	(1.912)	(2.037)	(2.058)
Ni	-0.0035	-0.0031	-0.0034	-0.0034	-0.0037	-0.0035
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
NYSE	-2.8606+	-3.8549*	-2.9261+	-3.7994*	-2.7951+	-2.8568+
	(1.556)	(1.540)	(1.555)	(1.541)	(1.553)	(1.552)
EGC	0.4844	0.9739	0.4196	1.1787	0.4034	0.3393
	(2.845)	(2.829)	(2.853)	(2.807)	(2.847)	(2.856)
Age	-0.0753**	-0.0741**	-0.0740**	-0.0769**	-0.0815**	-0.0803**
	(0.027)	(0.028)	(0.027)	(0.028)	(0.027)	(0.027)
Proceeds	-0.0020	-0.0019	-0.0019	-0.0020	-0.0021	-0.0021
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
HighTech	0.1578	2.0071	0.2944	1.8753	-0.2117	-0.0838
	(2.643)	(2.626)	(2.648)	(2.642)	(2.657)	(2.661)

Table 4.6: Main results (n = 2,532). *(Results continue on the next page.)*

	(1)	(2)	(3)	(4)	(5)	(6)
TechRisk	6.6861***		6.4173***		9.7773***	9.5096***
	(1.555)		(1.544)		(2.123)	(2.089)
OtherRisk		0.4767	0.2276			0.2150
		(0.311)	(0.310)			(0.308)
Patent				-0.1036	-0.3554	-0.3235
				(1.918)	(1.939)	(1.936)
TechRisk * Patent					-8.1664**	-8.1371**
					(3.028)	(3.023)
Observations	2,532	2,532	2,532	2,532	2,532	2,532
Year FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
R-squared	0.310	0.305	0.310	0.304	0.312	0.312

Notes: The dependent variable is *Underpricing* as a percentage of the offer price.
The *TechRisk* and *OtherRisk* variables are mean-centered and winsorized at the 1% level.
All models are estimated by OLS with robust standard errors, using the Huber-White estimator.
Industry fixed effects are Fama and French (1993) 12 industries.
Two-tailed tests: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1.

Table 4.6, continued: Main results. *(Results continued from the previous page.)*

range $[-1.24, 1.85]$. A unit increase in *TechRisk* is associated with an increase of approximately 6.42 percentage points of *Underpricing* and a percentage change of $\left(\frac{27.07+6.42}{27.07} - 1\right) \times 100 \approx 23.72\%$, on average, ceteris paribus. Given the mean *Proceeds*[12] of approximately USD 154 million, a unit increase in *TechRisk* is therefore associated with a USD 36,528,800 initial return (23.72% of USD 154 million), on average, ceteris paribus, suggesting an economically considerable association of technology risk disclosure.

[12] We do not take into account the exercise of overallotment options due to the unreliability of these data in the SDC Global New Issues database, in line with Loughran and Ritter (2002).

4.5.2 Patent Moderation

For evaluating the Main Hypothesis (HYP), we turn to hypothesis testing of our moderation condition. HYP specifies the condition that *TechRisk* might have a smaller association with *Underpricing* for IPO firms with at least one granted patent. We hypothesize that a firm's patents are associated with a decrease in the perceived risk of a firm for a given level of technology risk disclosure relative to a firm without granted patents. In other words, we expect the interaction coefficient of *TechRisk * Patent* to be negative. Before investigating the interaction effect, we note that Model (4) shows no significant association of *Patent* and *Underpricing*, in line with the results in Heeley et al. (2007).[13] The interaction term, *TechRisk * Patent*, in Table 4.6, Model (6), shows a significant negative association with *Underpricing* ($p = 0.007159$), providing evidence in support of HYP. Model (6) controls for variation in risk disclosure, excluding technology risk (*OtherRisk*) to show that the negative interaction in Model (5) negligibly decreases while remaining highly significant. The association of *TechRisk* and *Underpricing* is smaller for firms owning at least one granted *Patent*, reducing the association of *TechRisk* from 9.51 percentage points to 1.37 ($9.51 - 8.14$) percentage points, on average, ceteris paribus. Therefore, patent ownership reduces the *return-for-risk* association of *TechRisk* by 8.14 percentage points and $\left(\frac{27.07-8.14}{27.07} - 1\right) \times 100 \approx -30.07\%$ on average, ceteris paribus. Given the mean *Proceeds* of approximately USD 154 million, a unit increase in *TechRisk* is associated with a USD 54,100,200[14] initial return for firms without a *Patent* and a USD 7,792,400[15] initial return for firms with at least one granted *Patent*. It follows that *Patent* ownership reduces

[13] The effect of patent stock on Underpricing at the time of IPO can have ambiguous results, depending on the sample investigated. For example, Heeley et al. (2007) (1,413 companies) find no direct association. Hsu and Ziedonis (2013) (sample of 370 semiconductor companies) and Morricone et al. (2017) (sample of 130 semiconductor companies) find a negative association. Our sample is most similar to Heeley et al. (2007), who find no direct association between patents and underpricing.

[14] The percentage change in *Underpricing* for firms without a patent equals $\left(\frac{27.07+9.51}{27.07} - 1\right) * 100 \approx 35.13\%$ and 35.13% of USD 154 million equal USD 54,100,200.

[15] The percentage change in *Underpricing* for firms with a patent equals $\left(\frac{27.07+1.37}{27.07} - 1\right) * 100 \approx 5.06\%$ and 5.06% of USD 154 million equal USD 7,792,400.

the association of *TechRisk* and *Underpricing* by $\left(\frac{27.07-8.14}{27.07}-1\right)\times 100 \approx -30.07\%$, which amounts to a decrease of USD 46,307,800 million (30.07% of USD 154 million) in intial returns, on average, ceteris paribus.

An alternative explanation of the moderation effect acknowledges that technology risk disclosure (as opposed to patents) can be a mechanism to reduce information asymmetry, given that the firm owns at least one granted patent. The issuing firm decides how much technology-related risk to disclose when preparing for the IPO. The negative interaction effect of *TechRisk * Patent* suggests that the level of technology risk disclosure, rather than the patent stock, can reduce the information asymmetry when *Patent* equals one. As a consequence, one would expect that firms with patents and extensive technology risk disclosure are associated with lower *Underpricing*. Heeley et al. (2007) find that the effect of patent stock on underpricing is conditional on how transparent the link between patents and inventive returns is, where patent stock in a transparent industry is associated with lower underpricing. Our results are consistent with Heeley et al. (2007) in that disclosing more technology risk might increase the transparency of how the firm can use its patents to appropriate inventive returns, therefore reducing underpricing.

For the associations of control variables we interpret Model (6) because it contains all hypothesized variables. We find that firms with private equity investors (*PE*) are associated with significantly lower *Underpricing* ($p = 0.000879$). Firms issuing during the *Boom* period are associated with significantly higher *Underpricing* ($p < 0.000001$), in line with the results in Aggarwal et al. (2009). Firms working with underwriters with a high *Reputation* are significantly more underpriced ($p = 0.000041$) (Baron, 1982) as are firms high *Prominence* VC backing ($p = 0.000261$). Finally, firms with a higher *ROA* are associated with significantly higher *Underpricing* ($p = 0.002164$), and older firms (*Age*) exhibit significantly less *Underpricing* ($p = 0.003196$), consistent with the view that young (often high-tech) firms have greater growth opportunities (Younge, 2012).

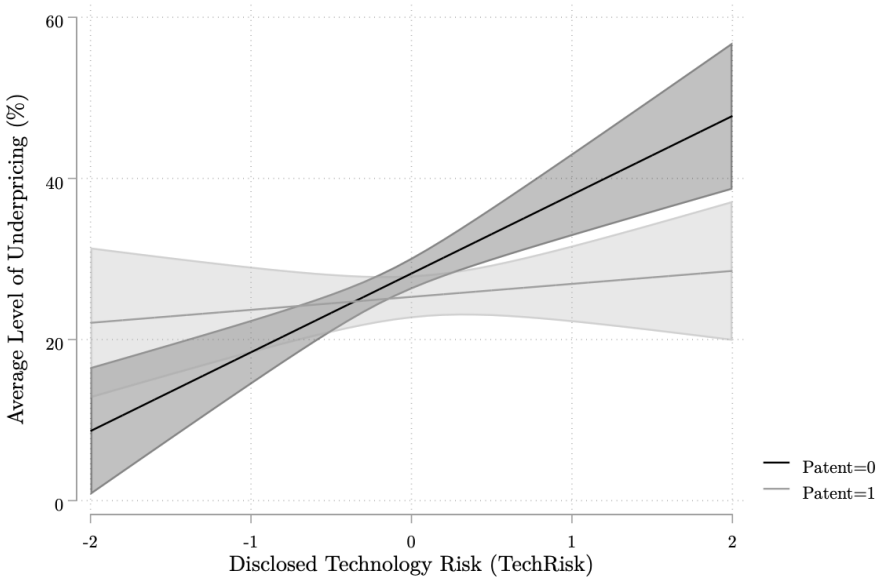

Figure 4.5: Marginal effect of *TechRisk* on *Underpricing* by *Patent* with 95% CIs.

Figure 4.5 shows the marginal effect of technology risk disclosure (*TechRisk*)[16] and *Underpricing* by *Patent* to show that owning at least one patent attenuates the positive *return-for-risk* association of *TechRisk*. The marginal effect when *TechRisk* is larger or equal to approximately 0.5 is statistically significantly higher for firms without a patent, suggesting that the *Patent* status only matters when the firm extensively discloses *TechRisk*. Secrecy considerations can explain this marginal effect as patents can mitigate the imitability concerns that might come with above-average technology risk disclosure. Consequently, below-average *TechRisk* poses little imitability concerns, reducing the importance of owning a *Patent* to deter imitation. Another explanation based on information asymmetry theory suggests that potential investors expect more information on the IPO firm's technology in the form of patents for firms with above-average *TechRisk*. However, the technology-related information asymmetry that patents could reduce is limited for firms with below-average *TechRisk*, resulting in an insignificant difference in the marginal effect by *Patent* for lower levels of *TechRisk*.

[16]The marginal effect estimations are supported by *TechRisk* in the range [-1.24, 1.85].

To summarize this section, the results show that technology-related risk disclosure is positively associated with *Underpricing* in a *return-for-risk* association, suggesting that the market expects higher initial returns for taking on more *TechRisk*. We also find that other risk disclosures are not significantly associated with *Underpricing*. Finally, owning at least one *Patent* attenuates the *return-for-risk* association.

4.5.3 Robustness Checks

Risk and Volatility

In describing the *return-for-risk* association, we argued that disclosing more *TechRisk* can increase investors' risk perception of the firm. We now conduct a brief robustness check to investigate this statement. Post-IPO stock return volatility has been used as a proxy for investors' perceived risk (Loughran and McDonald, 2011, 2013; Kravet and Muslu, 2013; Lowry et al., 2020). One perspective is that investors' estimates of firm value vary more for riskier firms than for less risky firms, which means that the spread of daily stock returns is higher for riskier firms. We therefore expect that *TechRisk* and post-IPO stock return *Volatility* are positively associated.

We set up three regression models where the dependent variable is *Volatility* and the explanatory variable is *TechRisk* with all control variables and fixed effects specified in Equation 4.3. We measure *Volatility* as three separate variables: the 15-day, 30-day, or 90-day rolling standard deviation from the first trading day forward (Kravet and Muslu, 2013). We find that *TechRisk* is positively significantly associated with *Vola15* ($p < 0.000001$), *Vola30* ($p = 0.000219$), and 90-day volatility ($p < 0.000001$), providing support for *TechRisk* as a proxy of investor risk perception at the offering.

Retraining with Varying Seeds

The output of a probabilistic topic model can change for different random seeds (Yang et al., 2016). Therefore, we retrain ten LDA models with varying, randomly drawn seeds and extract *TechRisk* from the resulting $K = 20$ topics following the steps outlined in Section 4.3.4. We use these alternative measures of *TechRisk* and *OtherRisk* to estimate the regression model in Equation 4.3. The results are robust to retraining the topic model with varying seeds. More specifically, the association of *TechRisk* remains significant at the 1% significance level, with coefficient sizes ranging from 6.4475 to 8.6686. The association of *OtherRisk* remains non-significant. The moderating effect of *TechRisk * Patent* remains significant at the 5% significance level, with coefficient sizes ranging from -6.2144 to -8.7969. As such, the main results and conclusions of our analysis remain qualitatively robust while showing marginal quantitative changes due to the probabilistic nature of the LDA model.

Year and Industry Normalization Groups

We now validate the robustness of our results with respect to the year and industry normalization groups for computing *TechRisk* and *OtherRisk*. Table 4.7 reports the robustness checks for the regression model in Equation 4.3. Models (1) through (6) report the associations when normalized with combinations of 2, 3, or 4 year groups and 5 or 12 Fama and French (1993) industries. Our baseline normalizing group is Model (1) and spans two issue years (i.e. 1996 and 1997, 1998 and 1999, 2000 and 2001, etc.) and 12 Fama and French (1993) industries (i.e. Business Equipment, Healthcare, Manufacturing, etc.). For estimating the associations of *TechRisk*, *OtherRisk*, and *TechRisk * Patent*, we fit the regression model in Equation 4.3. All models include the full set of control variables, year fixed effects, industry fixed effects, and robust standard errors. Table 4.7 shows that the results are robust to different normalizing groups. First, all specifications

for *TechRisk* exhibit positive associations with *Underpricing* at the 0.1% significance level (Validation). Second, all specifications of *OtherRisk* remain not significantly associated with *Underpricing*. Finally, the moderating effect of *TechRisk * Patent*, is negatively significantly associated with *Underpricing* at least at the 5% significance level across all specifications (HYP).

	(1)	(2)	(3)	(4)	(5)	(6)
	2 years	2 years	3 years	3 years	4 years	4 years
	12 ind.	5 ind.	12 ind.	5 ind.	12 ind.	5 ind.
TechRisk	9.5096***	9.4488***	9.2235***	9.3362***	9.2744***	9.2789***
	(2.089)	(2.038)	(2.057)	(2.011)	(2.121)	(2.084)
OtherRisk	0.2150	0.2138	0.1840	0.2422	0.1753	0.1936
	(0.308)	(0.306)	(0.297)	(0.298)	(0.304)	(0.304)
TechRisk * Patent	-8.1371**	-9.1279**	-7.7950**	-8.7848**	-7.7323*	-8.7041**
	(3.023)	(2.908)	(2.937)	(2.811)	(3.062)	(2.979)
Observations	2,532	2,532	2,532	2,532	2,532	2,532
Controls & FEs	Y	Y	Y	Y	Y	Y
R-squared	0.312	0.312	0.305	0.312	0.312	0.312

Two-tailed tests: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1.

Table 4.7: Robustness check for measuring risk disclosure with different normalization groups.

Patent

To account for variations in the timing and magnitude of patent stock of a firm, we vary the timing and the minimum number of granted patents when computing the *Patent* variable in Equation 4.3. Regarding the timing, we test alternative specifications of *Patent* to account for the possibility that potential IPO investors assess the patent portfolio 1, 2, or 3 months prior to the IPO. Table 4.8 shows that the associations of *TechRisk*, *OtherRisk*, *Patent*, and *TechRisk * Patent* with *Underpricing* are robust to alternative points in time when investors might evaluate the patent portfolio. The coefficient magnitudes for *TechRisk* decrease as the gap between the evaluation point in time and the IPO date increases from Model (1) to Model (4).

	(1) At IPO	(2) 1 month before	(3) 2 months before	(4) 3 months before
TechRisk	9.5096***	9.2646***	9.0407***	8.8544***
	(2.089)	(2.076)	(2.064)	(2.026)
OtherRisk	0.2150	0.2170	0.2181	0.2220
	(0.308)	(0.309)	(0.309)	(0.309)
Patent	-0.3235	-0.3455	-0.7536	-0.7247
	(1.936)	(1.950)	(1.916)	(1.912)
TechRisk * Patent	-8.1371**	-7.6088*	-7.0076*	-6.7692*
	(3.023)	(3.031)	(3.035)	(3.037)
Observations	2,532	2,532	2,532	2,532
Controls & FEs	Y	Y	Y	Y
R-squared	0.312	0.312	0.312	0.312

Two-tailed tests: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1.

Table 4.8: Robustness check for computing *Patent* at different points in time.

Concerning the number of granted patents, we test alternative specifications of *Patent* to counter the argument that a single granted patent might not substantially reduce the information asymmetry between the issuing firm and potential investors. We, therefore, increase the threshold such that *Patent* equals 1 only when the IPO firm has at least 2, 3, 4, or 5 granted patents at the time of the IPO, 0 otherwise. Table 4.9 shows that the associations of *TechRisk*, *OtherRisk*, *Patent*, and *TechRisk * Patent* with *Underpricing* are robust. The coefficient magnitudes for *TechRisk* decrease as the number of granted patents increases from Model (1) to Model (5).

To conclude this section, we find evidence that text-based risk disclosure matters for the underpricing of IPOs. Our results provide evidence for a *return-for-risk* association for risk disclosure, driven by the disclosure of technology-related risks. We find no evidence for an association of risk disclosure when excluding the disclosure about technological risks. Finally, we find evidence that patents can considerably attenuate the *return-for-risk* relationship between technology risk disclosure and underpricing.

	(1) At least 1	(2) At least 2	(3) At least 3	(4) At least 4	(5) At least 5
TechRisk	9.5096***	8.7226***	8.2453***	7.9517***	7.7822***
	(2.089)	(1.945)	(1.847)	(1.783)	(1.758)
OtherRisk	0.2150	0.2073	0.2071	0.2010	0.2015
	(0.308)	(0.309)	(0.309)	(0.309)	(0.309)
Patent	-0.3235	-0.1601	1.4391	2.2777	3.6484
	(1.936)	(2.039)	(2.234)	(2.203)	(2.321)
TechRisk * Patent	-8.1371**	-7.9989**	-7.8929*	-8.0298*	-7.8768*
	(3.023)	(3.041)	(3.256)	(3.300)	(3.451)
Observations	2,532	2,532	2,532	2,532	2,532
Controls & FEs	Y	Y	Y	Y	Y
R-squared	0.312	0.312	0.311	0.311	0.312

Two-tailed tests: *** p<0.001, ** p<0.01, * p<0.05, + p<0.1.

Table 4.9: Robustness check for computing *Patent* with different minimum counts.

4.6 Conclusion

Drawing on a cross-sectional dataset of IPO firms, we find evidence that information disclosure for technology firms matters at the IPO. Using a novel approach to measuring text-based risk, we validate the *return-for-risk* association for technology-related risk disclosures. Moreover, we find that patents attenuate this *return-for-risk* association.

Our paper makes several contributions. First, we combine research on the economics of technology assets (e.g., Corrado and Hulten (2010); Brynjolfsson et al. (2021)) with the IPO literature (e.g., Hanley and Hoberg (2012); Loughran and McDonald (2013); Morricone et al. (2017)) to contribute to the broader discussion around managing a firm's technology capabilities (Furr, 2021). Using a new approach to measuring text-based risk disclosures, our analysis suggests that the textual risk disclosures around technology impact firm value. We are the first to explicitly investigate technology risk disclosures in the IPO context to the best of our knowledge. Second, we contribute to examining the role of text-based risk disclosure at IPO. We add to this field with dispersed and

mixed evidence (Ritter and Welch, 2002) by connecting research on risk disclosure at IPO (e.g., Hanley and Hoberg (2010); Loughran and McDonald (2013)) and probabilistic text analysis with topic models (Egami et al., 2018; Hannigan et al., 2019) to focus on a particular risk topic, namely technology risk disclosure. By investigating the role of text-based risk at IPO, we further support the evidence that text-based risk disclosure contains signals that affect IPO outcomes. Third, we develop, apply, and open-source a new approach to measuring text-based risk disclosure that considers the time and industry dimensions of IPO firms, two key components in light of the dynamic nature of IPO phenomena (Ritter and Welch, 2002). The code is available as the *RiskyData-LDA* platform on GitHub.

The results in this study also have potentially important managerial implications. For example, technology-intensive firms without patents might face greater difficulties in raising financial resources when they "leave money on the table" (Ritter et al., 1998) to compensate IPO investors through underpricing or increasing the risk of a class-action lawsuit post IPO by withholding relevant risk factors in the IPO prospectus. Therefore, our results can have several consequences for the pricing of shares, patenting activity, and information disclosure when preparing for the IPO. In short, our findings suggest that IPO firms should limit technology risk disclosure and coordinate the IPO timing with their intellectual property activity.

Our findings come with limitations that open opportunities for future research. First, our static, cross-sectional research design compares firms at similar stages of their life-cycle on the one hand, but on the other hand, faces typical limitations of cross-sectional studies (Angrist and Pischke, 2010). As such, our results are mostly correlational and indicative. Hence, we take advice from Leamer (1983) to conduct robustness checks, investigating results under different variable constructions and model specifications. A possible next step to overcome some of the above limitations follows Sutton and Staw (1995), who point out that "one indication that a strong theory

has been proposed is that it is possible to discern conditions in which the major proposition or hypothesis is most and least likely to hold." We could investigate the drivers of the *return-for-risk* relation in more detail by looking at additional conditions under which the association strengthens or weakens. Second, our measure of text-based risk is one of many possible ways of quantifying the risk magnitude and risk topics disclosed by a firm (Bao and Datta, 2014; Hanley and Hoberg, 2019; Lopez-Lira, 2020). Despite a set of robustness checks, unsupervised learning methods lack a "true" model of the world, by definition, and can be difficult to apply to hypothesis testing (Egami et al., 2018). Future research might increase the robustness of risk topics by focusing on identifiability in topic models (e.g., Huang et al. (2016)) through extending the *RiskyData-LDA* platform on GitHub.

Chapter 5

Conclusion

In this dissertation, I set out to explore the central role of new technologies in how firms operate to create and capture value. Across the three essays, I have investigated automation, AI, and technology risk disclosure. The theory and evidence I developed underline that new technologies can affect organizing processes involving learning and adaptation, coordination under uncertainty, and risk disclosure at IPO in strategically important ways. More specifically, the results point to potential downsides of new technologies and suggest three mitigating solutions: dynamic routines to reduce the opportunity costs of automation through combining learning and automation (first essay), *Data Clans* to mitigate intraorganizational coordination costs when building AI capabilities (second essay), and intellectual property to enable the disclosure of technology information without threatening competitive advantage at the IPO (third essay).

The three topics selected are three of many aspects of technology strategy. Other relevant topics include the role of ecosystems and technological evolution in how organizations use new technologies, for example. However, the dissertation emphasized a connecting theme around the role of coordination and communication in an intangible, rapidly-changing business environment. The first essay argues that while replacing human labor with automation routines can decrease intraorganizational coordination costs, changing environments necessitate human learning to generate new knowledge that enables adaptation. The second essay argues that a *Data Clan* can reduce intraorganizational coordination costs between AI assets in an unpredictable environment through, e.g., a shared understanding and an increased tolerance for internal contract misspecification. Finally, the

results in the third essay suggest that communication between IPO firms and potential investors through technology risk disclosures and patents matters for efficiently raising capital through an IPO. These insights point to situations where coordination can inhibit or support creating and capturing value in a dynamic environment shaped by new technologies.

The dissertation contributes to the work on integrating uncertainty into strategic management research. For example, Furr and Eisenhardt (2021) argue that environmental uncertainty about the value of firm resources requires a focus on learning and cognition to understand the strategic logic of creating and capturing value. My work contributes to the field by investigating three topics on managing new technologies under uncertainty. The results in the first essay show that the value of automation strategies depends on the predictability of the environment and, consequently, the importance of learning and cognition. The inductive insights in the second essay suggest that a strong organizational culture can facilitate coordinating interdependent assets in the rapidly evolving context of enterprise AI technologies. Finally, the results in the third essay propose that investors perceive IPO firms disclosing more technology risk as riskier with important consequences for firms raising capital. Taken together, the dissertation aims to contribute to the strategic management literature by examining managerial complications that arise from exploiting new technologies under uncertainty.

The two methodological innovations in this dissertation might further impact the work of management scholars. First, the *OrgSim-RL* platform supports investigating firm-level interactions between the division of labor, coordination costs, and automation strategies in dynamic environments; interactions that are often non-linear and challenging to examine empirically. Second, the *RiskyData-LDA* platform allows measuring the risk topics and magnitudes for texts with paragraphs of risk factors beyond the IPO context, including annual and quarterly corporate SEC filings, for example. As such, the new measure of risk opens up avenues for future work on the strategic

consequences of disclosing particular risk topics. In short, the two open-source platforms encourage the adoption of computational methods to study new research questions in the field.

The insights developed in the dissertation can be beneficial to managers to make more informed decisions and establish realistic expectations about the value of new technologies. Managers can make more informed decisions about which organizational processes to automate, what assets to accumulate for building AI capabilities in incumbent organizations, and how to disclose technology risks and organize patenting activities at the IPO. Understanding the challenges of leveraging new technologies and potential remedies can help practitioners make better decisions about effectively allocating capital within the organization. Furthermore, managers can establish realistic expectations for new technology projects and avoid disappointing project outcomes due to high stakeholder expectations. Realistic expectations can further help to critically judge the value propositions of external technology vendors and "see through all the hype" (Von Krogh, 2018) that is often associated with automation and AI technologies.

The findings in this dissertation open up at least three avenues for future research. First, the future of work might change with the burgeoning adoption of automation in organizations (Frey and Osborne, 2017; Acemoglu and Restrepo, 2018b; Dogan and Yildirim, 2021). Future work might investigate questions such as the following ones. How might automation change the relative importance of skills and tasks necessary for a job in different industries? How might novel machine learning applications affect the conditions under which automation can be beneficial? The conceptualization of automation in the *OrgSim-RL* platform can serve as the starting point for investigating the above questions. Second, it remains an open question as to how the boundaries of the firm might change as new technologies such as artificial intelligence (AI) become increasingly central to organizational processes (Varian, 2018). Under what conditions are firms more or less likely to develop an AI solution internally versus acquiring it on the market? How accurately does

transaction cost theory describe empirical evidence on make-or-buy decisions? The results in the second essay (Chapter Three) suggest that organizational culture might matter for make-or-buy decisions in the context of AI solutions. Third, investors might increasingly focus on text-based technology risk disclosure when evaluating firms as new technologies permeate businesses (Bailey et al., 2022) and remain challenging to measure with existing accounting standards (Yang and Brynjolfsson, 2001). How might technology risk disclosure in annual reports impact firm value? What conditions amplify or attenuate the effects of technology risk disclosure? The *RiskyData-LDA* platform can serve as a basis to quantify text-based risk disclosures and investigate the above questions.

Appendix A

Diagnostics and Validation

A.1 Learning Trajectories

In this section, we report the learning trajectory results for all combinations of *DOL* (λ) and *Environmental Change* (δ). Figure A.1 shows the learning trajectories of episodic net rewards by *Environmental Change* (δ) and levels of automation (τ). Panel (A) shows a stable environment, where higher levels of automation are preferable because there is nothing to learn (except for the short, initial learning effort). By the end of a simulation run, high automation approaches a net reward of three, which is the highest possible net reward as noted in Section 2.4.7. Panel (B) plots the case of intermediate *Environmental Change* that is described in detail in Section 2.4.7. Panel (C) displays a very unstable environment, where higher levels of automation are preferable because there are no benefits to learning and adaptation. The convergence around a net reward of -2 suggests the prominence of the long path under the automation mode (a net reward of -1) and adaptation mode (a net reward of -2).

A.2 Opportunity Costs

Figure A.2 shows the opportunity costs per episode for the dynamic learning process in Panel (A) and the cumulative opportunity costs in Panel (B). For high automation, opportunity costs soar after the first change in the environment and remain elevated until the end of the simulation. We observe this behavior because the opportunity costs of lost learning are highest when attempting a

closed door as there are always two open doors in the grid world.

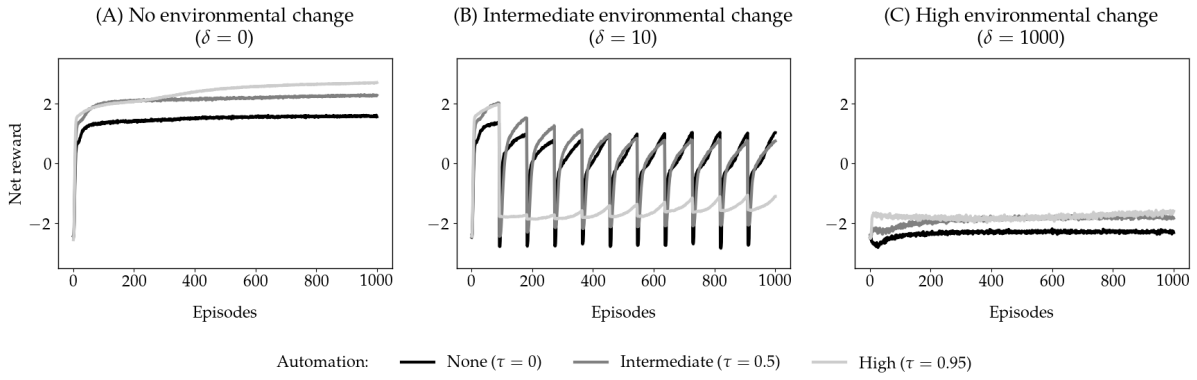

Figure A.1: Learning trajectories of episodic net rewards.

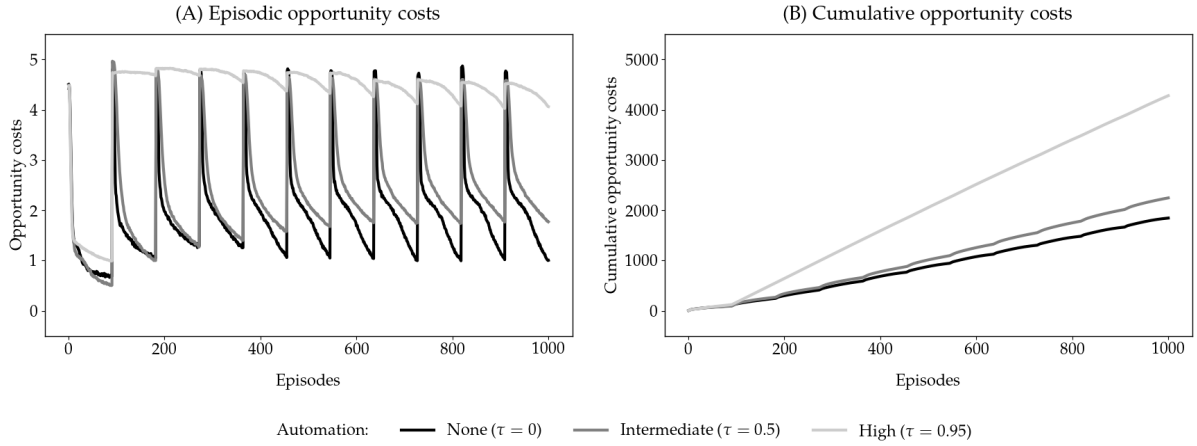

Figure A.2: Opportunity costs with intermediate *DOL* and intermediate *Environmental Change*.

A.3 Smooth Environmental Change

In this section, we change the operationalization of *Environmental Change* from punctuated changes to smooth changes in the aggregate. As a result, we expect to no longer see the punctuated learning trajectories but rather smooth learning trajectories that converge to a particular value. The results are robust to this alternative operationalization of *Environmental Change*.

Figure A.3 shows myopic automation under smooth *Environmental Change*. In Panel (A), one

can see that the learning trajectories converge to some net reward toward the end of the simulation. As expected, high automation initially performs well, but its performance rapidly deteriorates as doors start to open and close. Initially, intermediate automation outperforms no automation but converges to the same episodic net reward approximately halfway through the simulation. Panel (B) plots the cumulative episodic net rewards from Panel (A). In general, automation initially performs well, but the net rewards start to decline as environmental changes become more likely.

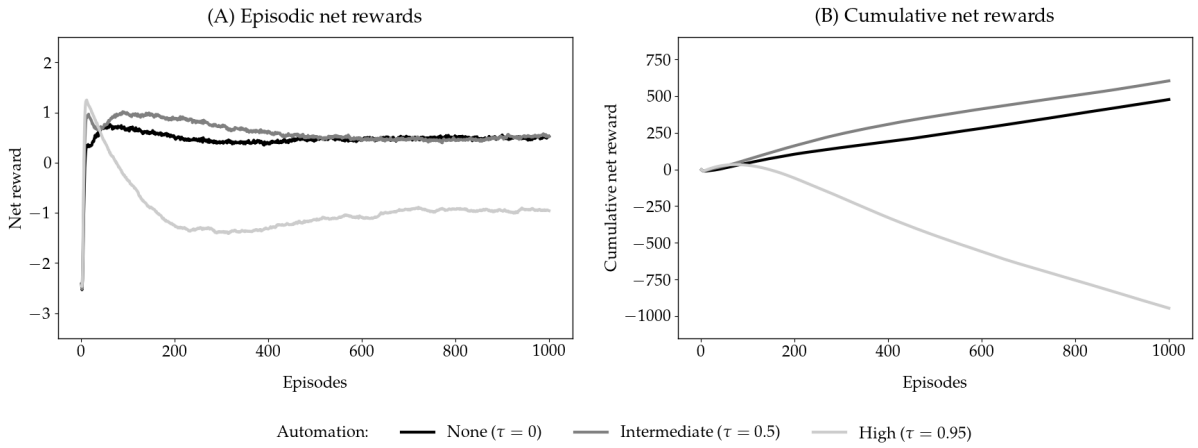

Figure A.3: Net rewards at an intermediate *DOL* and intermediate smooth *Environmental Change*.

Figure A.4 validates that automation incurs opportunity costs due to lost learning under smooth *Environmental Change*, in line with the opportunity costs under punctuated *Environmental Change* shown in Appendix A.2. Panel (A) shows that the average opportunity costs per episode equal approximately four for high automation, while for intermediate and no automation, they equal slightly less than two. Panel (B) plots the cumulative episodic opportunity costs from Panel (A).

Figure A.5 reveals that the high opportunity costs for high automation are due to the inability of automation to adapt away from the short path with a closed door. Dark grey represents the share of runs in which the agent attempted a closed door in a particular episode. Medium grey indicates the share of the long path and light grey the short path with an open door. Visually, the small "bump" in the dark area in Panel (A) at the beginning of the simulation grows for intermediate and high

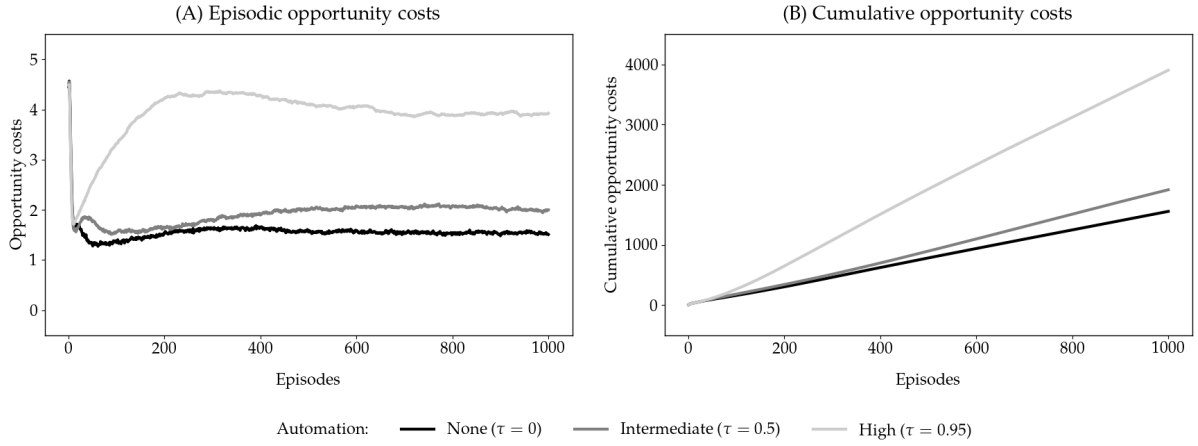

Figure A.4: Opportunity costs at an intermediate *DOL* and intermediate smooth *Environmental Change*.

automation, meaning that attempts at the closed door become more common with automation. The more elongated the bump, the slower the adaptation to a closing door. Furthermore, in Panel (C), one can see that high automation adapts away from attempting a closed door to the risk-free long path. When automation is lower in Panel (A) and Panel (B), the share of the short path with an open door is substantially higher by the end of the simulation than under high automation in Panel (C).

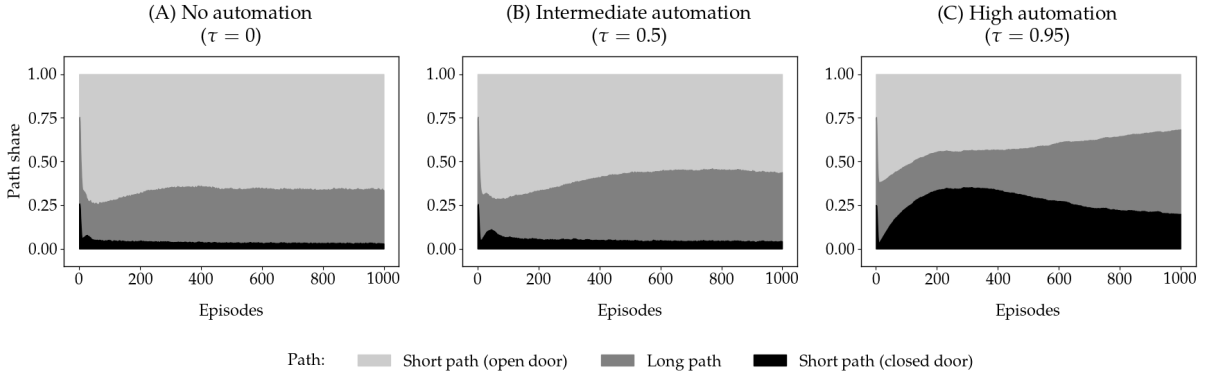

Figure A.5: Path shares at an intermediate *DOL* and intermediate smooth *Environmental Change*.

Appendix B

OrgSim-RL Platform

The *OrgSim-RL* platform is an open-source tool to simulate the returns to the *DOL* for an organization, available at `https://github.com/mxhofer/OrgSim-RL`. The tool aims to enable and accelerate future research. At the core, the simulation uses the tabular Dyna-Q (Sutton and Barto, 2018, p. 164) learning mechanism with additional economic and organizational parameters.

B.1 Architecture and Usage

While the entire code base runs on a local machine, we recommend using a cloud environment to cope with the considerable computational load. We architected *OrgSim-RL* to run in a Docker container on the Google Compute Engine (GCE). The Docker container facilitates running the simulation in other cloud environments such as Amazon Web Services (AWS) or Microsoft Azure. The simulation stores output in Google Cloud Storage, which we then ingest into Google BigQuery to efficiently query the output results. Finally, researchers can investigate results with a Streamlit dashboard, hosted and deployed with Google's Cloud Deployment Manager.

The codebase is written in Python with auxiliary YAML files and Unix shell scripts. Therefore, running the simulation requires familiarity with Python and the Unix command-line interface (CLI). For a detailed guide on using *OrgSim-RL*, consult the "readme.md" file in the root of the GitHub repository.

B.2 Diagnostic Capabilities

We implemented a range of diagnostic capabilities to monitor a simulation run and diagnose results ex-post. There are three verbosity levels for logging for monitoring a running simulation, including run-level, episode-level, and step-level logs. In the testing phase, we recommend granular step-level or episode-level logging. We recommend run-level logging to limit the console output size when running a full-scale simulation run. We also implemented an interactive grid world to visually step through the grid world with keyboard input. Figure B.1 shows an example screenshot of the diagnostic grid-world when the agent is at the start state (in red) in episode 128, which we selected to show the learned Q-values after the first door change. The values represent the maximum Q-value of the set of valid actions, and the arrows represent the action associated with the maximum Q-value. The hashes represent wall states, the states filled with green "X"'s represent door states, and the states with a blue "G" represent a goal state.

For diagnosing the simulation ex-post, the simulation outputs the net reward, coordination costs, opportunity costs, and all parameter values for each episode. In addition, we keep track of episodic diagnostic metrics, including the number of steps, the actions taken at the start state, the path taken, and the proportion of optimal actions relative to all actions taken, among others.

B.3 Future Work

We see various avenues for future research that the *OrgSim-RL* platform can support. What follows is a list of example avenues to explore. First, the nature of work is changing rapidly with the advent of automation and machine learning (Frey and Osborne, 2017; Acemoglu and Restrepo, 2018a; Raj and Seamans, 2019). How might automation change the relative importance of skills and tasks necessary for a job? How might automation affect the way that individuals coordinate with each

other? When and under what circumstances might individuals need to take over automated tasks? Researchers might investigate how varying transition costs of entering and existing automation can impact when and under what circumstances automation is desirable relative to human learning. The parameter ψ models transition costs in the *OrgSim-RL* platform.

Second, organizations might learn when to automate and when not to. How can organizations gauge when to automate and what knowledge to automate? How much better can organizations perform with such "smart automation"? To what extent can competitors imitate a dynamic capability to automate? Researchers could change the operationalization of automation in *OrgSim-RL* from an exogenous probability to automate to an endogenous mechanism that learns when to automate.

Third, an organization's current asset base might be more or less complementary to set up and maintain automation (Teece, 1986). How might different existing organizational assets change the effectiveness of automation? What assets make automation cheaper or more expensive to set up and maintain? What might be the barriers to imitating an automation routine for firms with the right complementary assets? To model variations in asset complementarity, one could change the parameter ϕ in the *OrgSim-RL* platform to control the carrying costs of automation.

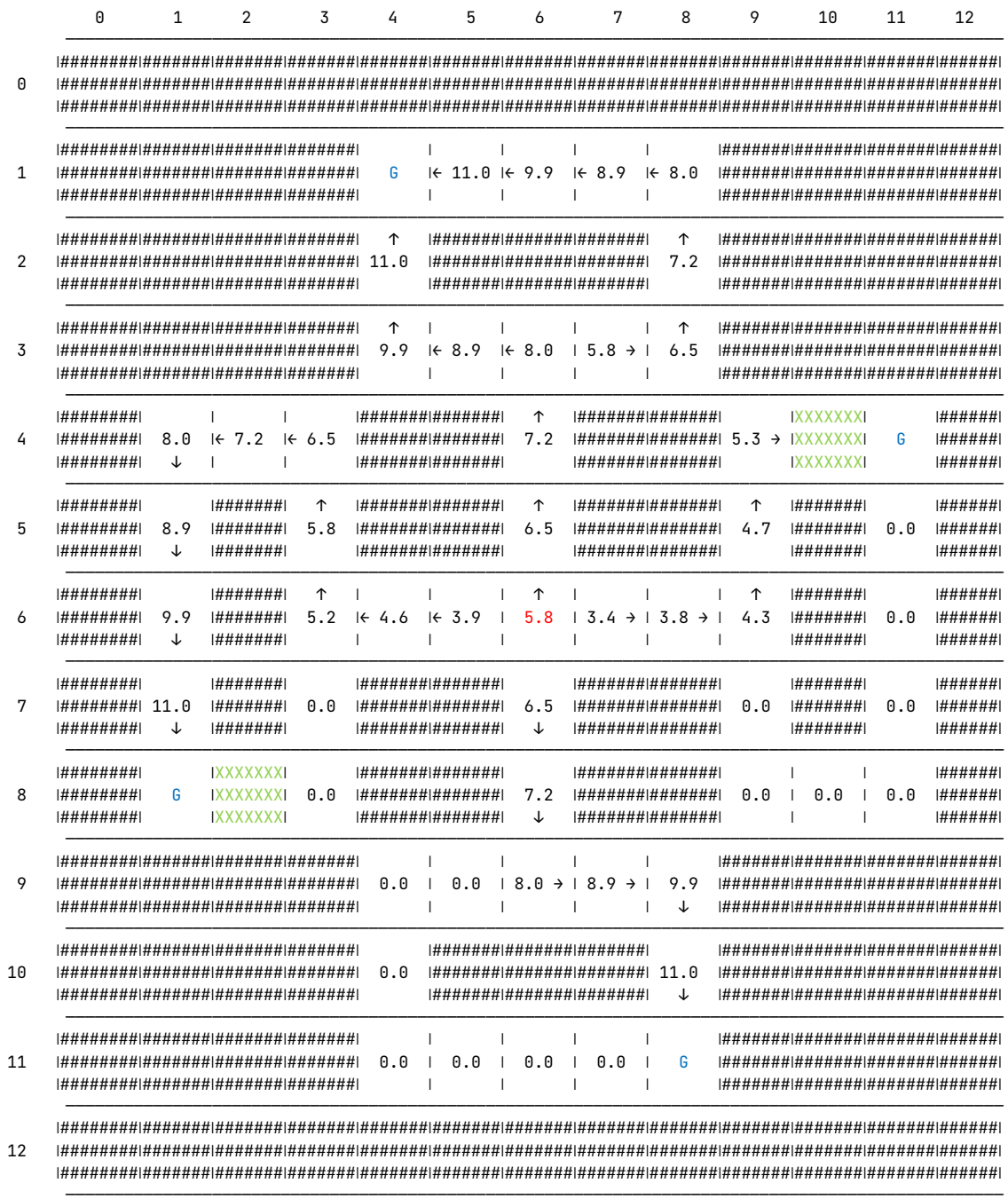

Figure B.1: Grid world output. Researchers can interactively control the simulation step-by-step.

Appendix C

Interview Questions

Characteristics of the interview participant

- What's your role at Firm X? Seniority level? Tenure? Business unit?
- Could you briefly tell me about your background (education, work experience, etc.)?
- Do you have the authority to allocate money for your department/team?
- What are artificial intelligence (AI) / big data (BD) for you?

The role of AI in the organization

- How important are AI/BD to your organisation? How does it compare to 2 years ago?
- How important are AI/BD to your business unit? How does it compare to 2 years ago?
- Does Firm X have an explicit AI/BD Strategy (or similar)? If so, can you tell me more?

Understanding resource allocations and decision making

- What are some of the recent AI/BD projects at Firm X?
- Can you tell me about the decision making process around allocating resources to AI/BD?
- What and who typically starts the allocation process?
- Which aspects of the allocation process tend to go well and which ones are difficult?
- Who's involved in making AI/BD allocation decision? Who makes the final decision? Who's accountable?
- If you could design the ideal AI/BD resource allocation decision making process, how would it look like?

Details on resource allocation

- What makes an AI/BD resource allocation successful? Financial component? Measurement(s)/metric(s)?
- How do you prioritize competing allocation opportunities?
- How do you decide which aspects to develop in-house and which ones to purchase?

Closing

- Do you think that there are any other relevant aspects that we haven't covered yet?
- What do you think are the most important topics in this discussion?

Appendix D

Labeling Topics

Table D.1 shows the hSBM topics and subtopics with their most likely word stems as visualized in Figure 3.2. We used these word stems, the link between topics and subtopics, and the most salient documents per topic to devise human-readable labels.

Topic		Most likely word stems	Subtopic		Most likely word stems
T1	Computation	tool intellig fund artifici risk relat qualiti equiti platform algorithm	**T 1.1**	Asset management	fund relat equiti privat asset public method document hedg sentiment
			T 1.2	Data	qualiti platform oh languag contamin amazon instrument shift textil toward
			T 1.3	Algorithms	tool intellig artifici algorithm forecast ad mind oppos promot academ
			T 1.4	Quantitative finance	risk perform test statist return care client account add robust
T2	Project management	use project case team build end valu cost three basic	**T 2.1**	Product design	user design exist devic extrem game instanc extract recent parti
			T 2.2	Convincing management	team person someon dedic concept c pipelin join entir blah
			T 2.3	Project sponsoring	initi obvious unit innov board present head respons bigger close
			T 2.4	Data scientists	three ago month potenti four p scientist role explor six
			T 2.5	Implementation	project basic implement defin
			T 2.6	Leadership	engin group experi cto vision skill offic hardwar cultur leadership
			T 2.7	Analytics strategy	analyt strategi whatev support top meet autom went educ singl
			T 2.8	Use case visibility	use case visibl
			T 2.9	SAFe agile framework	end valu cost discuss high stori task owner given epic
			T 2.10	Building applications	build question generat softwar requir general made type answer convinc
T3	Business impact	big tri product bit idea import technolog custom market trade	**T 3.1**	RPA	train rpa report valid code scale manual leader confer collabor
			T 3.2	Product	product technolog system r describ
			T 3.3	Trading	market trade trader hub short gas ineffici reduc layer fail
			T 3.4	Supply chain	chain suppli across insight phase grow assess analyst program outsid
			T 3.5	Business result	import result order improv expect l pretti faster life turbin
			T 3.6	Data culture	big got enabl water readi sudden plant allow afterward clean
			T 3.7	Prototyping	tri bit littl money sort effort plant seem
			T 3.8	Measuring impact	put ask success next tell number measur five sourc world
			T 3.9	Customer support	custom scienc sale quarter key integr interact critic absolut center
			T 3.10	Communicate value	show creat run cloud complex share almost et cetera accept
			T 3.11	Product	applic comput consum featur turn imag therefor sensor necessarili environ
			T 3.12	Prioritization	guy everyth piec everybodi huge listen adopt element criteria explain
			T 3.13	Supply chain	whether particular price factor vessel identifi china volum follow govern
			T 3.14	Planning	idea budget sometim alloc direct committe plan execut ceo formal
T4	Business needs	data say also need get differ kind start manag right	**T 4.1**	Machine learning	learn machin ai abl complet amount predict driven may access
			T 4.2	Data	data infrastructur collect
				Stop words	say let okay anoth usual
				Stop words	differ level give done cannot everi whole set capabl depend
			T 4.3	Cooperation	develop side might problem certain help solut togeth research topic
			T 4.4	Model compliance	get exampl look model opportun futur clear buy impact sell
				Stop words	kind right us understand term base inform better find stuff
			T 4.5	Network optimization	also take part mani cours optim less best appli network
			T 4.6	Business needs	need start manag first call said alreadi bring specif within
T5	Governance	one think peopl go realli busi like would know compani	**T 5.1**	Change management	compani lead transform never technic partner fact revenu robot rule
			T 5.2	Decision making	make mayb decis even good sure alway still happen sens
			T 5.3	Business processes	go busi thing way process come want chang talk yes
			T 5.4	Human resources	one like know year work actual two new move hire
			T 5.5	Motivation	think peopl realli mean someth time see lot yeah invest
			T 5.6	Uncertainty	would well much could resourc point around organ area probabl

Table D.1: Most likely word stems for all hSBM topics and subtopics in Figure 3.2.

Appendix E

Measuring Risk Disclosure

E.1 LDA Topic Model

LDA (Blei et al., 2003) is a fully probabilistic Bayesian factor model for discrete data. Let D represent all available documents that make up a corpus consisting of V unique words. As with other unsupervised algorithms, one must select the number of topics, K, a priori. There is no one correct value for K. Each topic $k \in K$ is a vector of probabilities $\boldsymbol{\beta}_k \in \Delta^{V-1}$ over the V unique words in the vocabulary. Intuitively, one can see a single topic k as a weighted word list where words belonging to the same latent theme have similar weights. We use $\boldsymbol{\beta}_k$ in Table 4.2 to give each topic an interpretable title based on the most likely words given that topic.

Each document $d \in D$ is a vector of probabilities $\boldsymbol{\theta_d}$ across all K topics, summing to one. Intuitively, each document can belong to multiple topics. For each document d and topic k, we get a value of θ_d^k, representing the "topic loading" of topic k in document d. The magnitude of a focal topic loading is *relative* to all other topics disclosed in the document. We use *relative* risk exposure, $\boldsymbol{\theta}$, in Section 4.3 to compute the aggregate risk disclosure for each IPO firm. Given the document-term matrix, we then try to optimize the overall likelihood given by Equation E.1, where $p_{d,v} = \Sigma_k \beta_k^v \theta_d^k$ represents the probability that a given word in document d is equal to term v in the vocabulary and $n_{d,v}$ counts the number of times word v appears in document d.

$$\prod_d \prod_v p_{d,v}^{n_{d,v}} \tag{E.1}$$

The subsequent inference problem requires approximating the posterior distributions of $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ for every topic k and every document d given K with the hyperparameters of the prior distributions.

E.2 Quantitative Model Evaluation

We implement the topic model in Python using the "gensim" package (Řehůřek and Sojka, 2010) and 3,396 firm-level documents consisting of 126,343 paragraphs (an average of 37 paragraphs per document). We use paragraphs as inputs for all topic models, meaning that the topic model does not know which paragraphs make up a "Risk Factors" section document.

Selecting an appropriate number of topics, K, can be challenging. Regarding quantitative evaluation methods, we focus on coherence as a measure of semantic coherence. Perplexity, also known as the predictive likelihood or the per-word likelihood bound, is a common measure of model fit in machine learning (Hoffman et al., 2010). However, perplexity is not strongly correlated to human judgment (Chang et al., 2009). Coherence is a proxy for how interpretable the topics are for a human evaluator, computed by how semantically similar words are grouped within the same topic. To compute coherence, we use the four-stage pipeline described in Röder et al. (2015), designed to maximize correlation with human topic rankings. The four stages are segmentation of the word space, probability estimation to measure sub-set quality, confirmation measure to compute the support of different combinations of sub-sets, and aggregation to a single coherence score. The pipeline uses a sliding window of 110 words for cross-validation, where we count co-occurrences for the given words to calculate the normalized point-wise mutual information (NPMI) and cosine similarity. Coherence is measured on a scale from 0 to 1, where 1 represents perfect coherence.

We run three-fold cross-validation on 80% of our data (i.e., 101,074 paragraphs) and compute out-of-sample coherence on the remaining 20% in each of the three folds. We then average the three

coherence measures to compare average performance and standard deviation. Given values for K used in existing literature, we conduct cross-validation for all topic models with $K \in [2, 30]$ with a step size of 2. We find the Mallet implementation of LDA (McCallum, 2002) with $K = 20$ to perform consistently well with an average coherence of approximately 0.575. Figure E.1 shows the three-fold cross-validated, out-of-sample coherence with standard errors across different numbers of topics, K. The highest coherence with a relatively narrow standard error is at $K = 20$.

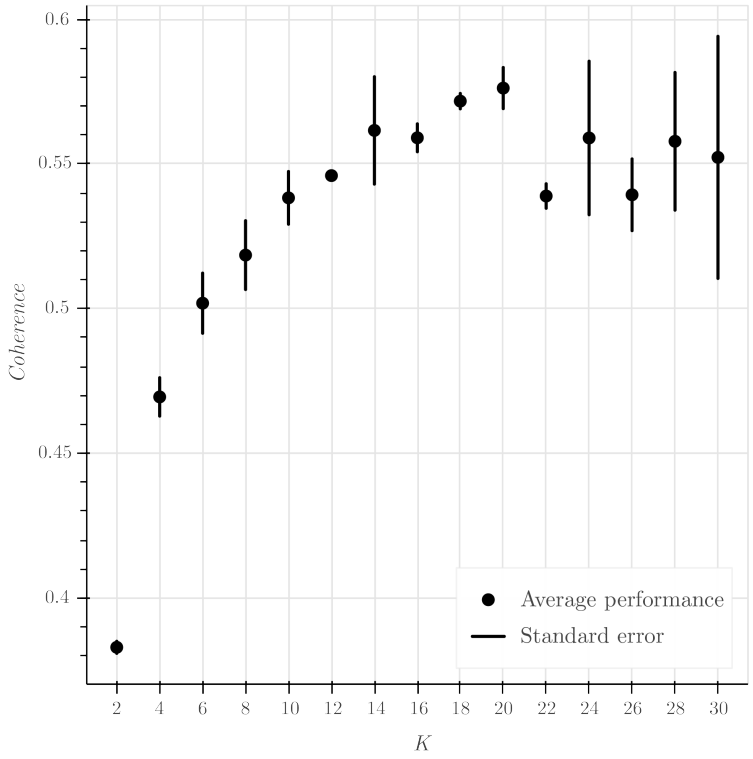

Figure E.1: Topic model coherence of the LDA Mallet model (McCallum, 2002).

E.3 Illustrative Example

Here, we illustrate how to compute aggregate risk disclosure with a toy example. The example consists of three topics ($K = 3$), four documents ($D = 4$), two-year groups and three industries. D_i represents document i, P_{ij} represents paragraph j in document i and T_k represents topic k.

First, we use a topic model to estimate topic loadings for all paragraphs, as shown in Table E.1. Standard LDA output has row-wise sums equal to one.

		T_1	T_2	T_3	Σ
D_1	P_{11}	0.2	0.1	0.7	1
	P_{12}	0.02	0.82	0.16	1
	P_{13}	0.16	0	0.84	1
	P_{14}	0.89	0.04	0.07	1
D_2	P_{21}	0.76	0.14	0.1	1
	P_{22}	0.09	0.51	0.4	1
D_3	P_{31}	0.82	0.1	0.08	1
	P_{32}	0.9	0.03	0.07	1
	P_{33}	0.12	0.79	0.09	1
D_4	P_{41}	0.91	0.08	0.01	1
	P_{42}	0.02	0.24	0.74	1
	P_{43}	0.12	0.3	0.58	1
	P_{44}	0.08	0	0.92	1
	P_{45}	0.02	0.2	0.78	1
⋮	⋮	⋮	⋮	⋮	⋮

Table E.1: Paragraph-level LDA output.

Second, we take the maximum topic loading for each paragraph and encode that topic-paragraph pair to one, zero otherwise[1]. Table E.2 shows the dominant topics.

		T_1	T_2	T_3	Σ
D_1	P_{11}	0	0	1	1
	P_{12}	0	1	0	1
	P_{13}	0	0	1	1
	P_{14}	1	0	0	1
D_2	P_{21}	1	0	0	1
	P_{22}	0	1	0	1
D_3	P_{31}	1	0	0	1
	P_{32}	1	0	0	1
	P_{33}	0	1	0	1
D_4	P_{41}	1	0	0	1
	P_{42}	0	0	1	1
	P_{43}	0	0	1	1
	P_{44}	0	0	1	1
	P_{45}	0	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table E.2: Dominant topics per paragraph.

[1]To encode a topic-paragraph pair to one, we might introduce a minimum loading value for the dominant topic or a minimum distance ahead of the next topic. The current implementation has no such restrictions.

Next, we aggregate paragraphs to documents by counting dominant topics for each document D_i. Each document contains the year grouping and Fama and French (1997) industry data as shown in Table E.3. Year and industry groups are required to adjust topic counts.

	T_1	T_2	T_3	year	industry
D_1	1	1	2	1996-1997	C
D_2	1	1	0	1996-1997	B
D_3	2	1	0	1996-1997	A
D_4	1	0	4	1998-1999	B
⋮	⋮	⋮	⋮	⋮	⋮

Table E.3: Counts of dominant topic paragraphs per document.

To compute aggregate risk disclosure relative to a focal IPO's year and industry group, we compute the year-industry mean and year-industry standard deviation of the dominant topic counts. The number of years to group and the number of industries are tunable hyperparameters. Increasing hyperparameter values increases the likelihood of not observing an IPO in a given year-industry group, which leads to a division-by-zero error. Hence, we add one (+1) to all standard deviations. Alternatively, we replace the year-industry standard deviations equal to zero with one, which does not change the results in a meaningful way.

Table E.4 shows the made-up mean of the counts of document-level dominant paragraphs by two-year and industry group. For example, in industry A, the average IPO in 1996-1997 disclosed 0.35 paragraphs about topic T_1. We came up with the values for means and standard deviations for illustrative purposes. Table E.5 shows the standard deviation of the counts of dominant paragraphs by two-year and industry group.

year	industry	T_1	T_2	T_3
1996-1997	A	0.35	0.63	1.12
	B	0.57	1.13	1.07
	C	0.4	2.63	0.83
1998-1999	A	0.66	1.17	1.2
	B	0.74	1.69	1.12
	C	0.48	0.54	0.06
⋮	⋮	⋮	⋮	⋮

Table E.4: Means of the counts of IPO document dominant paragraphs by normalization group.

year	industry	T_1	T_2	T_3
1996-1997	A	1.46	1.77	1.08
	B	1.95	1.93	1
	C	1.59	3.54	1.1
1998-1999	A	1.31	1.25	1.1
	B	1.91	1.04	1.52
	C	1.87	2.12	1.56
⋮	⋮	⋮	⋮	⋮

Table E.5: Standard deviations of paragraph counts of dominant topics by normalization group.

Finally, we z-score all topic loadings. Intuitively, a firm's exposure to a focal risk topic can increase (decrease) due to two reasons. First, the firm's year and industry group discuss the focal risk topic less (more) frequently. Second, the firm's industry discusses the focal risk topic with a lower (higher) variation. Table E.6 and Table E.7 show how aggregate risk disclosure is computed.

	T_1	T_2	T_3
D_1	=(1-0.4)/1.59	=(1-2.63)/3.54	=(2-0.83)/1.1
D_2	=(1-0.57)/1.95	=(1-1.13)/1.93	=(0-1.07)/1
D_3	=(2-0.35)/1.46	=(1-0.63)/1.77	=(0-1.12)/1.08
D_4	=(1-0.74)/1.91	=(0-1.69)/1.04	=(4-1.12)/1.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table E.6: Computing aggregate risk disclosure: intermediary step.

	T_1	T_2	T_3	$\Sigma = AggregateRisk$
D_1	0.38	-0.46	1.06	0.98
D_2	0.22	-0.07	-1.07	-0.92
D_3	1.13	0.21	-1.04	0.3
D_4	0.14	-1.625	1.89	0.41
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table E.7: Aggregate risk disclosure as the sum of all normalized risk topics.

The resulting normalized risk topics have a mean of zero and a standard deviation close to one. We can see in Table E.7 that firms with more paragraphs of risk disclosed (D_1 and D_4) have higher aggregate risk disclosure. We set up the toy example to work out this way to convey the intuition behind aggregate risk disclosure, of course.

E.4 Lorenz Curve

A core assumption of our approach to measuring aggregate risk disclosure is that, on average, one paragraph discusses one risk topic. Reading through numerous "Risk Factors" sections suggests that the assumption holds, but we investigate further using a Lorenz curve. Figure E.2 shows the cumulative share of topics from lowest to highest topic loading. The topic loadings are the raw outputs of the topic model. Intuitively, we expect one dominant topic for each paragraph (i.e., a solid line that goes through the bottom right) rather than a uniform distribution across all K topics (i.e., the dashed diagonal line of perfect equality). The Lorenz curve in Figure E.2 shows that the solid line is indeed relatively far away from the dashed diagonal, suggesting that paragraphs generally have a dominant topic loading.

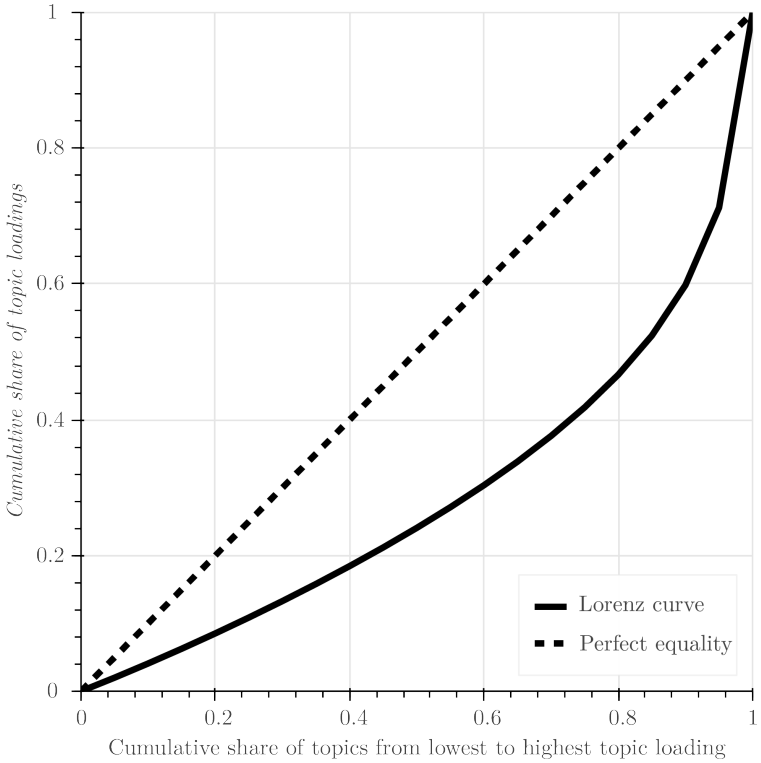

Figure E.2: Lorenz curve for paragraph-level topic loadings.

Appendix F

Data Pipeline

In this section, we describe the IPO data ingestion and processing pipeline.

Variable	Description	Source
Underpricing	Difference between the offer price and the first-day closing price (Beatty and Welch, 1996)	SDC, CRSP
Age	Firm age	SDC
Boom	Indicator for IPO boom years (1 January 1997 - 1 April 2000) (Aggarwal et al., 2009)	SDC
HighTech	Indicator equal to one if SDC classifies firm as high-tech	SDC
Hot	Hot markets measure as the count of IPOs in the same 4-digit SIC code in the previous year	SDC
NYSE	Indicator equal to one if the firms issued on New York Stock Exchange (NYSE)	SDC
PE	Indicator equal to one if the firm was funded by a private equity firm at the time of the IPO filing	SDC
Proceeds	Amount for the entire transaction plus overallotment amount (or green shoe) sold	SDC
Prominence	Indicator equal to one if the VC firm was among the top 30 investors in the prior year (Gulati and Higgins, 2003)	SDC
Reputation	Tombstone ranking of lead underwriter (maximum if more than one) from Carter and Manaster (1990)	SDC
VC	Indicator equal to one if the firm was funded by a venture capital firm at the time of the IPO filing	SDC
Assets	Total assets	Compustat
Book	Book value	Compustat
Ni	Net income	Compustat
Revenue	Total revenue	Compustat
ROA	Return on assets	Compustat
Patent	Indicator equal to one if the firm had at least one granted patent at the time of IPO	USPTO
EGC	Indicator equal to one if the firm self-classified as an emerging growth company (EGC) in the JOBS Act	SEC EDGAR
US	Indicator equal to one if the firms is headquartered in the US at the time of IPO	SEC EDGAR
Vola	Post-IPO volatility (15, 30, and 90 days)	CRSP
AggregateRisk	Aggregate risk disclosure based on the "Risk Factors" section, as described in Section 4.3.4	SEC EDGAR
TechRisk	Technology risk disclosure based on the "Risk Factors" section, as described in Section 4.3.4	SEC EDGAR
OtherRisk	Risk disclosure without technology risk based on the "Risk Factors" section, as described in Section 4.3.4	SEC EDGAR
Year	Issue year	SDC
Industry	Industry classification based on mapping in Fama and French (1997)	Compustat, CRSP

Table F.1: Summary of variables.

F.1 Sample Selection and Financial Meta-Data

The sample is defined using SDC Global New Issues (GNI) data. Figure F.1 outlines the major steps of the data pipeline.

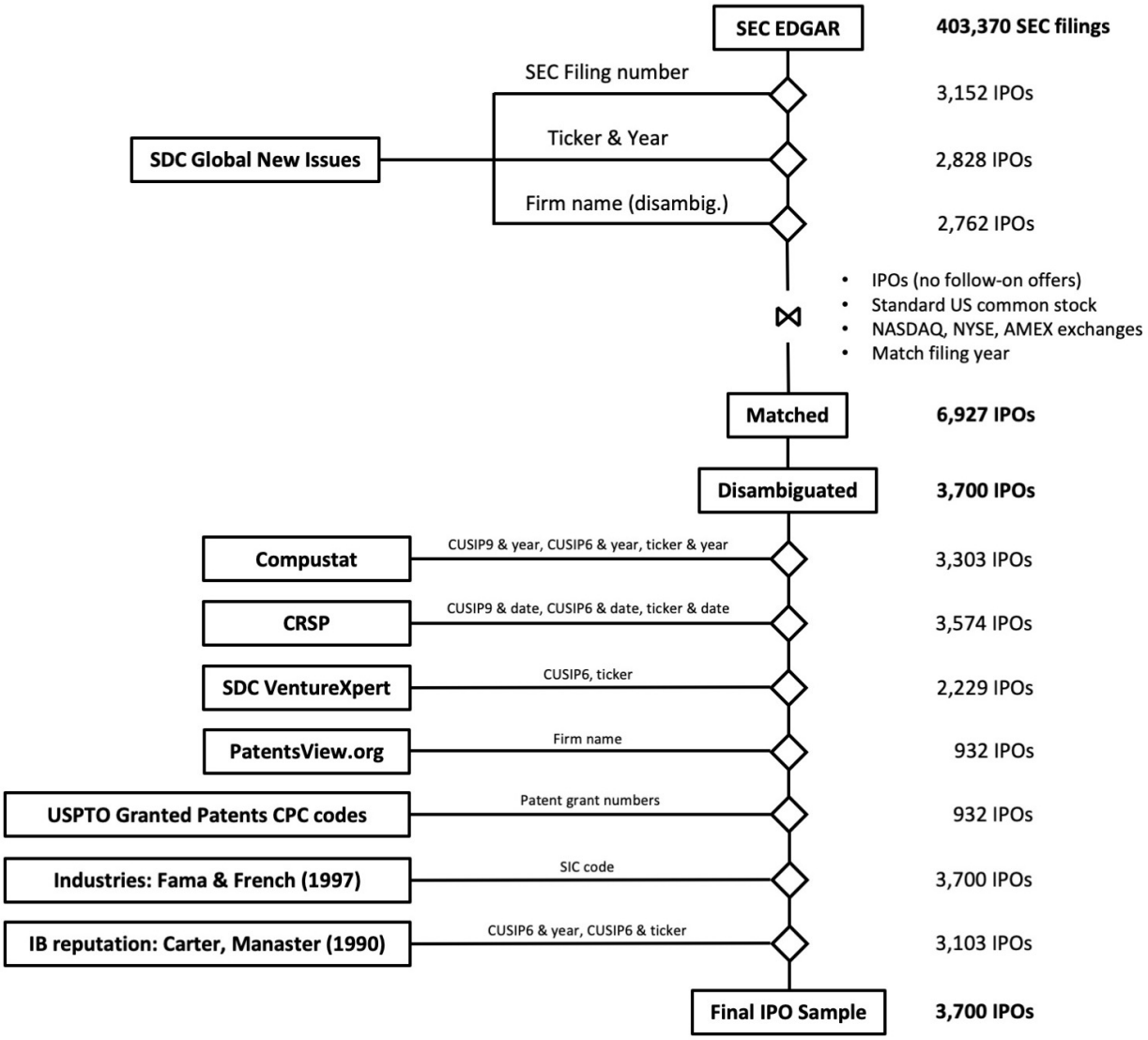

Figure F.1: Combining various data sources to build a sample of 3,700 IPOs.

First, we access SDC GNI and download all IPOs with common stock issued between 1.1.1996 and 31.12.2018. We download all corresponding meta-data fields available. Next, we define our

sample along the following dimensions. First, we IPOs are required to have the Standard Initial Public Offering Eligible Flag set to true, which eliminates non-underwritten transactions and transactions without a manager. Next, we require the Exchange Location to be in the US and restrict our sample to the main US stock exchanges: NASDAQ, NYSE, and AMEX. Finally, we only consider issues with Master Deal Type equal to C (i.e., US Common Stock, Class A Common Shares, Ordinary Shares). We count unique IPOs based on their CUSIP code throughout the entire data pipeline. The SDC GNI data consists of 4,935 IPOs.

Secondly, we augment SDC GNI data with filings from the US Securities and Exchange Commission (SEC) EDGAR database. Mergent, Inc., a subsidiary of the London Stock Exchange Group, provided us with all SEC 424(a) and 424(b)-type filings (also called *prospectuses*) for IPOs between 1996 and 2018. These data include 403,370 unique HTML and plain text documents. We combine three different merging approaches to join SEC EDGAR and SDC GNI data. Each approach disambiguates matched entries on the filing year. First, we match on disambiguated firm name and at least one of ZIP code, SEC code, or ticker symbol. First, we match on SEC filing number, which refers to the registration number used by the SEC. 2,353 IPOs match on the SEC filing number. Secondly, we match on ticker symbol and filing year. Each match on ticker and filing year must also match at least one of ZIP code or SEC filing numbers. 2,828 IPOs match on ticker and filing year. Finally, using a software tool developed by the TIS Lab at EPFL called *bizy*, we disambiguate firm names by removing suffixes, special symbols, and other ambiguous sub-strings, lowercase firm names, and run a fuzzy match based on the Levenshtein distance measure. Manual inspection at different cut-off points generates the set of tuples of matched firm names. 2,762 IPOs match on the firm name. Matched IPOs from all three approaches are then concatenated, resulting in 6,927 IPOs, before removing duplicates. Due to the importance of text data for downstream analyses, we require each IPO firm to have an entry in the SDC GNI and SEC data. The final, de-duplicated,

and disambiguated sample includes 3,700 IPOs.

Thirdly, we augment SDC GNI and SEC sample with these data sources: Compustat, Center for Research in Security Prices (CRSP), SDC VentureXpert, patents data from PatensView.org, the US Patent Office (USPTO), Fama and French (1997) industry classifications, and Loughran and Ritter (2004) investment banking underwriter reputation data as defined by Carter and Manaster (1990). The first three data sources are queried through the Wharton Research Data Services (WRDS), using the ticker symbol and CUSIP identifiers to increase the number of unique samples returned by WRDS. First, as explained earlier, we match Compustat data by combining three different approaches. We match on CUSIP9 and issue year, CUSIP6 and issue year, and ticker symbol and issue year. 3,303 Compustat entries match. Second, we match CRSP data on CUSIP9 and issue date, CUSIP6 and issue date, and ticker symbol and issue date to find 3,574 CRSP matches. Next, we match one-line SDC VentureXpert data on the industry statistics level on CUSIP6. These data include the VC backing dummy, for example. 2,229 VentureXpert entries match. Next, we query PatentsView.org (i.e., the rawassginee.tsv file) for a list of patent assignee firm names, which we disambiguate with the *bizy* software tool, developed by the TIS Lab at EPFL. Next, we compute firm-level patent portfolios *at the time of IPO*. To do so, another software tool developed by the TIS Lab at EPFL called *paty* automatically downloads and joins patent data from PatentsView.org for later use. For each of the 3,700 IPO firms in the sample, we find the patent grant numbers with a grant date on or before the IPO issue date. The patent grant numbers associated with each IPO firm come from *paty*/PatentsView.org. We find 932 patent IPO firms with 19,470 granted patents (i.e., an average of 21 patents per firm) at the time of IPO. We then concatenate Compustat and CRSP SIC codes to create a complete SIC code master field. We map the SIC code to Fama and French (1997) 5, 12, and 48 industries classifications, which are "defined with the goal of having a manageable number of distinct industries that cover all NYSE, AMEX, and NASDAQ

stocks" (Fama and French, 1997, p. 156). Finally, the underwriter reputation data from Carter and Manaster (1990), published in Loughran and Ritter (2004) is a score from 1 to 9, where more reputable IB firms have higher scores than less reputable IB firms. We successfully match 3,103 IPO firms using CUSIP6 and year, and CUSIP6 and ticker symbol.

To conclude this section, the final sample contains 3,700 IPO firms. In the absence of a static and universally unique identifier, the sample contains missing values, which we treat on a case-by-case basis.

F.2 Risk Factor Parsing

SEC 424(b)-type prospectuses include a "Risk Factors" section, which describes various types of risks investors face when purchasing shares in the issuing firm. Here, we describe the steps implemented to extract these sections from the SEC EDGAR data.

Due to the high volume of data of approximately 1 TB, we configure a distributed PySpark (a combination of Python and Spark) cluster on Google Cloud. More specifically, the risk factors parsing step takes approximately 30 minutes on a high-memory master node with 8 CPUs and 20 worker nodes with eight high-memory CPUs each. The key steps of Risk Factor Parsing are outlined in Figure F.2.

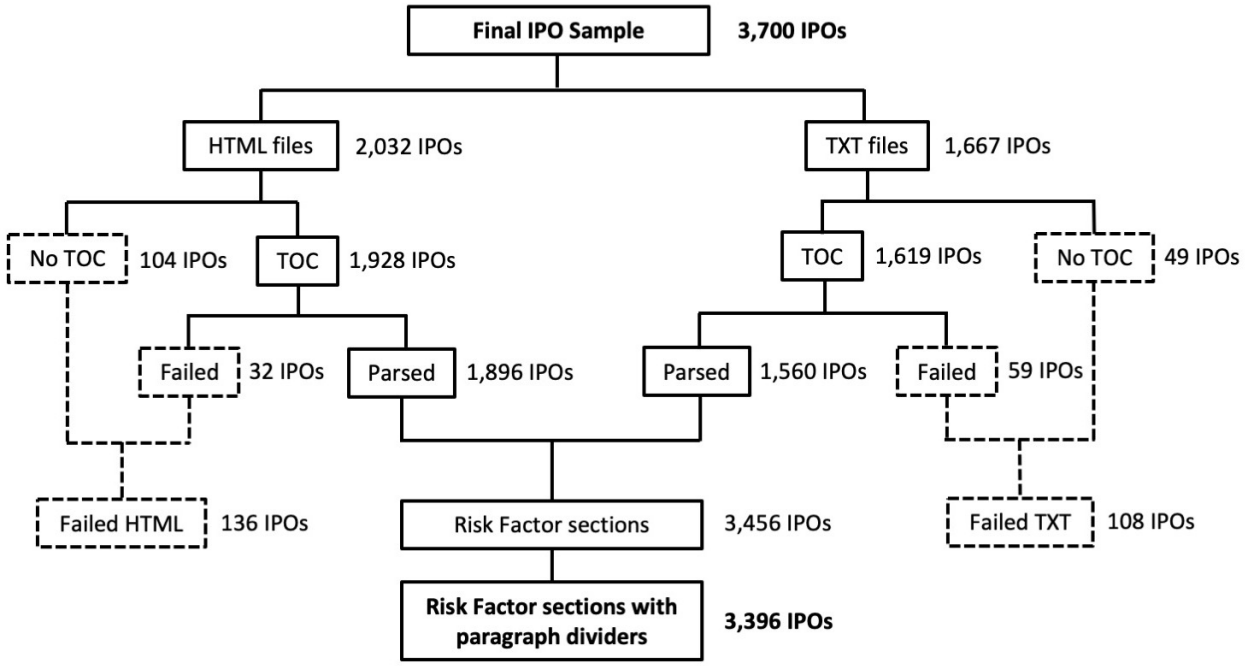

Figure F.2: Risk Factors parsing pipeline based on HTML and plain text SEC IPO prospectuses.

We split the data into HTML and plain text (txt) files to extract the Risk Factors section. Older prospectuses tend to be stored in plain text, while more recent ones tend to be stored in HTML format. We first identify the table of contents (TOC) for either type to identify the starting and ending page numbers. TOCs cannot be automatically identified in 104 HTML and 49 plain text prospectuses. All other prospectuses are parsed, resulting in 3,456 extracted Risk Factor sections. 244 prospectuses (136 HTML and 108 plain text) failed to be extracted automatically and will have to be reviewed manually.

Risk Factors sections are generally well-structured, consisting of a set of paragraphs. Extracting these paragraphs might be useful for applying advanced text analysis later on. We successfully extract 3,456 "Risk Factors" sections, of which 3,396 filings have machine-readable paragraph dividers.

Bibliography

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge (2017): "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

Aboody, David and Baruch Lev (2000): "Information asymmetry, R&D, and insider gains," *The Journal of Finance*, 55, 2747–2766.

Acemoglu, Daron and Pascual Restrepo (2018a): "Artificial intelligence, automation, and work," in *The economics of artificial intelligence: An agenda*, University of Chicago Press, 197–236.

——— (2018b): "Modeling automation," in *AEA Papers and Proceedings*, vol. 108, 48–53.

——— (2019): "Automation and new tasks: How technology displaces and reinstates labor," *Journal of Economic Perspectives*, 33, 3–30.

Adler, Paul S (1995): "Interdepartmental interdependence and coordination: The case of the design/manufacturing interface," *Organization Science*, 6, 147–167.

Agarwal, Sumit, Sudip Gupta, and Ryan D Israelsen (2017): "Public and private information: Firm disclosure, SEC letters, and the JOBS act," *Georgetown McDonough School of Business Research Paper*, 17–4.

Aggarwal, Rajesh, Sanjai Bhagat, and Srinivasan Rangan (2009): "The impact of fundamentals on IPO valuation," *Financial Management*, 38, 253–284.

Agrawal, Ajay K, Joshua S Gans, and Avi Goldfarb (2021): "AI Adoption and System-Wide Change," Tech. rep., National Bureau of Economic Research.

Aguinis, Herman, David B Audretsch, Caroline Flammer, Klaus E Meyer, Mike W Peng, and David J Teece (2022): "Bringing the Manager Back Into Management Scholarship," *Journal of Management*.

Angrist, Joshua D and Jörn-Steffen Pischke (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24, 3–30.

Argote, Linda (1982): "Input uncertainty and organizational coordination in hospital emergency units," *Administrative Science Quarterly*, 420–434.

Arnold, Tom, Raymond PH Fishe, and David North (2010): "The effects of ambiguous information on initial and subsequent IPO returns," *Financial Management*, 39, 1497–1519.

Arrow, Kenneth Joseph (1972): "Economic welfare and the allocation of resources for invention," *The RAND Journal of Economics*.

Arthur, W Brian (2009): *The nature of technology: What it is and how it evolves*, Simon and Schuster.

Autor, David (2015): "Why are there still so many jobs? The history and future of workplace automation," *Journal of Economic Perspectives*, 29, 3–30.

Autor, David H and David Dorn (2013): "The growth of low-skill service jobs and the polarization of the US labor market," *American Economic Review*, 103, 1553–97.

Autor, David H, Frank Levy, and Richard J Murnane (2003): "The skill content of recent technological change: An empirical exploration," *The Quarterly Journal of Economics*, 118, 1279–1333.

Bailey, Diane E, Samer Faraj, Pamela J Hinds, Paul M Leonardi, and Georg von Krogh (2022): "We are all theorists of technology now: A relational perspective on emerging technology and organizing," *Organization Science*, 33, 1–18.

Bao, Yang and Anindya Datta (2014): "Simultaneously discovering and quantifying risk types from textual risk disclosures," *Management Science*, 60, 1371–1391.

Barney, Jay (1986a): "Organizational culture: can it be a source of sustained competitive advantage?" *Academy of Management Review*, 11, 656–665.

——— (1986b): "Strategic factor markets: Expectations, luck, and business strategy," *Management Science*, 32, 1231–1241.

——— (1991): "Firm resources and sustained competitive advantage," *Journal of Management*, 17, 99–120.

Baron, David P (1982): "A model of the demand for investment banking advising and distribution services for new issues," *The Journal of Finance*, 37, 955–976.

Barry, Christopher B, Chris J Muscarella, John W Peavy Iii, and Michael R Vetsuypens (1990): "The role of venture capital in the creation of public companies: Evidence from the going-public process," *Journal of Financial Economics*, 27, 447–471.

Bartov, Eli, Partha Mohanram, and Chandrakanth Seethamraju (2002): "Valuation of internet stocks—an IPO perspective," *Journal of Accounting Research*, 40, 321–346.

Bean, Randy (2022): "Data & AI Leadership Executive Survey 2022," Tech. rep., NewVantage Partners.

Beatty, Randolph P and Ivo Welch (1996): "Issuer expenses and legal liability in initial public offerings," *The Journal of Law and Economics*, 39, 545–602.

Becker, Gary S and Kevin M Murphy (1992): "The division of labor, coordination costs, and knowledge," *The Quarterly Journal of Economics*, 107, 1137–1160.

Belford, Mark, Brian Mac Namee, and Derek Greene (2018): "Stability of topic modeling via matrix factorization," *Expert Systems with Applications*, 91, 159–169.

Bessen, James (2015): "Toil and technology: Innovative technology is displacing workers to new jobs rather than replacing them entirely," *Finance & Development*, 52.

Bird, Steven, Ewan Klein, and Edward Loper (2009): *Natural language processing with Python: analyzing text with the natural language toolkit*, O'Reilly Media, Inc.

Blei, David M (2012): "Probabilistic topic models," *Communications of the ACM*, 55, 77–84.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003): "Latent dirichlet allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Bloom, Nicholas, Philip Bunn, Paul Mizen, Pawel Smietanka, and Gregory Thwaites (2020): "The impact of Covid-19 on productivity," Tech. rep., National Bureau of Economic Research.

Bloom, Nicholas, Steven J Davis, and Yulia Zhestkova (2021): "Covid-19 shifted patent applications toward technologies that support working from home," in *AEA Papers and Proceedings*, vol. 111, 263–66.

Bloom, Nicholas and John Van Reenen (2007): "Measuring and explaining management practices across firms and countries," *The Quarterly Journal of Economics*, 122, 1351–1408.

Bolton, Patrick and Mathias Dewatripont (1994): "The firm as a communication network," *The Quarterly Journal of Economics*, 109, 809–839.

Brau, James C, James Cicon, and Grant McQueen (2016): "Soft strategic information and IPO underpricing," *Journal of Behavioral Finance*, 17, 1–17.

Bresnahan, Timothy F and Manuel Trajtenberg (1995): "General purpose technologies 'Engines of growth'?" *Journal of Econometrics*, 65, 83–108.

Brown, Shona L and Kathleen M Eisenhardt (1997): "The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations," *Administrative Science Quarterly*, 1–34.

Bruton, Garry D and Dev Prasad (1997): "Strategy and IPO market selection: Implications for the entrepreneurial firm," *Journal of Small Business Management*, 35, 1.

Brynjolfsson, Erik and Andrew McAfee (2011): *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*, Brynjolfsson and McAfee.

——— (2014): *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*, WW Norton & Company.

——— (2017): "The Business of Artificial Intelligence: What It Can—and Cannot—Do for Your Organization," *Harvard Business Review*, 7, 3–11.

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

Brynjolfsson, Erik and Tom Mitchell (2017): "What can machine learning do? Workforce implications," *Science*, 358, 1530–1534.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson (2021): "The productivity J-curve: How intangibles complement general purpose technologies," *American Economic Journal: Macroeconomics*, 13, 333–72.

Burgelman, Robert A (1991): "Intraorganizational ecology of strategy making and organizational adaptation: Theory and field research," *Organization Science*, 2, 239–262.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu (2020): "The structure of economic news," Tech. rep., National Bureau of Economic Research.

Campbell, John L, Hsinchun Chen, Dan S Dhaliwal, Hsin-min Lu, and Logan B Steele (2014): "The information content of mandatory risk factor disclosures in corporate filings," *Review of Accounting Studies*, 19, 396–455.

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L Zhang (2020): "How to talk when a machine is listening: Corporate disclosure in the age of AI," Tech. rep., National Bureau of Economic Research.

Carter, Richard and Steven Manaster (1990): "Initial public offerings and underwriter reputation," *The Journal of Finance*, 45, 1045–1067.

Castellanos, Sara (2020): "Tyson Takes Computer Vision to the Chicken Plant," *Wall Street Journal*.

Certo, S Trevis, Jeffrey G Covin, Catherine M Daily, and Dan R Dalton (2001): "Wealth and the effects of founder management among IPO-stage new ventures," *Strategic Management Journal*, 22, 641–658.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei (2009): "Reading tea leaves: How humans interpret topic models," in *Advances in Neural Information Processing Systems*, 288–296.

Charles, Kerwin Kofi, Erik Hurst, and Matthew Notowidigdo (2013): "Manufacturing decline, housing booms, and non-employment," *Chicago Booth Research Paper*.

Charmaz, Kathy (2006): *Constructing grounded theory: A practical guide through qualitative analysis*, SAGE Publications.

——— (2014): *Constructing grounded theory*, SAGE Publications.

Chenail, Ronald J (2011): "Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research." *Qualitative Report*, 16, 255–262.

Chondrakis, George, Carlos J Serrano, and Rosemarie H Ziedonis (2021): "Information disclosure and the market for acquiring technology companies," *Strategic Management Journal*, 42, 1024–1053.

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal (2020): "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 41, 1381–1411.

Choudhury, Prithwiraj, Dan Wang, Natalie Carlson, and Tarun Khanna (2019): "Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles," *Available at SSRN 3392448*.

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky (2021): "The state of AI in 2021," Tech. rep., McKinsey & Company.

Clemons, Eric K and Michael C Row (1991): "Sustaining IT advantage: The role of structural differences," *MIS Quarterly*, 275–292.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern (2018): "The impact of artificial intelligence on innovation," Tech. rep., National Bureau of Economic Research.

——— (2019): *The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis*, University of Chicago Press.

Cockburn, Iain M, Rebecca M Henderson, and Scott Stern (2000): "Untangling the origins of competitive advantage," *Strategic Management Journal*, 21, 1123–1145.

Cohen, Wesley M, Richard R Nelson, and John P Walsh (2000): "Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not)," Tech. rep., National Bureau of Economic Research.

Corley, Kevin G and Dennis A Gioia (2004): "Identity ambiguity and change in the wake of a corporate spin-off," *Administrative Science Quarterly*, 49, 173–208.

Corrado, Carol A and Charles R Hulten (2010): "How do you measure a" technological revolution"?" *American Economic Review*, 100, 99–104.

Creswell, John W and Vicki L Plano Clark (2017): *Designing and conducting mixed methods research*, SAGE Publications.

Crowston, Kevin (1997): "A coordination theory approach to organizational process design," *Organization Science*, 8, 157–175.

Daft, Richard L and Karl E Weick (1984): "Toward a model of organizations as interpretation systems," *Academy of Management Review*, 9, 284–295.

Davenport, Thomas H and Nitin Mittal (2020): "How CEOs Can Lead a Data-Driven Culture," *Harvard Business Review*.

Davis, Jason P, Kathleen M Eisenhardt, and Christopher B Bingham (2007): "Developing theory through simulation methods," *Academy of Management Review*, 32, 480–499.

Deming, David J (2017): "The growing importance of social skills in the labor market," *The Quarterly Journal of Economics*, 132, 1593–1640.

Denrell, Jerker, Christina Fang, and Daniel A Levinthal (2004): "From T-mazes to labyrinths: Learning from model-based feedback," *Management Science*, 50, 1366–1378.

Denrell, Jerker and James G March (2001): "Adaptation as information restriction: The hot stove effect," *Organization Science*, 12, 523–538.

Dierickx, Ingemar and Karel Cool (1989): "Asset stock accumulation and sustainability of competitive advantage," *Management Science*, 35, 1504–1511.

DiMaggio, Paul, Manish Nag, and David Blei (2013): "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding," *Poetics*, 41, 570–606.

Dogan, Mustafa and Pinar Yildirim (2021): "Managing automation in teams," *Journal of Economics & Management Strategy*.

Doidge, Craig, Kathleen M Kahle, G Andrew Karolyi, and René M Stulz (2018): "Eclipse of the public corporation or eclipse of the public markets?" *Journal of Applied Corporate Finance*, 30, 8–16.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart (2018): "How to make causal inferences using texts," *arXiv preprint arXiv:1802.02163*.

Eggers, Jamie P and Sarah Kaplan (2009): "Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change," *Organization Science*, 20, 461–477.

Eisenhardt, Kathleen M (1985): "Control: Organizational and economic approaches," *Management Science*, 31, 134–149.

——— (1989a): "Building theories from case study research," *Academy of Management Review*, 14, 532–550.

——— (1989b): "Making fast strategic decisions in high-velocity environments," *Academy of Management Journal*, 32, 543–576.

Epstein, Larry G and Martin Schneider (2008): "Ambiguity, information quality, and asset pricing," *The Journal of Finance*, 63, 197–228.

Ethiraj, Sendil K, Alfonso Gambardella, and Constance E Helfat (2016): "Replication in strategic management," *Strategic Management Journal*, 37, 2191–2192.

Evgeniou, Theodoros and Pal Boza (2020): "E.ON: Building a New AI Powered Energy World," *Insead*.

Fama, Eugene F (1970): "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, 25, 383–417.

Fama, Eugene F and Kenneth R French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

——— (1997): "Industry costs of equity," *Journal of Financial Economics*, 43, 153–193.

Fang, Christina and Daniel Levinthal (2009): "Near-term liability of exploitation: Exploration and exploitation in multistage problems," *Organization Science*, 20, 538–551.

Feldman, Martha S and Brian T Pentland (2003): "Reconceptualizing organizational routines as a source of flexibility and change," *Administrative Science Quarterly*, 48, 94–118.

Ferris, Stephen P, Qing Hao, and Min-Yu Liao (2012): "The effect of issuer conservatism on IPO pricing and performance," *Review of Finance*, 17, 993–1027.

Field, Laura Casares and Gordon Hanka (2001): "The expiration of IPO share lockups," *The Journal of Finance*, 56, 471–500.

Fleming, Nic (2018): "Computer-calculated compounds," *Nature*, 557, S55–S57.

Frey, Carl Benedikt and Michael A Osborne (2017): "The future of employment: How susceptible are jobs to computerisation?" *Technological Forecasting and Social Change*, 114, 254–280.

Furr, Nathan R (2021): "Technology Entrepreneurship, Technology Strategy, and Uncertainty," in *Strategic Management: State of the Field and Its Future*, Oxford University Press, chap. 3, 205.

Furr, Nathan R and Kathleen M Eisenhardt (2021): "Strategy and uncertainty: Resource-based view, strategy-creation view, and the hybrid between them," *Journal of Management*, 47, 1915–1935.

Gans, Joshua S and Scott Stern (2010): "Is there a market for ideas?" *Industrial and Corporate Change*, 19, 805–837.

Garicano, Luis and Yanhui Wu (2012): "Knowledge, communication, and organizational capabilities," *Organization Science*, 23, 1382–1397.

Gephart, Robert P Jr (1984): "Making sense of organizationally based environmental disasters," *Journal of Management*, 10, 205–225.

Gerlach, Martin, Tiago P Peixoto, and Eduardo G Altmann (2018): "A network approach to topic models," *Science Advances*, 4, eaaq1360.

Ghoshal, Sumantra and Christopher A Bartlett (1994): "Linking organizational context and managerial action: The dimensions of quality of management," *Strategic Management Journal*, 15, 91–112.

Gibson, Cristina B and Julian Birkinshaw (2004): "The antecedents, consequences, and mediating role of organizational ambidexterity," *Academy of Management Journal*, 47, 209–226.

Gioia, Dennis A, Kevin G Corley, and Aimee L Hamilton (2012): "Seeking qualitative rigor in inductive research: Notes on the Gioia methodology," *Organizational Research Methods*, 16, 15–31.

Glaser, Barney G and Anselm L Strauss (1967): *The discovery of grounded theory: strategies for qualitative research*, Aldine.

Greene, Jennifer C, Valerie J Caracelli, and Wendy F Graham (1989): "Toward a conceptual framework for mixed-method evaluation designs," *Educational Evaluation and Policy Analysis*, 11, 255–274.

Greenwald, Amy, Karthik Kannan, and Ramayya Krishnan (2010): "On evaluating information revelation policies in procurement auctions: A Markov decision process approach," *Information Systems Research*, 21, 15–36.

Griffiths, Thomas L and Mark Steyvers (2004): "Finding scientific topics," *Proceedings of the National academy of Sciences*, 101, 5228–5235.

Griliches, Zvi (1981): "Market value, R&D, and patents," *Economics Letters*, 7, 183–187.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart (2022): *Text as data: A new framework for machine learning and the social sciences*, Princeton University Press.

Gulati, Ranjay and Monica C Higgins (2003): "Which ties matter when? The contingent effects of interorganizational partnerships on IPO success," *Strategic Management Journal*, 24, 127–144.

Hanley, Kathleen Weiss (1993): "The underpricing of initial public offerings and the partial adjustment phenomenon," *Journal of Financial Economics*.

Hanley, Kathleen Weiss and Gerard Hoberg (2010): "The information content of IPO prospectuses," *The Review of Financial Studies*, 23, 2821–2864.

——— (2012): "Litigation risk, strategic disclosure and the underpricing of initial public offerings," *Journal of Financial Economics*, 103, 235–254.

——— (2019): "Dynamic Interpretation of Emerging Risks in the Financial Sector," *The Review of Financial Studies*.

Hannah, David R and Brenda A Lautsch (2011): "Counting in qualitative research: Why to conduct it, when to avoid it, and when to closet it," *Journal of Management Inquiry*, 20, 14–22.

Hannigan, Timothy R, Richard FJ Haans, Keyvan Vakili, Hovig Tchalian, Vern L Glaser, Milo Shaoqing Wang, Sarah Kaplan, and P Devereaux Jennings (2019): "Topic modeling in management research: Rendering new theory from textual data," *Academy of Management Annals*, 13, 586–632.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2017): "Transparency and deliberation within the FOMC: a computational linguistics approach," *The Quarterly Journal of Economics*, 133, 801–870.

Haskel, Jonathan and Stian Westlake (2018): *Capitalism without capital: The rise of the intangible economy*, Princeton University Press.

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun (2019): "Firm-level political risk: Measurement and effects," *The Quarterly Journal of Economics*, 134, 2135–2202.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015): "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021): "Deberta: Decoding-enhanced BERT with disentangled attention," in *International Conference on Learning Representations*.

Healy, Paul M and Krishna G Palepu (2001): "Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature," *Journal of Accounting and Economics*, 31, 405–440.

Heeley, Michael B, Sharon F Matusik, and Neelam Jain (2007): "Innovation, appropriability, and the underpricing of initial public offerings," *Academy of Management Journal*, 50, 209–225.

Hegde, Deepak, Baruch Lev, and Chenqi Zhu (2018): "Patent disclosure and price discovery," *Available at SSRN 3274837*.

Hermans, Raine and Ilkka Kauranen (2005): "Value creation potential of intellectual capital in biotechnology–empirical evidence from Finland," *R&D Management*, 35, 171–185.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010): "Online learning for latent dirichlet allocation," in *Advances in Neural Information Processing Systems*, 856–864.

Holmqvist, Mikael (2004): "Experiential learning processes of exploitation and exploration within and between organizations: An empirical study of product development," *Organization Science*, 15, 70–81.

Hsu, David H and Rosemarie H Ziedonis (2013): "Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents," *Strategic Management Journal*, 34, 761–781.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong Zheng (2018): "Analyst information discovery and interpretation roles: A topic modeling approach," *Management Science*, 64, 2833–2855.

Huang, Kejun, Xiao Fu, and Nikolaos D Sidiropoulos (2016): "Anchor-free correlated topic modeling: Identifiability and algorithm," *Advances in Neural Information Processing Systems*, 29.

Huang, Ke-Wei and Zhuolun Li (2011): "A multilabel text classification algorithm for labeling risk factors in SEC form 10-K," *ACM Transactions on Management Information Systems (TMIS)*, 2, 18.

Huber, Peter J (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221.

Israelsen, Ryan D (2014): "Tell it like it is: Disclosed risks and factor portfolios," *Available at SSRN 2504522*.

Itami, Hiroyuki and Thomas W Roehl (1991): *Mobilizing invisible assets*, Harvard University Press.

Jain, Bharat A and Omesh Kini (1994): "The post-issue operating performance of IPO firms," *The Journal of Finance*, 49, 1699–1726.

Jensen, Michael C and William H Meckling (1976): "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, 3, 305–360.

Jones, Benjamin F (2009): "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" *The Review of Economic Studies*, 76, 283–317.

Kahle, Kathleen M and René M Stulz (2017): "Is the US public corporation in trouble?" *Journal of Economic Perspectives*, 31, 67–88.

Kauffman, Stuart A et al. (1993): *The origins of order: Self-organization and selection in evolution*, Oxford University Press, USA.

Keisler, SB and L Sproull (1982): "Managerial Response to Changing Environments," *Administrative Science*.

Kobayashi, Vladimer B, Stefan T Mol, Hannah A Berkers, Gábor Kismihók, and Deanne N Den Hartog (2018): "Text mining in organizational research," *Organizational Research Methods*, 21, 733–765.

Kogut, Bruce and Udo Zander (1992): "Knowledge of the firm, combinative capabilities, and the replication of technology," *Organization Science*, 3, 383–397.

——— (1996): "What firms do? Coordination, identity, and learning," *Organization Science*, 7, 502–518.

Kravet, Todd and Volkan Muslu (2013): "Textual risk disclosures and investors' risk perceptions," *Review of Accounting Studies*, 18, 1088–1122.

Laney, Doug (2001): "3D data management: Controlling data volume, velocity and variety," *META Group Research Note*, 6, 1.

Lang, Mark and Russell Lundholm (1993): "Cross-sectional determinants of analyst ratings of corporate disclosures," *Journal of Accounting Research*, 31, 246–271.

Leamer, Edward E (1983): "Let's take the con out of econometrics," *The American Economic Review*, 73, 31–43.

Lee, Daniel D and H Sebastian Seung (2001): "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 556–562.

Leiblein, Michael J, Jeffrey J Reuer, and Todd Zenger (2018): "What makes a decision strategic?" *Strategy Science*, 3, 558–573.

Leone, Andrew J, Steve Rock, and Michael Willenborg (2007): "Disclosure of intended use of proceeds and underpricing in initial public offerings," *Journal of Accounting Research*, 45, 111–153.

Lev, Baruch (2000): *Intangibles: Management, measurement, and reporting*, Brookings institution press.

——— (2018): "Intangibles," *Available at SSRN 3218586*.

Lev, Baruch and Theodore Sougiannis (1996): "The capitalization, amortization, and value-relevance of R&D," *Journal of Accounting and Economics*, 21, 107–138.

Levinthal, Daniel and James G March (1981): "A model of adaptive organizational search," *Journal of Economic Behavior & Organization*, 2, 307–333.

Levinthal, Daniel A (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

Levinthal, Daniel A and James G March (1993): "The myopia of learning," *Strategic Management Journal*, 14, 95–112.

Li, Feng (2010): "Survey of the Literature," *Journal of Accounting literature*, 29, 143–165.

Lippman, Steven A and Richard P Rumelt (1982): "Uncertain imitability: An analysis of interfirm differences in efficiency under competition," *The Bell Journal of Economics*, 418–438.

Litov, Lubomir P, Patrick Moreton, and Todd R Zenger (2012): "Corporate strategy, analyst coverage, and the uniqueness paradox," *Management Science*, 58, 1797–1815.

Logue, Dennis E (1973): "On the pricing of unseasoned equity issues: 1965–1969," *Journal of Financial and Quantitative Analysis*, 8, 91–103.

Lopez-Lira, Alejandro (2020): "Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Loughran, Tim and Bill McDonald (2011): "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 66, 35–65.

——— (2013): "IPO first-day returns, offer price revisions, volatility, and form S-1 language," *Journal of Financial Economics*, 109, 307–326.

Loughran, Tim and Jay Ritter (2004): "Why has IPO underpricing changed over time?" *Financial Management*, 5–37.

Loughran, Tim and Jay R Ritter (2002): "Why don't issuers get upset about leaving money on the table in IPOs?" *The Review of Financial Studies*, 15, 413–444.

Lowry, Michelle, Roni Michaely, and Ekaterina Volkova (2020): "Information Revealed through the Regulatory Process: Interactions between the SEC and Companies ahead of Their IPO," *The Review of Financial Studies*.

Lowry, Michelle and Susan Shu (2002): "Litigation risk and IPO underpricing," *Journal of Financial Economics*, 65, 309–335.

MacCormack, Alan, Fiona Murray, and Erika Wagner (2013): "Spurring innovation through competitions," *MIT Sloan Management Review*, 55, 25.

Malone, Thomas W and Kevin Crowston (1994): "The interdisciplinary study of coordination," *ACM Computing Surveys (CSUR)*, 26, 87–119.

March, James and Herbert Simon (1958): *Organizations*, New York: Wiley.

March, James G (1991): "Exploration and exploitation in organizational learning," *Organization Science*, 2, 71–87.

Maritan, Catherine A (2001): "Capital investment as investing in organizational capabilities: An empirically grounded process model," *Academy of Management Journal*, 44, 513–531.

Markowitz, Harry (1952): "Portfolio Selection," *The Journal of Finance*, 7, 77–91.

Marr, Bernard (2018): "The Amazing Ways How Unilever Uses Artificial Intelligence to Recruit and Train Thousands of Employees," *Forbes*.

McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton (2012): "Big data: the management revolution," *Harvard Business Review*, 90, 60–68.

McCallum, Andrew Kachites (2002): "Mallet: A machine learning for language toolkit," `http://mallet.cs.umass.edu`.

Megginson, William L and Kathleen A Weiss (1991): "Venture capitalist certification in initial public offerings," *The Journal of Finance*, 46, 879–903.

Miles, Matthew B and A Michael Huberman (1984): "Drawing valid meaning from qualitative data: Toward a shared craft," *Educational Researcher*, 13, 20–30.

Mobius, Markus and Raphael Schoenle (2006): "The evolution of work," Tech. rep., National Bureau of Economic Research.

Mohr, John W and Petko Bogdanov (2013): "Introduction—Topic models: What they are and why they matter," *Poetics*.

Molina-Azorin, Jose F (2012): "Mixed methods research in strategic management: Impact and applications," *Organizational Research Methods*, 15, 33–56.

Mollick, Ethan (2012): "People and process, suits and innovators: The role of individuals in firm performance," *Strategic Management Journal*, 33, 1001–1015.

Morricone, Serena, Federico Munari, Raffaele Oriani, and Gaetan De Rassenfosse (2017): "Commercialization strategy and IPO underpricing," *Research Policy*, 46, 1133–1141.

Mueller, Hannes and Christopher Rauh (2018): "Reading between the lines: Prediction of political violence using newspaper text," *American Political Science Review*, 112, 358–375.

Mullainathan, Sendhil and Jann Spiess (2017): "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 31, 87–106.

Ocasio, William (1997): "Towards an attention-based view of the firm," *Strategic Management Journal*, 18, 187–206.

Ocasio, William and John Joseph (2005): "An attention-based theory of strategy formulation: Linking micro-and macroperspectives in strategy processes," in *Strategy Process*, Emerald Group Publishing Limited.

Okhuysen, Gerardo A and Beth A Bechky (2009): "10 coordination in organizations: An integrative perspective," *Academy of Management Annals*, 3, 463–502.

Orlikowski, Wanda J (1996): "Improvising organizational transformation over time: A situated change perspective," *Information Systems Research*, 7, 63–92.

Ouchi, William G (1980): "Markets, bureaucracies, and clans," *Administrative Science Quarterly*, 129–141.

O'Reilly III, Charles A and Michael L Tushman (2008): "Ambidexterity as a dynamic capability: Resolving the innovator's dilemma," *Research in Organizational Behavior*, 28, 185–206.

Park, Haemin Dennis and Pankaj C Patel (2015): "How does ambiguity influence IPO underpricing? The role of the signalling environment," *Journal of Management Studies*, 52, 796–818.

Penrose, Edith (1959): *The Theory of the Growth of the Firm*, Basil Blackwell.

Peteraf, Margaret A (1993): "The cornerstones of competitive advantage: a resource-based view," *Strategic Management Journal*, 14, 179–191.

Polanyi, Michael (1962): "Tacit knowing: Its bearing on some problems of philosophy," *Reviews of Modern Physics*, 34, 601.

Porter, Michael E (1980): "Industry structure and competitive strategy: Keys to profitability," *Financial analysts journal*, 36, 30–41.

Porter, Martin F et al. (1980): "An algorithm for suffix stripping." *Program*, 14, 130–137.

Posen, Hart E and Daniel A Levinthal (2012): "Chasing a moving target: Exploitation and exploration in dynamic environments," *Management Science*, 58, 587–601.

Powell, Thomas C and Anne Dent-Micallef (1997): "Information technology as competitive advantage: The role of human, business, and technology resources," *Strategic Management Journal*, 18, 375–405.

Raisch, Sebastian and Sebastian Krakowski (2020): "Artificial Intelligence and Management: The Automation-Augmentation Paradox," *Academy of Management Review*.

Raj, Manav and Robert Seamans (2019): "Primer on artificial intelligence and robotics," *Journal of Organization Design*, 8, 1–14.

Ransbotham, Sam, Shervin Khodabandeh, Ronny Fehling, Burt LaFountain, and David Kiron (2019): "Winning with AI," *MIT Sloan Management Review*, 61180.

Reagans, Ray, Linda Argote, and Daria Brooks (2005): "Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together," *Management Science*, 51, 869–881.

Reeds, R and RJ De Filippi (1990): "Causal ambiguity, barriers to imitation and sustainable advantage," *Academy of Management Review*, 15, 88–102.

Řehůřek, Radim and Petr Sojka (2010): "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: Elra, 45–50.

Repenning, Nelson P (2002): "A simulation-based approach to understanding the dynamics of innovation implementation," *Organization Science*, 13, 109–127.

Ritter, Jay R (1984): "The" hot issue" market of 1980," *Journal of Business*, 215–240.

Ritter, Jay R and Ivo Welch (2002): "A review of IPO activity, pricing, and allocations," *The Journal of Finance*, 57, 1795–1828.

Ritter, Jay R et al. (1998): "Initial public offerings," *Contemporary Finance Digest*, 2, 5–30.

Robert Baum, J and Stefan Wally (2003): "Strategic decision speed and firm performance," *Strategic Management Journal*, 24, 1107–1129.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand (2014): "Structural topic models for open-ended survey responses," *American Journal of Political Science*, 58, 1064–1082.

Rock, Daniel (2019): "Engineering value: The returns to technological talent and investments in artificial intelligence," *Available at SSRN 3427412*.

Rock, Kevin (1986): "Why new issues are underpriced," *Journal of Financial Economics*, 15, 187–212.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015): "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, Acm, 399–408.

Romanelli, Elaine and Michael L Tushman (1994): "Organizational transformation as punctuated equilibrium: An empirical test," *Academy of Management Journal*, 37, 1141–1166.

Rosokha, Yaroslav and Kenneth Younge (2020): "Motivating innovation: The effect of loss aversion on the willingness to persist," *Review of Economics and Statistics*, 102, 569–582.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo (2021): ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI," in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Schein, Edgar H (1981): "Does Japanese management style have a message for American managers?" *Societal Culture and Management*, 23, 55–67.

Sculley, David, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison (2015): "Hidden technical debt in machine learning systems," *Advances in Neural Information Processing Systems*, 28, 2503–2511.

Seamans, Robert and Manav Raj (2018): "AI, labor, productivity and the need for firm-level data," Tech. rep., National Bureau of Economic Research.

Shannon, Claude E (1948): "A mathematical theory of communication," *The Bell System Technical Journal*, 27, 379–423.

Sharpe, William F (1964): "Capital asset prices: A theory of market equilibrium under conditions of risk," *The Journal of Finance*, 19, 425–442.

Simon, Herbert A (1955): "A behavioral model of rational choice," *The Quarterly Journal of Economics*, 69, 99–118.

Smircich, Linda and Charles Stubbart (1985): "Strategic management in an enacted world," *Academy of Management Review*, 10, 724–736.

Smith, Adam (1965): *The Wealth of Nations*, Modern Library.

Srivastava, Anup (2014): "Why have measures of earnings quality changed over time?" *Journal of Accounting and Economics*, 57, 196–217.

Sutton, Robert I and Barry M Staw (1995): "What theory is not," *Administrative Science Quarterly*, 371–384.

Sutton, Richard S (1990): "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Machine Learning Proceedings*, Elsevier, 216–224.

Sutton, Richard S and Andrew G Barto (2018): *Reinforcement learning: An introduction*, MIT press.

Tambe, Prasanna (2014): "Big data investment, skills, and firm value," *Management Science*, 60, 1452–1469.

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

——— (1998): "Capturing value from knowledge assets: The new economy, markets for know-how, and intangible assets," *California Management Review*, 40, 55–79.

——— (2007): "Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance," *Strategic Management Journal*, 28, 1319–1350.

Teece, David J, Gary Pisano, and Amy Shuen (1997): "Dynamic capabilities and strategic management," *Strategic Management Journal*, 18, 509–533.

Teh, Yee W, Michael I Jordan, Matthew J Beal, and David M Blei (2005): "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Advances in Neural Information Processing Systems*, 1385–1392.

Thomke, Stefan and Walter Kuemmerle (2002): "Asset accumulation, interdependence and technological change: evidence from pharmaceutical drug discovery," *Strategic Management Journal*, 23, 619–635.

Thompson, Neil, Shuning Ge, and Yash M Sherry (2020): "Building the Algorithm Commons: Who discovered the algorithms that underpin computing in the modern enterprise?" *Global Strategy Journal*.

Toutanova, Kristina, Dan Klein, Christopher D Manning, and Yoram Singer (2003): "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for computational Linguistics, 173–180.

Treleaven, Philip and Bogdan Batrinca (2017): "Algorithmic regulation: automating financial compliance monitoring and regulation using AI and blockchain," *Journal of Financial Transformation*, 45, 14–21.

Tushman, Michael L and Philip Anderson (1986): "Technological discontinuities and organizational environments," *Administrative Science Quarterly*, 439–465.

Tushman, Michael L and Charles A O'Reilly III (1996): "Ambidextrous organizations: Managing evolutionary and revolutionary change," *California Management Review*, 38, 8–29.

Tushman, Michael L and Elaine Romanelli (1985): "Organizational evolution: A metamorphosis model of convergence and reorientation." *Research in Organizational Behavior*.

Valente, Marco (2008): "Pseudo-NK: an enhanced model of complexity," Tech. rep., LEM Working Paper Series.

Varian, Hal (2018): "Artificial intelligence, economics, and industrial organization," Tech. rep., National Bureau of Economic Research.

Vise, David A and Mark Malseed (2006): *The Google Story*, Pan Books.

Von Krogh, Georg (2018): "Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing," *Academy of Management Discoveries*.

Von Krogh, Georg, Shiko Ben-Menahem, and Yash Raj Shrestha (2021): "Artificial intelligence in strategizing: Prospects and challenges," .

Walsh, Mike (2020): "AI Should Change What You Do — Not Just How You Do It," *Harvard Business Review*.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2019): "Superglue: A stickier benchmark for general-purpose language understanding systems," *arXiv preprint arXiv:1905.00537*.

Webb, Eugene and Karl E Weick (1979): "Unobtrusive measures in organizational theory: A reminder," *Administrative Science Quarterly*, 24, 650–659.

Weick, KE and RL Daft (1983): "The Effectiveness of Interpretation Systems." in *Organizational Effectiveness: A Comparison of Multiple Models*, 71–93.

Weick, Karl E (1995): *Sensemaking in organizations*, vol. 3, SAGE Publications.

Wernerfelt, Birger (1984): "A resource-based view of the firm," *Strategic Management Journal*, 5, 171–180.

White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 817–838.

Wilkins, Alan L and William G Ouchi (1983): "Efficient cultures: Exploring the relationship between culture and organizational performance," *Administrative Science Quarterly*, 468–481.

Williamson, Oliver E (1973): "Markets and hierarchies: some elementary considerations," *The American Economic Review*, 63, 316–325.

——— (2002): "The theory of the firm as governance structure: from choice to contract," *Journal of Economic Perspectives*, 16, 171–195.

Wilson, H James and Paul R Daugherty (2018): "Collaborative intelligence: Humans and AI are joining forces," *Harvard Business Review*, 96, 114–123.

Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi (2007): "The increasing dominance of teams in production of knowledge," *Science*, 316, 1036–1039.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng (2013): "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 1445–1456.

Yang, Shinkyu and Erik Brynjolfsson (2001): "Intangible assets and growth accounting: evidence from computer investments," *Unpublished paper. MIT*, 85, 28.

Yang, Yi, Shimei Pan, Yangqiu Song, Jie Lu, Mercan Topkara, and JW Player (2016): "Improving Topic Model Stability for Effective Document Exploration," in *International Joint Conference on Artificial Intelligence*, 4223–4227.

Younge, Kenneth A (2012): "Science and the Mobility Discount: Evidence from Initial Public Offerings," Tech. rep., UC Berkeley Mimeo.

Zhang, Daniel, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Sakhaee Ellie, Yoav Shoham, Jack Clark, and Raymond Perrault (2022): "The AI Index 2021 Annual Report," Tech. rep., Stanford University, Human-Centered AI Institute.

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher (2021): "The AI Economist: Optimal Economic Policy Design via Two-level Deep Reinforcement Learning," *arXiv preprint arXiv:2108.02755*.

Zollo, Maurizio and Sidney G Winter (2002): "Deliberate learning and the evolution of dynamic capabilities," *Organization Science*, 13, 339–351.

Zott, Christoph (2003): "Dynamic capabilities and the emergence of intraindustry differential firm performance: insights from a simulation study," *Strategic Management Journal*, 24, 97–125.

Curriculum Vitae

MAXIMILIAN HOFER
Scheuchzerstrasse 47, 8006 Zurich, Switzerland +41 77 5346294 LinkedIn maximilian.hofer@epfl.ch

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland – *PhD in Management of Technology* **Oct 2018 – Present**
- Supervised by Professor Kenneth A. Younge, Chair of Technology & Innovation Strategy (TIS)
- My thesis uses computational methods to investigate how organizations manage technological resources and risks

University of Oxford, United Kingdom – *MSc in Advanced Computer Science* **Oct 2017 – Aug 2018**
- MSc project titled "Named Entity Recognition from Medical Text with Deep Neural Networks" (Distinction)
- Presented my MSc project at the Oxford Computer Science Conference 2018 (arXiv:1811.05468)

University College London (UCL), United Kingdom – *BSc in Management Science (First class)* **Sep 2014 – Jun 2017**
- BSc thesis: "Product Usability at Quid" in collaboration with Quid Inc, San Francisco (Distinction)
- Graduated at the top of my class; founded the UCL Austrian Society; led Marketing at the Economics & Finance Society

WORK EXPERIENCE

EPFL, Lausanne – *Teaching Assistant & Instructor* **Oct 2018 – Present**
- Led all practical programming parts of seven iterations of the 1-week "Data Science for Managers" executive education course, working with > 200 managers in Switzerland to solve practical use cases with modern data science methods
- Initiated and organized the 2019 "Computational Methods for Economists" summer school, winning a grant of CHF 15k

Sigma Squared Technologies, Zurich – *Technology Consultant (part-time)* **Jun 2018 – Present**
- Incorporated a sole proprietorship to provide data science, business analytics, and digital consulting services
- Organized a two-day workshop with the C-suite of a EUR XB revenue company on how to build analytics capabilities
- Developed a binary prediction model with 98.7% accuracy and deployed it as a Google Kubernetes Engine API

Lakestar Advisors, Zurich – *PhD VC Intern* **May 2021 – July 2021**
- Supported the investment team in sourcing, evaluating, and executing seed and Series A investments in Europe
- Conducted technology due diligence on the Aleph Alpha deal (raised a total of $27 million, Series A)
- Developed a multi-class prediction model to estimate the overall fit of an investment target with 90.12% AUC

Amazon, Munich – *Business Intelligence Intern* **Jun 2017 – Aug 2017**
Quid Inc, San Francisco – *Product Management Intern* **Jun 2016 – Aug 2016**
Credit Suisse, London – *Investment Banking Spring Week Analyst* **Apr 2015**

POSITIONS OF RESPONSIBILITY

NEO Network, Lausanne – *President Hub Lausanne (part-time)* **Sep 2018 – Aug 2019**
- Launched the Lausanne Hub of NEO Network, a student-run technology think-tank and network
- Organized keynotes with up to 85 participants and speakers from the World Economic Forum, Swisscom, EY, and PwC

University of Oxford, Oxford – *Student Representative* **Oct 2017 – Aug 2018**
UCL Austrian Society, London – *Founder & President* **Apr 2015 – June 2017**
Ski School Lech, Lech, Austria – *Ski and Snowboard Instructor (during winter breaks)* **Dec 2013 – Jan 2017**

AWARDS & ACHIEVEMENTS

EPFL, Lausanne – *Award for exceptional PhD performance* **Sep 2019**
UCL, London – *Dean's List for outstanding academic performance* **Jul 2017**
UCL Union, London – *Societies Colors Commendation for 2 years of dedicated service to university life* **Mar 2016**

RELEVANT SKILLS, LANGUAGES, INTERESTS & INFORMATION

- *Technical Skills*: Proficient in Python, shell scripting, Google Kubernetes Engine, MS Office
- *Design Skills*: Familiar with Agile project management, design thinking, Photoshop, InDesign, Illustrator, and Final Cut
- *Languages*: German (native tongue), English (level C2), French (level B2)
- *Interests*: Skiing (certified Austrian Snowsport Instructor), landscape photography, and cooking
- *Personal information*: born on 18 May 1995, Austrian citizenship, B permit