How to Talk When a Machine is Listening? Corporate Disclosure in the Age of AI*

Sean Cao† Wei Jiang‡ Baozhong Yang§ Alan L. Zhang¶

First draft: August 2020

This draft: June 2022

*The authors have benefited from discussions with Rui Albuquerque, Elizabeth Blankespoor (discussant), Emilio Calvano (discussant), Lauren Cohen (discussant), Will Cong (discussant), Ilia Dichev, Arup Ganguly (discussant), Jillian Grennan, Bing Han, Kathleen Hanley (discussant), Rebecca Hann, Gerard Hoberg (discussant), Byoung-Hyoun Hwang (discussant), Chris Hennessy, Alan Huang (discussant), Bin Ke (discussant), Michael Kimbrough, Leonid Kogan, Augustin Landier (discussant), Tim Loughran (discussant), Song Ma, Ville Rantala (discussant), Max Rohrer (discussant), Gustavo Schwenkler (discussant), Kelly Shue, Suhas Sridharan, Isabel Wang (discussant), Teri Yohn, Gwen Yu, Dexin Zhou, and comments and suggestions from participants in seminars and conferences at Columbia, ECB, EDHEC, Emory, Georgia State, Harvard, London Business School, Maryland, Michigan, Michigan State, Peking University, Renmin University, Stockholm Business School, Toronto, Utah, Washington, the NBER Economics of Artificial Intelligence Conference, the NBER Big Data and Securities Markets Conference, AFA 2022, the Pacific Center for Asset Management, the SOAR Symposium at Singapore Management University, the Third Bergen FinTech Conference at the NHH Norwegian School of Economics, Machine Learning and Business Conference at University of Miami, RCFS Winter Conference, 11th Financial Markets and Corporate Governance Conference, the China FinTech Research Conference 2021, the Adam Smith Workshop 2021, the Conference on Financial Innovation at Stevens Institute of Technology, FIRS 2021, the Cambridge Alternative Finance Sixth Annual Conference, 2021 CAPANA Research Conference, CICF 2021, and NFA 2021.

†J. Mack Robinson College of Business, Georgia State University. Email: scao@gsu.edu

‡Columbia Business School, NBER, and ECGI. Email: wj2006@columbia.edu

§J. Mack Robinson College of Business, Georgia State University. Email: bzyang@gsu.edu

¶College of Business, Florida International University. Email: alan.zhang@fiu.edu

How to Talk When a Machine is Listening? Corporate Disclosure in the Age of AI

ABSTRACT

Growing AI readership (proxied by machine downloads and ownership by AI-equipped investors) motivates firms to prepare filings that are friendlier to machine processing and to mitigate linguistic tones that are unfavorably perceived by algorithms. Loughran and McDonald (2011) and BERT (2018) serve as event studies supporting attribution of the decrease in the measured negative sentiment to increased machine readership. This relationship is stronger among firms with higher benefit (e.g., external financing needs) or lower cost (e.g., litigation risk) of sentiment management. This is the first study exploring the *feedback effect* on corporate disclosure in response to technology.

JEL Classification: D83, G14, G30

Key Words: Machine Learning, AI, Corporate Disclosure, Textual Analysis, Speech Analysis, Feedback Effect

1. Introduction

The annual report (and other regulatory filings) is more than a legal requirement for public companies; it provides an opportunity to communicate financial health, promote the culture and brand, and engage with a full spectrum of stakeholders. How those readers process this wealth of information significantly affects their perception of and, hence, participation in the business. Warren Buffett's annual letters to shareholders in Berkshire Hathaway's annual reports are often considered Corporate American writing at its best. "Be fearful when others are greedy and greedy when others are fearful," Buffett wrote in the 2007 report. "When it's raining gold, reach for a bucket, not a thimble," he added in 2009. That is an entire business philosophy in 23 words.

However, there are many reasons why Buffett's writing is envious but hard to emulate. Added to this list is the evolving potential readership in the age of artificial intelligence (AI). Increasingly, companies realize that the target audience of their mandatory and voluntary disclosures no longer consists solely of human analysts and investors. A substantial amount of buying and selling of shares is triggered by recommendations made by robots and algorithms that process information with machine learning tools and natural language processing kits.[1] Both the technological progress and the sheer volume of disclosures make the trend inevitable.[2] Companies that wish to accomplish the desired outcome of communication and engagement with stakeholders need to adjust how they talk about their finances, brands, and forecasts in the age of AI. In other words, they should heed the unique logic and techniques underlying the rapidly evolving analysis of language and sentiment facilitated by large-scale machine-learning techniques, such as automated computational processes that identify pos-

[1]For example, Kensho (acquired by S&P in 2018 in the largest AI-driven acquisition deal at the time) developed an algorithm named Warren (after Warren Buffett) that provides a simple interface allowing investors to ask complex questions in plain English. Kensho provides answers by searching through millions of market data points. (Source: "Wall Street tech spree: With Kensho acquisition S&P Global makes largest A.I. deal in history," Antoine Gara, *Forbes*, March 6, 2018). A leading hedge fund, the Man Group, has begun to manage substantial portions of its assets using AI and algorithmic trading. (Source: "The massive hedge fund betting on AI," Adam Satariano and Nishant Kumar, *Bloomberg*, September 27, 2017.)

[2]Cohen, Malloy, and Nguyen (2020) document that the length of 10-Ks increases by five times from 2005 to 2017 and that the number of textual changes over previous filings increases by over 12 times.

itive, negative, and neutral opinions in a whole corpus of firm disclosures that is beyond the processing ability of human brains. While the literature is catching up with and guiding investors' rising aptitude to apply machine learning and computational tools to extract qualitative information from disclosures and news, there has not been an analysis exploring the *feedback effect*: how companies adjust the way they talk knowing that machines are listening. This paper fills this void.

Our analysis starts with a diagnostic test that connects how machine-friendly a company composes its disclosures (measured by *Machine Readability* following Allee, DeAngelis, and Moon, 2018) and the expected extent of machine readership for a company's SEC filings on EDGAR, for which we develop multiple proxies. Using historical information, the first variable, *Machine Downloads*, is constructed by tracking IP addresses that conduct downloads in large batches. Machine request is a precursor and a necessary condition for machine reading, and the sheer volume of machine-downloaded documents makes it unlikely for them to be processed by human readers alone. Because machine reading is adopted by technology-sophisticated investors, we also construct a measure based on share ownership by institutional investors with AI capabilities, *AI Ownership*, tracked from their job postings (following Abis and Veldkamp, 2022). Finally, we proxy investor technology capacity by calculating the ownership-weighted *AI Talent Supply* available to institutional investors, based on the state-year-level proportion of the working-age population with IT degrees where the investors are headquartered. Because asset manager headquarters were mostly chosen before the AI era and bear no direct relation to portfolio firms, the last variable is likely to be orthogonal to omitted variables explaining *Machine Readability*.

We show that, in the cross section of filings with firm and year fixed effects, a one standard deviation change in expected machine downloads is associated with a 0.24 standard deviation increase in the *Machine Readability* of the filing. On the other hand, other (non-machine) downloads do not bear a meaningful correlation with machine readability, validating *Machine Downloads* as a proxy for machine readership. The alternative proxies *AI*

Ownership and *AI Talent Supply* bear similar economic and statistical significance. We further validate the economic mechanism underlying our main variables by showing that trades follow more quickly after a filing becomes public when *Machine Downloads* is higher, with even stronger interactive effect with better *Machine Readability.* Such a result demonstrates the real impact of machine processing on information dissemination.

After establishing a positive association between a high AI reader base and machine-friendlier disclosure documents, we next explore how firms manage the "sentiment" and "tone" perceived by machines. It is well documented that corporate disclosures attempt to strike the right sentiment and tone with (human) readers without being explicitly dishonest or overtly noncompliant (Loughran and McDonald, 2011; Kothari, Shu, and Wysocki, 2009). Hence, we expect a similar strategy catering to machine readers. While researchers and practitioners have long relied on the Harvard Psychosociological Dictionary (especially the Harvard-IV-4 TabNeg file) to count and contrast "positive" and "negative" words to construct "sentiment" as perceived by (mostly human) readers, the publication of Loughran and McDonald (2011, "LM" hereafter) presents an instrumental event to test our hypothesis pertaining to machine readers. This is not only because the paper presented a specialized finance dictionary of positive/negative words and words that are informative about prospects and uncertainty, but also because the word lists that came with the paper have served as a leading lexicon for algorithms to sort out sentiments in both the industry and academia.[3] The differences in both the timeline and the context of the new dictionary allow us to trace out the impact of AI readership on sentiment management by corporations.

As a first step, we establish that firms which expect high machine downloads avoid LM-negative words but only post 2011 (the publication year of the LM dictionary). Such a structural change is absent with respect to words deemed negative by the Harvard dictionary. As a result, the difference, *LM – Harvard Sentiment*, follows the same path as

[3]The LM dictionaries have had a far-reaching influence in the academic literature; e.g., see our discussion of the literature using the LM dictionary at the end of the introduction. For examples of industry uses, see "Natural language processing in finance: Shakespeare without the monkeys," Slavi Marinov, Man Group, July 2019, and "NLP in the stock market," Roshan Adusumilli, *Medium*, February 1, 2020

LM Sentiment. For a tighter identification, we further confirm a parallel pre-trend in *LM – Harvard Sentiment* between firms with high and low (top and bottom terciles of) machine downloads up to 2010. Post 2011 saw a clear divergence where the "high" group significantly reduced, relative to the "low" group, the use of negative words from the LM dictionary as opposed to those from the Harvard dictionary. Given the quasi-randomness of the exact timing of publication, the difference-in-differences in the sentiment expression is more likely to be attributable to firms' catering to their AI readers than to an alternative hypothesis that the publication was a side show of a pre-existing and continuing trend.

The documented relation raises intriguing equilibrium implications. If firms can "positify" language without cost and constraint in order to impress machine and human readers, the signals would quickly lose relevance. To remain in an equilibrium in which investors extract information from disclosures, we hypothesize that firms derive and incur heterogeneous benefits and costs from managing sentiment and tone. On the benefit side, we find that firms facing imminent external financing needs are more likely to suppress LM (2011) negative words and to disclose in more machine-readable format so as to ensure that the positive signals are well received. On the cost side, firms facing higher litigation risk are more moderated in their word-mincing.

The rapid evolution of AI technology, even during the writing and revision of this paper, provides "out-of-sample" tests to affirm that the relation we identified off the publication of LM (2011) is not a lone incidence. First, we resort to the emergence of Bidirectional Encoder Representations from Transformers (BERT) developed by Google in 2018 (Devlin, Chang, Lee, and Toutanova, 2018), the state-of-art for machine processing of textual data. We show that BERT-measured negative sentiment drops more post 2018 for firms with higher AI readership, measured by *AI Ownership* and *AI Talent Supply*.[4] Second, we take the study about "how to talk when a machine is listening" literally into the speech setting. Earlier work (Mayew and Ventakachalam, 2012) finds that managers' vocal expressions, as assessed

[4] The data coverage for *Machine Downloads* ended in the first half of 2017.

by vocal analytic software, can convey incremental information valuable to analysts covering the firm. Thus, managers should recognize that their speeches need to impress bots as well as humans. Applying the software to extract two emotional features well-established in the psychology literature, valence and arousal (corresponding to positivity and excitedness of voices), from managerial speeches in conference calls, we find that managers of firms with higher expected machine readership exhibit more positivity and excitement in their vocal tones, echoing the anecdotal evidence that managers increasingly train or even seek professional help to improve their vocal performances along the quantifiable metrics.[5]

Our study builds on an expanding literature on information acquisition and dissemination via SEC-filing downloads,[6] opting for a new angle on the consequences of and human reactions to machine processing. A central theme from the rapidly growing literature on textual analysis is that qualitative information from and the writing quality of disclosures predict asset returns and corporate performance.[7] The computational textual analyses have been steadily advanced by more-modern machine-learning techniques[8] and have been extended to non-text data such as the audio of conference calls (Mayew and Ventakachalam, 2012) and video of startup pitch presentations (Hu and Ma, 2020). Our study departs from the extant literature as we explore managerial disclosure strategies in response to the growing

[5]Sources: "Listening without prejudice: How the experts analyze earnings calls for lies, bluffs, and other flags," Sterling Wong, *Minyanville*, April 18, 2012; and "How to listen for the hidden data in earnings calls," Alina Dizik, *Chicago Booth Review*, May 25, 2017.

[6]Recent studies analyzing downloads of SEC filings include Bernard, Blackburne, and Thornock (2020), Cao, Du, Yang, and Zhang (2021), Chen, Cohen, Gurun, Lou, and Malloy (2020), and Crane, Crotty, and Umar (2021).

[7]Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Hanley and Hoberg (2010) pioneered applying psychological dictionaries to financial texts to give content to sentiments. LM (2011) developed capital-market-specific dictionaries which have since been applied to large-scale computation of tone and sentiment in financial texts, e.g., Dow Jones newswires (Da, Engelberg, and Gao, 2011), New York Times financial articles (Garcia, 2013), 10-K and IPO prospectuses (Jegadeesh and Wu, 2013), corporate press releases (Ahern and Sosyura, 2014), earnings conference calls (Jiang, Lee, Martin, and Zhou, 2019), and wire news from Factiva (Huang, Tan, and Wermers, 2020). Hwang and Kim (2017) directly connect the writing quality of filings to valuation in the context of closed-end funds. See also the survey article Loughran and McDonald (2016).

[8]Applications of more recent techniques in finance research include support vector regressions (Manela and Moreira, 2017), word embedding and latent Dirichlet allocation (Li, Mai, Shen, and Yan, 2020; Hanley and Hoberg, 2019; Cong, Liang and Zhang, 2019), and neural networks (Chen, Wu, and Yang, 2019). See also the survey article Cong, Liang, Yang, and Zhang (2021).

presence of AI analytical tools in both the industry and academia.

Our study thus connects to a distinct literature on the "feedback effect": while the financial markets reflect firm fundamentals, market perception also influences managers' information sets and decision making (see a survey by Bond, Edmans, and Goldstein, 2012). We uncover a novel "feedback effect" of machine learning about firm fundamentals on corporate disclosure decisions in the era of AI. As long as the encoded rules are not completely opaque—and thus are transparent, observable, or reverse-engineerable to at least some degree—agents impacted by machine learning decisions have the incentive to manipulate inputs in order to game a more desirable outcome. Though a relation between evaluation metrics and agent behavior is not new,[9] it is fairly recent that the machine learning community formalizes the matter as one of "strategic classification" (Hardt, Megiddo, Papadimittriou, and Wootters, 2016; Dong, Roth, Schutzman, and Waggoner, 2018; Milli, Miller, Dragan, and Hardt, 2019) and that anecdotal evidence surfaces that companies' investor relations departments resort to algorithmic systems to test draft versions of disclosures for optimal effects.[10] We present the first large-sample empirical evidence of the feedback effect from algorithmic assessment to corporate behavior.[11] While some adaptive behavior, such as making disclosures more machine-reading friendly, is innocuous or even welcome, other algorithm-induced changes, such as the expression of sentiment and tone, highlight the increasing challenge on machine learning to be "manipulation proof" in that the algorithms learn to anticipate the strategic behavior of informed agents without observing it in training samples (see theoretical analyses in Bjorkegren, Blumenstock, and Knight, 2020; Hennessy and Goodhart, 2020).

[9]In their classical work, Goodhart's (1975) Law and Lucas's (1976) Critique generalize the phenomenon in the setting of macro policy interventions.

[10]The circulation of this study as a working paper also has raised awareness. See, e.g., "Sweet-talking CEOs are starting to outsmart the robot analysts," Gregor Stuart Hunter, *Bloomberg*, October 20, 2020; "Robo-surveillance shifts tone of CEO earnings calls," Robin Wigglesworth, *Financial Times*, December 5, 2020; and "Companies are now writing reports tailored for AI readers – and it should worry us," John Naughton, *The Guardian*, December 5, 2020. All these articles featured our research in the context of the new phenomenon.

[11]LM (2011) acknowledged, without providing evidence, the theoretical possibility that "[k]nowing that readers are using a document to evaluate the value of a firm, writers are likely to be circumspect and avoid negative language."

2. Hypothesis Development

The experience of Man Group chief executive, Luke Ellis, provides a fitting motivation to our hypothesis development. Realizing that his speech could be systematically and instantaneously scraped by quant investors with natural language processing tools, Mr. Ellis decided to be coached to avoid certain words and phrases that algorithms could pick up on and thus affect Man's stock price. He was quoted as saying "There's always been a game of cat and mouse, in CEOs trying to be clever in their choice of words. But the machines can pick up a verbal tick that a human might not even realise is a thing."[12] The episode suggests that some firms are adjusting their external communications in order for the right message to be sent to, or the right impression to be made on, a machine audience.

To formalize the hypothesis, we develop a stylized model[13] that connects firm disclosures targeting machine readers to securities trading and pricing. In disclosures, a firm manages two additive terms to the true quality of firm fundamentals. The first is "tone"—a more positive tone, other things equal, elicits a higher perception of firm fundamentals; the second is "noise" seen by machine readers, capturing information lost due to imperfect machine readability. The higher the machine readability, the lower the signal's noise. Due to costly technology, there is increasing marginal cost to reach higher levels of machine readability.

The trading game consists of a "machine trader" (i.e., an AI-equipped speculator who trades on machine-parsed information from the disclosure), a noise trader, and a market maker who sets price according to the Kyle (1985) model (see also Kim and Verrecchia, 1994; Foster and Viswanathan, 1996). The firm's utility is a sum of three terms. The first is increasing in the current stock price, capturing the reality that managerial payoffs or firms' gains from external financing tend to be an increasing function of stock price. The second term captures the cost of manipulating tones in disclosure, which can result in reputation and litigation risk. The last term reflects the costs to maintain a given level of machine

[12]For the full story, see "Robo-surveillance shifts tone of CEO earnings calls," Robin Wigglesworth, *Financial Times*, December 5, 2020.

[13]For details, see Internet Appendix 1.

readability. Such costs could be technology driven. Note that higher machine readability or more precise machine signals lead to more machine-driven trades, which in turn increase the impact of tones on prices. Therefore, under such an objective function, the firm desires, from an initial level, to adopt more positive tones and higher machine readability but is eventually constrained by the costs in mispricing[14] and technology upgrades.

Empirical tests in Section 4 and Section 5 demonstrate these first-order effects. In Section 5.3, we further test the empirical relation between machine-targeted disclosure management and the proxies for costs (e.g., litigation risk). After extending the model to multiple human and machine traders, we show that firms are motivated in maintaining higher levels of tone management and machine readability when the machine traders are more numerous. Our model further shows that stock liquidity (market depth) decreases with the increasing presence of machine readers. The intuition here is that providing machine traders a more accurate signal increases the information asymmetry between the machine traders and the market maker, forcing the latter to increase price sensitivity while trading in order to avoid being taken advantage of by the machine traders. We present the empirical test on this relation in Section 4.3.

3. Data, Variable Construction, and Sample Overview

3.1. Data sources

The primary data source of this study is the Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system and the associated Log File Data Set. Since 1994, the SEC has provided the public with access to securities filings containing value-relevant and market-moving information through its EDGAR system, available through the SEC's website and WRDS SEC Analytics Suite.

While EDGAR is a content archive, its Log File tracks requests and downloads. More

[14]Though firms desire high stock price, they are also constrained in setting investor expectations because a large deviation of price from fundamental value leads to reputation damage and litigation risk.

specifically, it comprises all records of the requests of SEC filings from EDGAR since January 2003 to June 2017. Each observation in the original dataset contains information on the visitor's Internet Protocol (IP) address, timestamp, and the unique accession number of the filing that the visitor downloads. In preprocessing the raw Log File, we exclude requests that land on index pages because such requests do not download actual company filings. We then match the accession number with the SEC master filing index to select all the 10-K and 10-Q filings.[15] This procedure yields a total of 438,752 filings (119,135 10-K and 319,617 10-Q). After matching to CRSP/Compustat, our final sample of raw filings consists of 359,819 filings (90,437 10-K and 269,382 10-Q), filed by 13,763 unique CIKs, between 2003 and 2016.[16]

Needless to say, regulatory filings are one of the venues through which firms communicate to the marketplace. Alternatively, firms can host corporate events such as conference calls, corporate presentations, and non-deal roadshows. Regulatory filings have the advantage that their audience composition is mostly exogenous to firms' own decisions, which is less true in the other settings. For example, managers can invite a selected audience to corporate events, while regulatory filings are open to everyone (Cohen, Lou, and Malloy, 2020). For these considerations, we focus on these two most important SEC filings for public companies.

3.2. Construction of main variables

3.2.1. Proxies for machine readership

Several constructed variables are fundamental to our analyses; we describe those in detail here. The first key variable measures the frequency of machine downloads of corporate filings, which serves as an upper bound as well as a proxy for the presence of "machine readers." Despite the advent of multiple data sources, the SEC EDGAR website remains

[15]We do not include amendments and other variant filings because these documents likely mirror the original filings.

[16]The end point of the sample period was dictated by the fact that the SEC stopped publishing the more recent Log File Data Set after June 2017.

the earliest and most authoritative source for company filings to be publicly released.[17] With the advances in computing power and data availability, some large hedge funds and asset managers have started big-data-driven programs to process and analyze unstructured data including corporate filings and news.[18] Recent academic studies also provide evidence that investment companies rely on machine downloads of EDGAR filings for some of their trading strategies. Crane, Crotty, and Umar (2021) find that hedge funds that employ robotic downloads perform better than those that do not. Cao, Du, Yang, and Zhang (2021) show that machine downloaders exhibit skills in identifying profitable copycat trades from their peers' disclosures.

To measure machine downloads, we identify an IP address downloading more than 50 unique firms' filings on any given date as a machine (i.e., robot) visitor and classify its requests on that day as machine downloads, the same criterion as used by Lee, Ma, and Wang (2015).[19] In addition, we include requests that are attributed to web crawlers in the SEC Log File Data as machine-initiated. All remaining requests are labeled as "other" requests. Finally, we aggregate machine requests and other requests, respectively, for each filing within seven days (i.e., days [0,7]) after it becomes available on EDGAR; the majority of requests occur during this period.

Figure 1 shows an exponential growth of machine downloads since 2003. The number of machine downloads of corporate 10-K and 10-Q filings increased from 360,861 in 2003 to 165,318,719 in 2016.[20] During the same period, machine downloads have also become the predominant force among all EDGAR requests: the number of machine downloads as

[17]There was a multi-year episode of early leakage, which was largely resolved in mid-2015. See Bolandnazar, Jackson, Jiang, and Mitts (2020).

[18]See, e.g., "Cohen's Point72 hires 30 people for big data investing," Simone Foxman, *Bloomberg*, March 10, 2015, and "BlackRock uses big data for big gains," Sarah Max, *Barron's*, December 26, 2015.

[19]Loughran and McDonald (2017) proposed an alternative and more aggressive approach to classify those daily IP addresses having more than 50 requests as robot visitors. Because this approach tends to classify almost all downloads as machine-driven in the most recent years, we resort to the more stringent measure by Lee, Ma, and Wang (2015). We nevertheless present the results using the Loughran and McDonald (2017) classification, which is qualitatively similar, in sensitivity checks.

[20]There are other filings, notably 8-K, that are of strong interest to the market. We do not include 8-K filings mainly because they, unlike 10-K/Qs, do not follow a standard structure, making it difficult to compare readability and writing styles in the cross section.

a fraction of all downloads increased from 39% in 2003 to 78% in 2016. The dip in 2016 appears to be temporary. The fraction recovers to 92% during the first half of 2017—the last time period (but incomplete year) for which the SEC log information is available.

[Insert Figure 1 Here]

The variable *Machine Downloads* measures the propensity of machine downloads of a particular filing using ex ante information only. For a firm's (indexed by i) filing (indexed by j) at time t, *Machine Downloads* is the natural logarithm of the average number of machine downloads of firm i's filings during the $[t-4, t-1]$ quarters (we only include the machine downloads of a historical filing within seven days of posting on EDGAR, as explained earlier).[21] *Other Downloads* (the remainder) and *Total Downloads* (the sum) are constructed analogously.[22] In addition to download activities, firms may learn from a combination of sources (e.g., investor relations) about the audience of their communications. We expect firms to be informed to various degrees about the AI capacity of their institutional shareholders. Thus, we construct two more variables capturing machine readership via the AI capabilities of firms' investors. The first such measure is *AI Ownership*, which is the percentage of shares outstanding held by investment companies with AI capabilities. We classify an investment company as such if it has AI-related job postings in the past five years according to data from Burning Glass, following Abis and Veldkamp (2022). *AI Ownership* is the aggregate ownership measured at the firm level and in the quarter before the firm's current filing.

Both *Machine Downloads* and *AI Ownership* involve choices made by investors; those choices could be jointly determined with firms' disclosure choices. To form a sharper causal inference from investor base to disclosure choices, we construct a third proxy for machine

[21]SEC log files are posted on a quarterly basis with a six-month delay (Chen, Cohen, Gurun, Lou, and Malloy, 2020), and quicker discovery could be made with a Freedom of Information Act (FOIA) request. To err on the safe side that a firm obtains the SEC log file information at time t, we conduct a sensitivity check by repeating the main analyses using *Machine Downloads* constructed with a lag of one year. Results, reported in Table IA.2 in the Internet Appendix, are qualitatively similar.

[22]Further, results using *%Machine Downloads*, defined as the ratio of *Machine Downloads* to *Total Downloads* (without taking the natural logarithm for either variable), are reported in Table IA.1 in the Internet Appendix.

readership, *AI Talent Supply*, based on local AI talent supplies where investors are headquartered, which is mostly exogenous to firms and investors. In the first step, we retrieve the number of people between 18 and 64 with college or graduate school degrees in information technology, scaled by population at the state-year level, using data from Integrated Public Use Microdata Series (IPUMS) surveys.[23] Second, for each firm and during the quarter prior to the current filing, we aggregate *AI Talent Supply* over all states based on the headquarters of the investors, weighted by their ownership. Because the headquarters locations for most investors were determined before the AI era and bear no inputs from the portfolio firms, the resulting *AI Talent Supply* should be exogenous to the omitted variables in firm disclosures.

3.2.2. Machine Readability

The second key variable pertains to the "machine readability" of a 10-K or 10-Q filing, which measures the ease with which a filing can be "understood"—that is, processed and parsed—by an automated program. Recent literature in accounting and finance has studied various concepts of (e.g., Hodge, Kennedy, and Maines, 2004; Blankespoor, 2019; Blankespoor, deHaan, and Marinovic, 2020; Gao and Huang, 2020) and proposed metrics for (Allee, DeAngelis, and Moon, 2018) information processing costs related to either machine or human processing (or both). After reviewing the existing research, we adopt multiple metrics developed in Allee, DeAngelis, and Moon (2018) that we believe to best summarize the important attributes distinctly related to machine readability:[24] (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators);

[23] A recent paper by Jiang, Tang, Xiao, and Yao (2021) provides a detailed description of the data.

[24] We thank the authors of Allee, DeAngelis, and Moon (2018) for sharing these component variables from their paper. We adopt a subset of the measures developed therein as we focus solely on components that matter mostly for machine readability (e.g., whether numbers and tables are parsable) and do not include components that may affect both machine parsing and human understanding (e.g., whether a document is separated into different sections).

(iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters. In our main specification, each attribute is standardized to a Z-score before being averaged to form a single-index *Machine Readability* measure. We present sensitivity checks (and demonstrate robustness) using the first principal component[25] of the five attributes as well as the individual underlying attributes.[26]

[Insert Figure 2 Here]

Figure 2 shows the trend of *Machine Readability* from 2004 to 2015. *Machine Readability* saw steep ascendance till 2008, followed by modest growth before leveling off around 2011. The increasing trend per se is prima facie evidence that companies are not following a fixed template for financial filings, but instead have been adapting the format of their filings to a changing environment.[27]

3.2.3. (Negative) sentiment and tones

The third class of key variables aims at measuring "sentiments," which broadly refer to the use of natural language processing, text analysis, and computational linguistics to systematically identify, extract, and quantify subjective information. Because a primary interest of this study is to contrast the sentiment as perceived by human and machine readers, we resort to two established lexica that guide sentiment classification by the two types of

[25] For details, see Table IA.1 in the Internet Appendix.

[26] Figure IA.1 in the Internet Appendix provides a visualization of the *Machine Readability* variable by showing two sample filings with a low and high score, with explanations of how features of the filings are related to the machine readability scoring.

[27] On April 13, 2009, the SEC released a mandate on "Interactive Data to Improve Financial Reporting" (see `https://www.sec.gov/info/smallbus/secg/interactivedata-secg.htm`) as a regulatory effort in adapting disclosures to machine readers. This mandate applies to financial reports of all companies and was implemented over the period from 2009 to 2011. It requires companies to provide financial statements in an interactive data format using the eXtensible Business Reporting Language (XBRL). The release states describes the primary purpose of the amendments to be "[making] financial information easier for investors to analyze and to assist companies in automating regulatory filings and business information processing."

readers. The first lexicon is the Harvard General Inquirer IV-4 psychological dictionary. This comprehensive dictionary assigns 77 psychological intonations or categories to English words. For each corporate filing, we count the number of words that fall into the "Negative" category and normalize it by the total number of words in the textual part of a 10-K/Q filing with all tags, tables, and exhibits removed. This procedure follows the common practice in the literature, for example, LM (2011) and Cohen, Malloy, and Nguyen (2020). The resulting measure, expressed in percentage points, is termed *Harvard Sentiment*. The average filing in our sample contains four Harvard General Inquirer negative words per 100 words. The second lexicon is developed by LM (2011), who create dictionaries of positive and negative words that are specific to the context of financial documents. We count the number of LM-negative words and scale it by the length of the document. The resulting measure, expressed in percentage points, is the *LM Sentiment*. We consider only the negative sentiment related to both dictionaries because the previous literature, including Tetlock (2007), LM (2011), and Cohen, Malloy, and Nguyen (2020), finds that positive sentiment is not as informative.[28] An average (median) filing uses 1.63 (1.54) LM-negative words in every 100 words. The interquartile range is from 1.19 to 1.98 words per 100 words. Finally, we form the difference, *LM – Harvard Sentiment*, to capture the contrast.

LM (2011)'s list of measures for sentiment goes broader to include litigiousness, uncertainty, weak modal and strong modal words, all in financial contexts. More specifically, *Litigious* is the number of litigation-related words (such as "claimant" and "tort") divided by the length of the document, expressed in percentage points. The other measures are constructed analogously. Uncertainty words capture a general notion of imprecision (such as "approximate" and "contingency"), and weak modal and strong modal words convey levels of confidence (such as "always" and "must" as strong, and "possibly" and "could" as weak). In an average filing, every 100 words contain 0.97 (1.43, 0.52, and 0.30) litigious (uncertainty, weak modal, and strong modal) words. We confirm LM (2011)'s findings that the frequency

[28]Replacing the negative sentiment measure by a net sentiment measure does not change our results qualitatively.

of words in these categories in firm filings is associated with stock market reactions and real outcomes, hence constitute motive for firms to manage the wording that could lead to tone inference.

The emergence of Bidirectional Encoder Representations from Transformers (BERT), a transformer-based machine-learning technique for natural language processing developed by Google in 2018, offers us an additional—and recent—setting to test the same economic mechanism. The BERT model provides an integral treatment of sentences that take into account the meaning, order, and interactions of words. More specifically, we use FinBERT (Yang, Siy UY, and Huang, 2020), a version of BERT trained with financial disclosure data (including 10-K, conference call transcripts, and analyst reports) and thus more tailored to our setting, to classify the sentiment of individual sentences in 10-Ks to be positive or negative. We construct the *BERT Sentiment* measure as the ratio of the number of BERT-negative sentences to the total number of sentences (or total number of words) in 10-K sections.[29]

3.2.4. Vocal emotions

Though the focus of this study rests on 10-K and 10-Q filings, we extend to conference calls between firms and the public. The last set of key variables thus concerns audio quality. We build a web crawler using *Selenium-Python* to obtain the audios of conference calls from 2010 to 2016 from EarningsCast.[30] After matching with CRSP/Compustat, our sample consists of 43,462 audio files from 3,290 unique firms (*gvkey*).

Anecdotal evidence suggests that executives have become aware that their speech patterns and emotions, evaluated by humans or software, impact their assessment by investors

[29]To economize on computation time, we focus on the key 10-K section most relevant to our context: Item 7 ("Management Discussion & Analysis (MD&A)"). We also conduct a sensitivity check which also includes Item 1 ("Business (a description of the company's operation)").

[30]EarningsCast is a commercial aggregator for company earnings calls, calendar feeds, and podcast feeds. Its website is `https://earningscast.com`. *Selenium-Python* is an open-source software package that allows us to program a specific mouse-clicking sequential pattern for a particular website so that we can automate web browsing and internet data retrieval from the website; see `https://selenium-python.readthedocs.io`.

and analysts.[31] A pioneer academic study by Mayew and Ventakachalam (2012) finds that when analysts make stock recommendations, they incorporate managers' emotions during conference calls. One of the most prominent models of emotion, the circumplex model, originally developed by Russell (1980), suggests that emotions are distributed in a two-dimensional space defined by valence and arousal. Following Hu and Ma (2020), we rely on a pretrained Python machine learning package *pyAudioAnalysis*[32] (Giannakopoulos, 2015) to code the vocal emotion of each conference call. *Emotion Valence* describes the extent to which an emotion is positive or negative, with a larger value indicating greater positivity. *Emotion Arousal* refers to the intensity or strength of the associated emotion state; a greater (lower) value suggests that the speaker is more excited (calmer). Both measures are bounded between –1 and 1.

3.2.5. Firm characteristics

As usual, the firm characteristics variables (serving as control variables) are retrieved or based on information from standard databases accessed via WRDS, such as CRSP/Compustat and Thomson Reuters Ownership Database. In this category of variables, *Size* is the market capitalization in the natural logarithm. *Tobin's Q* is the natural logarithm of the ratio of the sum of market value of equity and book value of debt to the sum of book value of equity and book value of debt. *ROA* is the ratio of EBITDA to assets. *Leverage* is the ratio of total debt to assets at book value. *Growth* is the average sales growth of the past three years. *IndAdjRet* is the monthly average SIC three-digit industry-adjusted stock returns over the past year. *InstOwnership* is the ratio of the total shares of institutional ownership to shares outstanding. *Analyst* is the natural log of one plus the number of IBES analysts covering the stock. *IdioVol* is the annualized idiosyncratic volatility (using daily data) from

[31] See, e.g., "Can executives' speech patterns provide a good investment guide?" Katherine Heires, *Institutional Investor*, March 22, 2012, and "Listening without prejudice: How the experts analyze earnings calls for lies, bluffs, and other flags", Sterling Wong, *Minyanville*, April 18, 2012.

[32] The open-source *pyAudioAnalysis* is available at https://github.com/tyiannak/pyAudioAnalysis.

the Fama-French three-factor model. *Turnover* is the monthly average of the ratio of trading volume to shares outstanding. *Segment* is the number of business segments and measures the complexity of business operations, following Cohen and Lou (2012). All control variables are constructed annually using information available at the previous year-end. All potentially unbounded variables are winsorized at the 1% extremes.

[Insert Table 1 Here]

Appendix A hosts the definitions of all variables, and Table 1 reports summary statistics. Because some variables require historical information, the sample for our regression analyses starts in 2004 and consists of a total of 324,607 filings (81,075 10-K and 243,532 10-Q).

4. AI Readership and Machine Readability of Disclosures

4.1. Validating Machine Downloads as proxy for AI readership

Our analyses critically depend on *Machine Downloads* being an effective proxy for the presence of AI readership. We thus conduct two tests that support the validity of this key empirical proxy. First, tracing the downloads to the identities of the downloaders would help ascertain that the large-batch downloads are indeed a likely precursor for machine processing. To this end, we use the ARIN Whois database to manually match the IP addresses that have the highest volumes of machine downloads to the universe of investors who ever appear as a 13F filer in the Thomson Reuters 13F database during the sample period. Table 2 reports the identities of the top 20 machine downloaders and the types of institutions they are. Half of the top ten on the list are prominent quantitative hedge funds: Renaissance Technologies, Two Sigma, Point 72, Citadel, and D.E. Shaw. This revelation confirms the anecdotal evidence that quant funds are major players in integrating big data and unstructured data analyses in making investment decisions. The remaining institutions are mostly brokers and investment banks with significant asset management businesses.

[Insert Table 2 Here]

Second, we connect *Machine Downloads* to its primary suspect, hedge funds that adopt AI strategies. Following Guo and Shi (2020), we classify a hedge fund to be AI-prone if at least one employee has been involved in AI projects based on their LinkedIn profiles.[33] We then define *AI Hedge Fund* to be the percentage of shares outstanding that is held by such hedge funds at the firm-quarter level, based on the 13F filings via the Thomson Reuters Ownership Database. We find that *AI Hedge Fund* significantly (at the 5% level) predicts *Machine Downloads* inclusive of all the control variables introduced in Section 3.2.5.[34]

4.2. Relation between Machine Downloads and Machine Readability

As more and more investors use AI tools such as natural language processing and sentiment analyses, we hypothesize that companies adjust the way they talk in order to communicate effectively to readers what they put in the reports. A diagnostic test is thus to relate *Machine Readability* to *Machine Downloads* in the cross section and over time. The first four columns of Table 3 report the results from the following regression at the filing level, indexed by firm(i)-filing(j)-date(t), with both year and firm (or industry) fixed effects, in addition to the slew of control variables ($Control$, as introduced in Section 3.2.5):[35]

$$Machine\ Readability_{i,j,t} = \beta Machine\ Readership_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \gamma Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \epsilon_{i,j,t} \quad (1)$$

[Insert Table 3 Here]

[33]We thank Xuxi Guo and Zhen Shi for sharing the data of hedge funds with AI-experienced employees. AI projects are identified based on both job title and descriptions of experience/responsibility.

[34]For detailed results, please see the last two columns of Table IA.3 in the Internet Appendix.

[35]Table IA.3 in the Internet Appendix reports regressions for the determinants of *Machine Downloads*. Results show that machine downloads tend to be higher for large firms with more firm-specific developments (e.g., high trading turnover, high idiosyncratic volatility). Because our research question concerns the consequence of machine readership, the magnitude of machine downloads (instead of the percentage) is the more pertinent metric and hence our default measure.

In Table 3 Panel A, *Machine Downloads* serves as the proxy for machine readership. It shows that the expected machine downloads for company's filing, whether measured as the volume or percentage of machine downloads, significantly (at the 1% level) and positively predicts machine-reading friendliness across all specifications. With the standard deviations of *Machine Downloads* and *Machine Readability* being 1.763 and 0.584 respectively (see Table 1), the first four columns show that a one standard deviation increase in *Machine Downloads* is associated with a 0.18 to 0.24 standard deviation increase in *Machine Readability*. The effects are almost invariant with or without the control variables, indicating that other firm characteristics have little confounding effect.[36] Presumably, non-machine downloads could serve as a natural placebo test. Indeed, all four coefficients on *Other Downloads* (Columns (1) to (4)) turn out to be indistinguishable from zero, economically and statistically.

In reality, firms are unlikely to manage the level of machine readability of their disclosures back and forth from year to year. Instead, increasing machine readability is usually an outcome of a technology upgrade which firms conduct every once in a while when they observe the rise of machine readership of their published filings. To capture such a mechanism, we present a new machine readability *upgrade* analysis based on intertemporal differencing (instead of firm fixed effects). More specifically, we define an "upgrade" event at the filing (i, j, t) level if $Machine\ Readability_{i,j,t}$ incurs a significant (i.e., one standard deviation of the full sample) increase over the previous year's $Machine\ Readability_{i,t-1}$. We then regress the indicator variable $MR\ Upgrade_{i,j,t}$ on lagged changes in *Machine Downloads* from $t-2$ to $t-1$, $\Delta Machine\ Downloads_{i,t-1}$.

Results are reported in the last two columns of Panel A of Table 3. We show that past growth in machine downloads is a significant predictor of machine readability upgrades. Such a dynamic upgrading analysis affords a byproduct of tighter causal identification: While a regression with firm fixed effects (Columns (2) and (4)) helps with identification when endo-

[36]To ensure that the intertemporal persistence of *Machine Downloads* is not impacting statistical inference, we adopt the Driscoll and Kraay (1998) standard errors to account for serial dependence. In addition, we present a sensitivity check for standard errors double clustered by industry and time. Results, reported in Table IA.4 in the Internet Appendix, are robust.

geneity due to firm-level heterogeneity is time-invariant, the intertemporal differencing (i.e., $MR\ Upgrade_{i,j,t}$ and $\Delta Machine\,Downloads_{i,t-1}$) relaxes the assumption such that the unobserved firm-level heterogeneity is only required to be stable during the differencing window, or two years, which is plausible. Moreover, this specification also mitigates the concerns for the intertemporal persistence of the key independent variable *Machine Downloads* in levels, because the upgrades do not exhibit persistence in our sample.

Panel B of Table 3 breaks down *Machine Readability* into its five components: *Table Extraction*, *Number Extraction*, *Table Format*, *Self-Containedness*, and *Standard Characters*. Results show that high expected machine downloads increase all five sub-metrics of machine readability significantly (at the 1% level). Again, the coefficients of *Other Downloads* do not have consistent signs across the five attributes.

Panel C of Table 3 examines the relation between machine readability with the two alternative measures for machine readership. *AI Ownership* is the percentage of shares outstanding of a given firm during the quarter before the filing that are owned by "AI-equipped" institutional investors based on their job postings. *AI Talent Supply* is the state-level information technology talent (as percentage of population) aggregated at the firm level based on the headquarter locations of its investors. Both variables are described in Section 3.2.1). Results show that a one standard deviation increase in *AI Ownership* (*AI Talent Supply*) is associated with a 0.04 (0.12) standard deviation increase in *Machine Readability* (all significant at the 5% level). The consistent relation using all machine-readership proxies provides confidence in the inferences. Moreover, the results associated with *AI Talent Supply* are particularly helpful for causal inferences, as state-level AI talent supply where firms' investors are headquartered, mostly decided before the AI era, is likely to be exogenous to any omitted variables in the regression.

Finally, a strand of accounting literature documents that sometimes firms may want to downplay bad news with obfuscated language (Asay, Libby, and Rennekamp, 2018). To demonstrate a consistent incentive, we verify that there is a correlation between linguistic obfuscation and complexity (Loughran and McDonald, 2014; Kim, Wang, and Zhang, 2019) and low *Machine Readability*, which could be interpreted as technical/formatting obfuscation; moreover, firms exhibiting greater linguistic complexity are less likely to have an upgrade in machine readability.[37]

4.3. The effect of Machine Downloads and Machine Readability on trading and information dissemination

The primary advantage machines enjoy is their capacity and information processing speed. When disclosures are read more by machines, and when filings are made more machine readable, we hypothesize that trades motivated by the information in the disclosures should materialize faster and the speed of information dissemination should be faster. The testing of such a hypothesis is operationalized into a duration analysis connecting "time to trade" and "time to quote change" to the key independent variables. Using high-frequency data in NYSE Trade and Quote (TAQ) databases, we first conduct the following regression at the filing level, indexed by firm(i)-filing(j)-date(t), with year and firm (or industry) fixed effects:

$$
\begin{aligned}
Time\,to\,Trade_{i,j,t} = {} & \beta_1 Machine\ Downloads_{i,j,t} \times Machine\ Readability_{i,j,t} \\
& + \beta_2 Machine\ Downloads_{i,j,t} + \beta_3 Machine\ Readability_{i,j,t} \\
& + \delta Other\ Downloads_{i,j,t} + \gamma Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \epsilon_{i,j,t}
\end{aligned} \tag{2}
$$

There are two versions for the dependent variable: *Time to the First Trade* and *Time to the First Directional Trade*, the construction of which follows Bolandnazar, Jackson, Jiang, and Mitts (2020). *Time to the First Trade* is the length of time, in seconds, between the time stamps of the EDGAR posting and the first subsequent trade of the issuer's stock. *Time to the First Directional Trade* adds a requirement that the trade needs to be profitable

[37]For details, see Table IA.5 in the Internet Appendix.

(before any transaction cost) based on the price at the end of the 15^{th} minute post filing. That is, the first directional trade is the first buy (sell) trade at a price below (above) the "terminal value," where buy- and sell-initiated trades are classified by the Lee and Ready (1991) algorithm. As in Bolandnazar, Jackson, Jiang, and Mitts (2020), we focus on the 15-minute window in order to isolate the effect of the filing; hence, the duration variables are censored at the end of the time window.

The results, reported in Table 4 Panel A, support the prediction that high *Machine Downloads* are associated with faster trades after a filing becomes publicly available. A one standard deviation increase in *Machine Downloads* saves 8.6 to 14.7 seconds for the first trade and 13.3 to 21.8 seconds for the first directional trade. All coefficients associated with directional trades (in the last four columns) are significant at the 1% level, while the coefficients lose significance with *Time to the First Trade* when firm fixed effects are included. Moreover, the relation between *Machine Downloads* and the Time to Trade variables is indeed significantly stronger when *Machine Readability* is higher.

[Insert Table 4 Here]

In addition to trades, we examine how *Machine Downloads* affects the quote changes around filings, a more direct test for information dissemination. We define a directional quote change as an increase (decrease) in the ask (bid) price if the price at the end of the 15^{th} minute post filing is higher (lower) than the latest price prior to filing. We then replace the dependent variable in Equation (2) to be *Time to the First Directional Quote Change*, classified as the first increase in ask price upon favorable news or the first decrease in the bid price upon unfavorable news, where the direction is determined by stock price 15 minutes post filing. We find similar but statistically weaker results.[38]

[38]Detailed results are reported in Table IA.6 in the Internet Appendix. It is worth noting that the relation we study herein is different from the setting in Allee, DeAngelis, and Moon (2018), which combines humans' and machines' information processing costs. We make more strict empirical choices to focus on machine readability. Such a difference could explain why Allee, DeAngelis, and Moon (2018) show limited evidence on the speed of news dissemination.

While the previous tests suggest that machines speed up information dissemination, it remains unknown whether such a change improves or dampens liquidity. The theoretical literature on disclosure overall concludes that disclosure quality generally increases liquidity and, as a result, reduces cost of capital for the disclosing firms (e.g., Diamond and Verrecchia, 1991; Verrecchia, 2001; Easley and O'Hara, 2004; Balakrishnan, Billings, Kelly, and Ljungqvist, 2014, and review by Goldstein and Yang, 2017). Machine readability effectively enhances the disclosure quality, but only for a subset of readers. Hence the liquidity effect is *a priori* not clear when investors are a mix of those with and without AI tools. Moreover, when firms provide information in a way that allows certain traders—in this case, machine-equipped investors—to make judgments about a firm's fundamentals more efficiently than others, information asymmetry worsens (Kim and Verrecchia, 1994 and 1997).

Following the common practice in the market microstructure literature, we test the impact of machine readers on information asymmetry and hence trading liquidity by exploring the bid-ask spread before and after a filing. Specifically, we conduct the following regression at the firm(i)-filing(j)-minute(m) level with both filing and minute fixed effects:

$$
\begin{aligned}
\textit{Bid-Ask Spread}_{i,j,m} = {} & \beta \textit{Machine Downloads}_{i,j} \times \textit{After}_{i,j,m} + \\
& \gamma \textit{Machine Readability}_{i,j} \times \textit{After}_{i,j,m} + \alpha_{i,j} + \alpha_m + \epsilon_{i,j,m}.
\end{aligned} \tag{3}
$$

The samples cover from 15 minutes before each filing to 15 minutes afterwards. The dependent variable, *Bid-Ask Spread*, is constructed using the latest pair of lowest ask price and highest bid price within each minute following the National Best Bid and Offer (NBBO) rule, and is scaled by the midpoint of the bid price and ask price. *After* is a dummy variable equal to one if minute m occurs after the filing is posted. When both filing ($\alpha_{i,j}$) and minute-level time (α_m) fixed effects are included, the single-variable terms (including *Machine Downloads* and *Machine Readability*) and the control variables are all subsumed because firm characteristics do not change during the 30-minute window.

The most important coefficient from the results, reported in Table 4 Panel B, is the coefficient associated with *Machine Downloads* × *After*. Panel B shows that *Bid-Ask Spread* widens more for filings with higher expected *Machine Downloads* after filings become publicly available. The coefficient is significant at the 1% level across all specifications. From the result in Column (2), the incremental increase in the spread associated with a one standard deviation increase of *Machine Downloads* amounts to 14 basis points, or about 19% (3.3%) of the median (average) spread in our sample. However, files that score higher on *Machine Readability* do not experience significant spread expansion post filing, despite positive coefficients on *Machine Readability* × *After*.

Because firm characteristics variables are subsumed by high-dimensional fixed effects, we explore the cross-sectional effects by sorting firms into two subsamples by the median value of *Turnover* (defined in Appendix A), an important variable characterizing a firm's trading environment. Results in the last two columns of Table 4 Panel B show that the trading environment has little impact on the relation between *Machine Downloads* and *Bid-Ask Spread*. The two coefficients are not materially different from each other, economically or statistically.

The overall evidence is consistent with the prediction that machine-equipped (hence quicker-informed) investors are able to update their judgments about a firm's fundamentals more efficiently than others, which worsens information asymmetry.

5. Managing Sentiment and Tone with Machine Readers

5.1. Textual sentiment

While truthfulness in disclosure reports is expected and required, managers usually want to portray their business activities and prospects in a positive light to attract or gain from stakeholders (creditors, employees, suppliers, and customers). Earlier literature has quantified the information content from sentiment by counting positive and negative words in corporate reports, based on respectable lexicons such as the Harvard Psychosociological Dictionary, specifically, the Harvard-IV-4 TagNeg (H4N) file. Such word lists were originally developed for human readers and for general purposes, and over time they have come to serve as an objective standard for researchers to analyze the sources and consequences of tone and sentiment, as perceived by the general readership, in corporate disclosures and new media (Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Hanley and Hoberg, 2010). However, the meaning and tone of English words are highly context- and discipline-specific, and a general word categorization scheme might not translate effectively into a specialized field such as finance. This motivated the influential work by LM (2011), which presented a specialized dictionary of positive and negative words that fits the unique text of financial situations. According to LM (2011), almost three-fourth of the words identified by the Harvard dictionary as negative (such as "liability") are words typically not considered negative in financial contexts. The LM (2011) dictionary has since become the leading lexicon used in algorithms for sentiment calibration.[39]

The timeline of Harvard General Inquirer dictionary (existing since 1996) and the Loughran-McDonald dictionary (since 2011)[40] and their differential adoption by human versus machine readers provide a unique setting for us to test how the writing of corporate filings adjusts to AI readers. We consider the following regression at the filing level, indexed by firm(i)-filing(j)-date(t), with year and firm (or industry) fixed effects:

$$Negative\,Sentiment_{i,j,t} = \beta_1 Machine\,Downloads_{i,j,t} \times Post_t + \beta_2 Machine\,Downloads_{i,j,t} + \delta Other\,Downloads_{i,j,t} + \gamma Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \epsilon_{i,j,t} \quad (4)$$

[39]For example, as of April 2022, the LM paper has been cited more than 3,700 times by researchers. And their word list has been adopted for the WRDS SEC Sentiment Data. The dictionary has been frequently featured in industry white papers and technical reports, such as in "Natural language processing in finance: Shakespeare without the monkeys" by the Man Group in July 2019.

[40]The paper was in public distribution, e.g., posted on the SSRN, since 2009. Google citation counts show that LM (2011) was cited 10 times prior to 2011 and 243 times by 2013; citations have grown exponentially to 2,716 times as of January 2021.

There are three versions of the dependent variable *Negative Sentiment* in the equation above: the *LM Sentiment*, the *Harvard Sentiment*, and their difference *LM – Harvard Sentiment*, as defined in Section 3.2.3. We only consider the prevalence of negative words because earlier research (Tetlock, 2007; LM, 2011; Cohen, Lou, and Malloy, 2020) indicates that positive words are not informative of firm future outcomes or stock returns. *Post* is an indicator variable for years that came after the publication of LM (2011) and is equal to one for filings in 2012 and onwards, and zero otherwise. Filings in 2011 are excluded from the analysis. The year fixed effect subsumes the variable *Post* on its own.

Under the hypothesis that AI readers employed by algorithmic investors shape the style and quality of corporate writing, we expect the difference-in-differences coefficient β_1 to be significantly negative for *LM Sentiment* but not for *Harvard Sentiment*. That is, there should be a differential relation between *LM Sentiment* and *Machine Downloads* during the *Post* period (after the publication of LM (2011)) relative to before, but a similar change around 2011 should be absent for *Harvard Sentiment*. Such an exclusive set of effects is confirmed by results in Table 5.

[Insert Table 5 Here]

Table 5 shows an unambiguous contrast before and after 2011, the year when the paper was published, on the effect of measures related to LM (2011). Post 2011, a one standard deviation increase in *Machine Downloads* is associated with a 9 to 11 basis point incremental decrease in *LM Sentiment*, on top of an insignificant (Column (3) with industry fixed effect) or much smaller (Column (4) with firm fixed effects) effect during the pre-2011 period. The incremental effect post 2011, significant at the 1% level, represents about 5% of the sample mean of *LM Sentiment*, or 0.15 standard deviations. In contrast, the coefficient on *Harvard Sentiment* is positive in both Columns (5) and (6), and even statistically significant in Column (6) with refined firm fixed effects. This evidence is suggestive of a substitution effect; that is, managers use negative words from the LM list in place of synonyms from the Harvard list. Finally, Columns (1) and (2) show that the relation between *LM – Harvard*

Sentiment and *Machine Downloads* conforms to that of *LM Sentiment*, confirming that the differential effect is mainly driven by reduced *LM Sentiment*.

Results in Table 5 keep the possibility open that the publication of LM (2011) merely reflects a general trend of a strengthening relation between the machine downloads and avoiding using words that are perceived to have negative connotations in the finance context. Such a possibility still supports the general thesis that machine readership impacts disclosure quality; nevertheless, a "parallel pre-trend" would allow a sharper identification on the impact of a new lexicon available to machine reading. Figure 3 illustrates the structural break, instead of a pre-existing and continuing trend, around 2011. More specifically, we aggregate the *LM – Harvard Sentiment* at the annual level separately for filings that are in the top and bottom terciles of *Machine Downloads* in each year. Figure 3 plots the time series of the incremental tendency to use LM-negative words over Harvard-negative words by the two groups of filings.

[Insert Figure 3 Here]

Figure 3 shows a parallel pre-trend of the two groups till 2011 and then a clear divergence afterward. Before 2011, filings in the top and bottom terciles of *Machine Downloads* exhibit clustered movements in the *LM – Harvard Sentiment*. Afterwards, the top tercile's sentiment trends down relative to that of the bottom tercile. We note a general trend, among all firms, to use fewer negative words in disclosures, which may reflect a growing awareness among firms of the perception induced by linguistic sentiments after the first generation of textual research. After the LM (2011) list was published, clearer and more practical guidance became available. Figure 3 suggests that firms with high machine readership were more motivated to avoid negative words that could feed into machine reading, leading to the divergence.

Given the quasi-randomness of the event year 2011 due to the long and unpredictable time period for finance research to appear in print,[41] it is unlikely that the publication of

[41] A recent paper by Dai, Donohue, Drechsler, and Jiang (2022) shows that the typical eventually published finance paper takes about three years to come to publication fruition, with a standard deviation of 1.8 years.

LM (2011) perfectly timed a structural break in the tone management by corporations that would have materialized in the paper's absence. In other words, it is implausible that the LM dictionary summarizes the practice that was already in place, and that it serves as a coincidentally concurrent sideshow. Table 5 and Figure 3 thus provide more support to the hypothesis that corporate writing has been adjusted to serve machine readers and this shift was impacted by the availability of the LM dictionary.

Given the aggregate evidence that firms avoid words that are likely to be classified as negative by algorithms, we are curious to further uncover which words have become the least welcome. Out of all words classified as negative by the LM dictionary but not the Harvard dictionary, we are able to compare the frequencies they appear in filings pre- (2004–2010) and post-2011 (2012–2016). Sorted by the reduction in the average frequency per filing, the ten most avoided words are: "restructuring," "termination," "restatement," "declined," "correction," "misstatement," "terminated," "late," "alleged," and "omitted." The reduction amounts to 0.15 times to 0.35 times per filing. Sorted by the percentage reduction, that is, the reduction in frequency scaled by the frequency in the pre-2011 period,[42] the ten most avoided words are "restatement," "declined," "misstatement," "closure," "late," "dismissed," "inquiry," "alleged," "omitted," and "restructuring." The reduction in these words amounts to 10% to 35%.[43]

5.2. Managing other textual tones with machine readers

In addition to providing lists of sentiment-related words, LM (2011) also constructs lists of "tone" words, tailored to the financial context, aiming to capture litigiousness, uncer-

[42]Some words which show up infrequently before 2011 but never appear after 2011 would have a percentage reduction of -100%. We only consider words with an average frequency per filing of no less than 0.5 times.

[43]To further address the concern that reduction in the LM-negative words could be part of the general trend that firms experienced fewer negative events coming out of the Financial Crisis since 2010, we present a comparison for pre- and post-2011 of major "negative" events, forced CEO turnover (the construction of which follows Peters and Wagner (2014) and is obtained from Florian Peters' website), restatement, and restructuring. If we add these events (as indicator variables) as control variables, results remain robust (see Table IA.7 in the Internet Appendix).

tainty, and weak and strong modality. The expanded dictionary allows machines to assess more dimensions of a document's connotations. LM (2011) discovers that the stock market responds less positively to disclosures using more negative, uncertain, strong modal, and weak modal words, and that firms with a high proportion of negative or strong modal words are more likely to report material weakness. Given the market reaction, it is reasonable to expect managers to adjust tone along these dimensions after the methodology became publicly known. We re-estimate Equation (4) by replacing the dependent variable with *Litigious*, *Uncertainty*, *Weak Modal*, and *Strong Modal*, which are all defined in Section 3.2.3 as well as in Appendix A:

$$Tone_{i,j,t} = \beta_1 Machine\ Downloads_{i,j,t} \times Post_t + \beta_2 Machine\ Downloads_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \gamma Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \epsilon_{i,j,t} \quad (5)$$

If managers have adjusted the frequency of LM-negative words based on their knowledge about investor reaction to sentiment, they should then be expected to also understand the impact of other tones documented in LM (2011). Given LM (2011)'s discovery that the frequency of all four tones were met with negative stock market reactions, we conjecture that managers of firms with high expected machine readership should moderate these words after 2011. Results in Table 6 support such a prediction. The coefficients associated with *Machine Downloads* × *Post* are significant (at 5% level or less) for all four dependent variables. That is, post-2011 corporate reports expecting more machine readers are more likely to avoid conveying a sentiment, as evaluated by an algorithm, that is predictive of legal liabilities, that is indicative of uncertain prospects, or that exhibits too little or too much confidence and surety. Taking the coefficient from Column (2), a one standard deviation increase in *Machine Downloads* predicts a 0.19 standard deviation decrease in the *Litigious* tone.

[Insert Table 6 Here]

5.3. *Equilibrium and cross-sectional effects*

The empirical findings in the previous sections generate intriguing equilibrium implications. In order for corporate disclosures to remain informative to investors in equilibrium, the language used must be, to some extent, constrained to honesty and transparency. If firms can "positify" language unlimitedly in order to impress machine and human readers, the signals would quickly lose relevance, resulting in a babbling equilibrium (Crawford and Sobel, 1982).[44] To remain in an equilibrium in which investors extract information from disclosures, we hypothesize that firms face heterogeneous costs and derive heterogeneous benefits when deviating from truthful and transparent language.

We test the hypothesis in two cross-sectional settings. The first test explores motives underlying positive disclosures by sorting firms by upcoming external financing needs, defined as the net total issuance in a given year in excess of that in last year. The net total issuance is calculated as the sum of the net debt issuance (change in current and long-term debt) and the net equity issuance, scaled by book assets. We single out firms that fall into the top quartile of external financing needs and compare them with the rest of the sample firms. Results in Columns (1) and (2) of Table 7 show that firms facing high external financing needs, which presumably present greater incentives to convey clear and positive communications to investors, are indeed more likely to increase machine readability. They are also more likely to economize on words that would be perceived negatively by textual analyzers (Columns (3) and (4)).

The second test builds on the premise that firms under tighter regulatory scrutiny or higher litigation risk are more constrained in mincing words. To sort on litigation risk, we follow Bertomeu, Cheynel, Floyd, and Pan (2021), which developed a measure of machine-learning-predicted probability of litigation at the industry level using a broad set of variables

[44] Indeed, we find that the return predictability based on LM sentiment diminishes after 2011, consistent with an evolving "cheap talk" effect. However, such diminishing returns are also commonly associated with publication of return predictability based on publicly observable signals (McLean and Pontiff, 2016).

capturing accounting, capital markets, governance, and auditing conditions.[45] Based on the predicted probability, we classify firms in the top-quartile industries as embodying high litigation risk, while the rest of the firms serve as controls. Columns (5) and (6) in Table 7 show that the reduction in the use of negative words after 2011 is significantly less pronounced among high litigation risk firms, presumably because such firms are more constrained in manipulating language in disclosures.

[Insert Table 7 Here]

6. Out-of-Sample Tests: Recent Technology and Audio Tone

Despite the extensive tests conducted based on LM (2011), we have results based on a single event. Fortunately, the rapid evolution of AI technology provides us "out-of-sample tests" to support the inferences developed in earlier sections. This section explores disclosure adaptation to newer natural language processing technology and AI audio analyzers.

6.1. Managing sentiment in response to recent technology (BERT)

In the first test, we study managerial disclosure adaptation to the Bidirectional Encoder Representations from Transformers (BERT), the current state-of-art for machine processing of text data. BERT was introduced in 2018 by a group of researchers at Google (Devlin, Chang, Lee, and Toutanova, 2018), who also open-sourced the associated codes and model. BERT considers the sequential relations of words inside sentences and produces superior results in understanding the meanings of sentences.

Because the EDGAR Log File Data Set stopped in 2017 and BERT was published in 2018, our *Machine Downloads* variable is not available for this test. Instead, we resort to *AI Ownership* and *AI Talent Supply* developed in Section 3.2.1 as the key independent variables, both of which are proxies for the percentage of firm stocks held by investment companies

[45] We gather the data from Jeremy Bertomeu's website.

with high potential for AI capabilities. The coverage of our key independent variables ends in 2019, hence we focus on a relatively close window, between 2016 and 2019, around the publication of BERT. We consider the following regression at the firm-year level, indexed by firm(i)-year(t), with year and firm fixed effects:

$$BERT\ Sentiment_{i,t} = \beta AI\ Readership_{i,t} \times Post\text{-}BERT_t + \delta AI\ Readership_{i,t} + \gamma Control_{i,t} + \alpha_i + \alpha_t + \epsilon_{i,t} \quad (6)$$

The dependent variable, *BERT Sentiment*, is the ratio of the number of negative sentences (based on BERT) to the total number of sentences in the key 10-K section most relevant to our context: Item 7 ("Management Discussion & Analysis (MD&A)"). That section is considered to be the focal place where management provides investors with its view of the financial performance and condition of the company.[46] The key independent variable *AI Readership* is either *AI Ownership* or *AI Talent Supply*. In a difference-in-differences setting, reported in Table 8, we find that firms with higher *AI Ownership* or *AI Talent Supply* reduce the representation of negative sentences significantly, relative to firms with lower AI-equipped investors, after the introduction of BERT in 2018.

[Insert Table 8 Here]

6.2. *Managing audio quality in conference calls with machine readers*

Though the textual quality of disclosures is this study's focus, voice analytics, enabled by the development of modern machine-learning methods, provides an out-of-sample test for our hypothesis that corporate disclosure caters to machines. Starting around 2008, voice analytic software, such as the commercial Layered Voice Analysis (LVA) software and open-source

[46] It is a common practice for researchers to focus on this item of 10-K for textual analysis so as to optimize on the ratio of informative disclosure to boilerplate language, as well as economizing on computation time. See, e.g., Loughran and McDonald (2011) and Cohen, Malloy, and Nguyen (2020). We conduct a sensitivity check which also includes Item 1 ("Business (a description of the company's operations)"). The results, reported in Table IA.8 in the Internet Appendix, are indistinguishable from those in the main specification.

software on GitHub, have gained attention among investors looking for an edge in information processing. Such software has enabled researchers to study managers' vocal expressions and their implications on capital markets (Mayew and Ventakachalam, 2012; Hu and Ma, 2020). If managers are aware that their disclosure documents could be parsed by machines, they should have realized that their machine readers may also be using voice analyzers to extract signals from vocal patterns and emotions contained in managers' speeches.

This section explores whether management adjusts the way they talk (on conference calls) when they expect that machines are listening, based on a sample of audio data of earnings-related conference calls from 2010 to 2016, as described in Section 3.2.4. The choice of the sample is motivated by two factors. First, conference calls are staged events that allow firms to interact with stock analysts and institutional investors. Importantly, Huang and Wermers (2020) find that institutional investors significantly react to the tone of calls in their trades and holdings of stocks, and hence these calls should be the right venue to test any feedback effect. Second, vocal tones are inevitably affected by fundamentals: managers are more likely to exhibit positivity and excitement when firm fundamentals are strong and outlooks bright. By analyzing earnings calls, we can control for the underlying fundamentals by including earnings surprise in the regressions.

Since there are no data on downloads of conference calls, we keep *Machine Downloads* of a firm's filings as the proxy for the prevalence of "machine listeners," based on the premise that *Machine Downloads* represents investors' propensity to deploy AI tools in analyzing corporate disclosures. Table 9 reports the results from the following regression at the conference call level, indexed by firm(i)-call(k)-date(t), with year and firm (or industry) fixed effects:

$$\begin{aligned} Emotion_{i,k,t} = \beta Machine\ Downloads_{i,k,t} + \delta Other\ Downloads_{i,k,t} \\ + \gamma Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \epsilon_{i,k,t} \end{aligned} \quad (7)$$

We measure emotion along two dimensions developed in psychology, *Valence* and *Arousal*, that capture positivity and intensity of vocal tones, respectively (Russell, 1980).

[Insert Table 9 Here]

The first four columns of Table 9 show that higher *Machine Downloads* is associated with higher *Valence*, or positivity in vocal emotion. A one standard deviation increase in *Machine Downloads* is associated with a 0.28 standard deviation higher *Valence*. The last four columns of Table 9 indicate a positive, but much weaker, relation between *Machine Downloads* and *Arousal*, a more exciting emotion in conference calls. In Columns (4) and (8), *Control* further includes *Earnings Surprise*, defined as the difference between actual earnings and the median analyst forecast.[47] The coefficients associated with *Machine Downloads* barely change.

Based on videos of entrepreneurs pitching investors for funding, Hu and Ma (2020) show that venture capitalists are more likely to invest in start-ups whose founders give pitches that are rated high in valence and arousal. Reactions by VC investors to vocal emotion may well apply to the general capital markets. Our findings support the hypothesis that managers are motivated to manipulate their vocal expressions to achieve a more favorable effect on investors that rely on machine processing, and also justify the anecdotal evidence that managers increasingly seek professional coaching in order to improve vocal performances.[48]

7. Concluding Remarks

This paper presents the first study showing how corporate disclosure in writing and speaking has been reshaped by machine readership employed by algorithmic traders and quantitative analysts. Our findings indicate that increasing AI readership motivates firms to prepare filings that are friendlier to machine parsing and processing, highlighting the growing

[47] Calculating the *Earnings Surprise* variable requires analyst coverage (tracked by the IBES analyst data), which results in a much smaller sample.

[48] Sources: "Listening without prejudice: How the experts analyze earnings calls for lies, bluffs, and other flags", Sterling Wong, *Minyanville*, April 18, 2012, and "How to listen for the hidden data in earnings calls", Alina Dizik, *Chicago Booth Review*, May 25, 2017.

roles of AI in the financial markets and their potential impact on corporate decisions. Firms manage sentiment and tone perception that caters to AI readers by, for example, differentially avoiding words perceived as negative by algorithms, as compared to those perceived as such by human readers. CEOs also aim to present with the vocal qualities that are favorably rated by software. While the literature has shown how investors and researchers apply machine learning and computational tools to extract information from disclosures and news, our study is the first to identify and analyze the *feedback effect*: how companies adjust the way they talk knowing that machines are listening. Such a feedback effect can lead to unexpected outcomes, such as manipulation and collusion (Calvano, Calzolari, Denicolo, and Pastorello, 2020). The technology advancement calls for more studies to understand the impact of and induced behavior by AI in financial economics and in the broad society.[49]

[49] Sports provide an analogous example in a non-finance setting. The English Premier League decided not to let Video Assistant Referee (VAR) overpower referee judgment. One main reason is that players will reverse-engineer and play to the rules underlying the VAR decisions, which will likely lead to undesirable outcomes such as more "low grade" (to the machine) but atrocious (to humans) fouls. See "Why has the introduction of video technology gone so badly in soccer?" James Reade, *Forbes*, December 10, 2020.

Appendix A: Definitions of Variables

Variable	**Definition**
After	An indicator variable equal to one if the time m occurs after a filing is publicly released on EDGAR. It is defined within the [-15, 15]-minute window, where minute 0 is the filing time.
AI Hedge Fund	The percentage of shares outstanding owned by AI hedge funds, classified based on employees' work experience in AI-related projects disclosed on their LinkedIn profiles (Guo and Shi, 2020). It is computed at the stock-quarter level from 13F holdings of hedge funds.
AI Ownership	The firm-year-level aggregate ownership of AI-equipped investment company shareholders in the quarter before the firm's current filing. We classify an investment company as having AI capacity if it has AI-related job postings in the past five years using the job posting data between 2011 and 2018 from Burning Glass.
AI Talent Supply	We first retrieve the number of people between 18 and 64 with college or graduate school degrees in information technology, scaled by population at the state-year level, using data between 2011 and 2018 from Integrated Public Use Microdata Series (IPUMS) surveys. Second, for each firm and during the quarter prior to the current filing, we aggregate state-year-level AI talents over all states based on the headquarters of the investors, weighted by their ownership.
BERT Sentiment	The number of negative sentences in Item 7 of a 10-K filing, scaled by the total number of sentences or the total number of words.
Bid-Ask Spread	The difference between the ask price and the bid price scaled by the midpoint of them, expressed in percentage points and calculated at the minute level following the NBBO rule.
Earnings Surprise	The difference between the actual quarterly earnings and the median earnings forecast of IBES analysts, scaled by the stock price.
Emotion Arousal	The excitedness of speech emotion, calculated from a pretrained Python machine learning package *pyAudioAnalysis.*
Emotion Valence	The positivity of speech emotion, calculated from a pretrained Python machine learning package *pyAudioAnalysis.*
External Financing Needs	The net total issuance in a given year in excess of that in the previous year. The net total issuance is calculated as the sum of the net debt issuance (change in current and long-term debt) and the net equity issuance, scaled by book assets.
Growth	The average sales growth of the past three years.
Harvard Sentiment	The number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing, expressed in percentage points.
IdioVol	The annualized idiosyncratic volatility (using daily data) from the Fama-French three-factor model.
IndAdjRet	The monthly average SIC3-adjusted stock returns over the past year.

(continued)

Variable	Definition
InstOwnership	The ratio of the total shares of institutional ownership to shares outstanding.
Leverage	The ratio of total debt to assets.
Litigation Risk	The machine-learning-predicted probability of litigation at the industry level using a broad set of variables capturing accounting, capital markets, governance, and auditing conditions, developed by Bertomeu, Cheynel, Floyd, and Pan (2021).
Litigious	The number of Loughran-McDonald (LM) litigation-related words in a filing divided by the total number of words in the filing, expressed in percentage points.
LM Sentiment	The number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing, expressed in percentage points.
LM – Harvard Sentiment	*LM Sentiment* minus *Harvard Sentiment.*
Log(#analyst)	The natural logarithm of one plus the number of IBES analysts covering the stock.
Machine Downloads	For a firm's filing at time t, *Machine Downloads* is the natural logarithm of the average number of machine downloads of the firm's historical filings during the $[t-4, t-1]$ quarters. To measure machine downloads, we identify an IP address downloading more than 50 unique firms' filings daily as a machine visitor, the same criterion used by Lee, Ma, and Wang (2015). In addition, we include requests attributed to web crawlers in the Log File Data as machine-initiated. Machine requests are aggregated for each filing within seven days (i.e., days [0, 7]) after it becomes available on EDGAR.
Δ*Machine Downloads*	For a firm's filing at time t, the change in *Machine Downloads* (before taking the natural logarithm) from the previous-year average. Δ*Machine Downloads* is the natural logarithm of the change (A constant is added to ensure the number is positive before taking the natural logarithm).
Machine Readability	The average of five filing attributes, including (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators); (iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters. Each attribute is standardized to a Z-score before being averaged to form a single-index *Machine Readability* measure.

(continued)

Variable	**Definition**
MR Upgrade	An "upgrade" event at the filing (i, j, t) level equal to one if *Machine Readability*, $MR_{i,j,t}$, incurs a significant (i.e., one standard deviation) increase over the previous-year average, $MR_{i,t-1}$, and zero otherwise.
Other Downloads	For a firm's filing on day t, *Other Downloads* is the natural logarithm of the average number of non-machine downloads of the firm's historical filings during the $[t-4, t-1]$ quarters.
Post	An indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before (filings in 2011 are excluded from the analysis).
Post-BERT	An indicator variable equal to one for filings after 2018, and zero otherwise (filings in 2018, when BERT was published, are excluded from the analysis).
ROA	The ratio of EBITDA to assets.
Segment	The number of business segments, following Cohen and Lou (2012). It measures the complexity of business operations.
Size	The natural logarithm of the market capitalization.
Strong Modal	The number of Loughran-McDonald (LM) strong modal words in a filing divided by the total number of words in the filing, expressed in percentage points.
Time to the First Directional Trade	The length of time, in seconds, between the EDGAR publication time stamp and the first directional trade after a filing is publicly released, censored at the end of a 15-minute window. The first directional trade is the first buy (sell) trade at a price below (above) the terminal value at the end of the window, where buy- and sell-initiated trades are classified by the Lee and Ready (1991) algorithm.
Time to the First Trade	The length of time, in seconds, between the EDGAR publication time stamp and the first trade of the issuer's stock, censored at the end of a 15-minute window.
Tobin's Q	The natural logarithm of the ratio of the sum of market value of equity and book value of debt to the sum of book value of equity and book value of debt.
Turnover	The monthly average of the ratio of trading volume to shares outstanding, multiplied by 12.
Uncertainty	The number of Loughran-McDonald (LM) uncertainty-related words in a filing divided by the total number of words in the filing, expressed in percentage points.
Weak Modal	The number of Loughran-McDonald (LM) weak modal words in a filing divided by the total number of words in the filing, expressed in percentage points.

References

Abis, Simona and Laura Veldkamp, 2022, The changing economics of knowledge production, Working paper, Columbia University.

Ahern, Kenneth R., and Denis Sosyura, 2014, Who writes the news? Corporate press releases during merger negotiations, *Journal of Finance* 69, 241–291.

Allee, Kristian D., Matthew D. DeAngelis, and James R. Moon Jr, 2018, Disclosure "scriptability", *Journal of Accounting Research* 56, 363–430.

Asay, H. Scott, Robert Libby, and Kristina Rennekamp, 2018, Firm performance, reporting goals, and language choices in narrative disclosures, *Journal of Accounting and Economics* 65, 380–398.

Balakrishnan, Karthik, Mary Brooke Billings, Bryan Kelly, and Alexander Ljungqvist, 2014, Shaping liquidity: On the causal effects of voluntary disclosure, *Journal of Finance* 69, 2237–2278.

Bernard, Darren, Terrence Blackburne, and Jacob Thornock, 2020, Information flows among rivals and corporate investment, *Journal of Financial Economics* 136, 760–779.

Bertomeu, Jeremy, Edwige Cheynel, Eric Floyd, and Wenqiang Pan, 2021, Using machine learning to detect misstatements, *Review of Accounting Studies* 26, 468–519.

Björkegren, Daniel, Joshua E. Blumenstock, and Samsun Knight, 2020, Manipulation-proof machine learning, Working paper, Brown University and U.C. Berkeley.

Blankespoor, Elizabeth, 2019, The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate, *Journal of Accounting Research* 57, 919–967.

Blankespoor, Elizabeth, Ed deHaan, and Iván Marinovic, 2020, Disclosure processing costs, investors' information choice, and equity market outcomes: A review, *Journal of Accounting and Economics* forthcoming.

Bolandnazar, Mohammadreza, Robert J. Jackson Jr, Wei Jiang, and Joshua Mitts, 2020, Trading against the random expiration of private information: A natural experiment, *Journal of Finance* 75, 5–44.

Bond, Philip, Alex Edmans, and Itay Goldstein, 2012, The real effects of financial markets, *Annual Review of Financial Economics* 4, 339–360.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello, 2020, Artificial intelligence, algorithm pricing, and collusion, *American Economic Review* 100, 3267–3297

Cao, Sean Shun, Kai Du, Baozhong Yang, and Alan L. Zhang, 2021, Copycat skills and disclosure costs: Evidence from peer companies' digital footprints, *Journal of Accounting Research* 59, 1261–1302.

Chen, Huaizhi, Lauren Cohen, Umit Gurun, Dong Lou, and Christopher Malloy, 2020, IQ from IP: Simplifying search in portfolio choice, *Journal of Financial Economics* 138, 118-137.

Chen, Mark A., Qinxi Wu, and Baozhong Yang, 2019, How valuable is FinTech innovation? *Review of Financial Studies* 32, 2062–2106.

Cohen, Lauren, and Dong Lou, 2012, Complicated firms, *Journal of Financial Economics* 104, 383–400.

Cohen, Lauren, Dong Lou, and Christopher Malloy, 2020, Casting conference calls, *Management Science* 66, 5015–5039.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *Journal of Finance* 75, 1371–1415.

Cong, Lin William, Tengyuan Liang, Baozhong Yang, and Xiao Zhang, 2021, Analyzing textual information at scale, *Information to Facilitate Efficient Decision Making: Big Data, Blockchain and Relevance* (ed. Kashi Balachandran), World Scientific Publishers, 239–272

Cong, Lin William, Tengyuan Liang, and Xiao Zhang, 2019, Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, Working paper, Chicago University and Cornell University.

Crane, Alan D., Kevin Crotty, and Tarik Umar, 2021, Hedge funds and public information acquisition, Working paper, Rice University.

Crawford, Vincent P., and Joel Sobel, 1982, Strategic information transmission, *Econometrica* 50, 1431–1451.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461–1499.

Dai, Rui, Lawrence Donohue, Qingyi Drechsler, and Wei Jiang, 2022, Dissemination, publication, and impact of finance research: When novelty meets conventionality, *Review of Finance* forthcoming.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, BERT: Pre-training of deep bidirectional transformers for language understanding, Working paper, Google AI Language.

Diamond, Douglas W., and Robert E. Verrecchia, 1991, Disclosure, liquidity, and the cost of capital, *Journal of Finance* 46, 1325–1359.

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu, 2018, Strategic classification from revealed preferences, *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Driscoll, John C., and Aart C. Kraay, 1998, Consistent covariance matrix estimation with spatially dependent panel data, *Review of Economics and Statistics* 80, 549–560.

Easley, David, and Maureen O'Hara, 2004, Information and the cost of capital, *Journal of Finance* 59, 1553–1583.

Foster, F. Douglas, and S. Viswanathan, 1996, Strategic trading when agents forecast the forecasts of others, *Journal of Finance* 51, 1437–1478.

Gao, Meng, and Jiekun Huang, 2020, Informing the market: The effect of modern information technologies on information production, *Review of Financial Studies* 33: 1367–1411.

García, Diego, 2013, Sentiment during recessions, *Journal of Finance* 68, 1267–1300.

Giannakopoulos, Theodoros, 2015, pyAudioAnalysis: An open-source Python library for audio signal analysis, PloS ONE 10, e0144610.

Goldstein, Itay, and Liyan Yang, 2017, Information disclosure in financial markets, *Annual Review of Financial Economics* 9, 101–125.

Goodhart, Charles, 1975, Problems of monetary management: The UK experience, *Papers in Monetary Economics* 1, Reserve Bank of Australia.

Guo, Xuxi, and Zhen Shi, 2020, The impact of AI talents on hedge fund performance, Working paper, Georgia State University.

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

Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic interpretation of emerging risks in the financial sector, *Review of Financial Studies* 32, 4543–4603.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters, 2016, Strategic classification, *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111–122.

Hennessy, Christopher A., and Charles A. E. Goodhart, 2020, Goodhart's Law and machine learning: A structural approach, Working Paper, London Business School and London School of Economics.

Hodge, Frank D., Jane Jollineau Kennedy, and Laureen A. Maines, 2004, Does search-facilitating technology improve the transparency of financial reporting? *The Accounting Review* 79: 687–703.

Hu, Allen, and Song Ma, 2020, Human interactions and financial investment: A video-based approach, Working paper, Yale University.

Huang, Alan Guoming, Hongping Tan, Russ Wermers, 2020, Institutional trading around corporate news: Evidence from textual analysis, *Review of Financial Studies* 33, 4627-4675.

Huang, Alan Guoming, and Russ Wermers, 2020, Who listens to corporate conference calls? The effect of "soft information" on institutional trading, Working paper, University of Waterloo and University of Maryland.

Hwang, Byoung-Hyoun, and Hugh Hoikwang Kim, 2017, It pays to write well, *Journal of Financial Economics* 124, 373–394.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word power: A new approach for content analysis, *Journal of Financial Economics* 110, 712–729.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou, 2019, Manager sentiment and stock returns, *Journal of Financial Economics* 132, 126–149.

Jiang, Wei, Yuehua Tang, Rachael Jiqiu Xiao, and Vincent Yao, 2021, Surviving the FinTech disruption, NBER Working paper 28668.

Kim, Chansog, Ke Wang, and Liandong Zhang, 2019, Readability of 10-K reports and stock price crash risk, *Contemporary Accounting Research* 36, 1184–1216.

Kim, Oliver, and Robert E. Verrecchia, 1994, Market liquidity and volume around earnings announcements, *Journal of Accounting and Economics* 17, 41–67.

Kim, Oliver, and Robert E. Verrecchia, 1997, Pre-announcement and event-period private information, *Journal of Accounting and Economics* 24, 395–419.

Kothari, Sabino P., Susan Shu, and Peter D. Wysocki, 2009, Do managers withhold bad news? *Journal of Accounting Research*, 47, 241–276.

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Lee, Charles MC, Paul Ma, and Charles CY Wang, 2015, Search-based peer firms: Aggregating investor perceptions through internet co-searches, *Journal of Financial Economics* 116, 410–431.

Lee, Charles MC, and Mark J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733–746.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan, 2020, Measuring corporate culture using machine learning, *Review of Financial Studies* 34, 3265–3315.

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

Loughran, Tim, and Bill McDonald, 2014, Measuring readability in financial disclosures, *Journal of Finance* 69, 1643–1671.

Loughran, Tim, and Bill McDonald, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187–1230.

Loughran, Tim, and Bill McDonald, 2017, The use of EDGAR filings by investors, *Journal of Behavioral Finance* 18, 231–248.

Lucas, Robert E, 1976, Econometric policy evaluation: A critique, *Carnegie-Rochester Conference Series on Public Policy* 1, 19–46.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Mayew, William J., and Mohan Venkatachalam, 2012, The power of voice: Managerial affective states and future firm performance, *Journal of Finance* 67, 1–43.

McLean, R. David, and Jeffrey Pontiff, 2016, Does academic research destroy stock return predictability? *Journal of Finance* 71, 5–32.

Milli, Smitha, John Miller, Anca D. Dragan, and Moritz Hardt, 2019, The social cost of strategic classification, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

Peters, Florian S., and Alexander F. Wagner, 2014, The executive turnover risk premium, *Journal of Finance* 69, 1529–1563.

Russell, James A, 1980, A circumplex model of affect, *Journal of Personality and Social Psychology* 39, 1161–1178.

Tetlock, Paul C, 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

Yang, Yi, Mark Christopher Siy UY, and Allen Huang, 2020, FinBERT: A pretrained language model for financial communications, Working paper, Hong Kong University of Science and Technology.

Verrecchia, Robert E, 2001, Essays on disclosure, *Journal of Accounting and Economics* 32, 97–180.

Figure 1 Trend of Machine Downloads

This figure plots the annual number of machine downloads (blue bars and left axis) and the annual ratio of machine downloads to total downloads (red line and right axis) across all 10-K and 10-Q filings from 2003 to the first half of 2017 (after which the SEC Log File Data Set stopped coverage). Machine downloads are defined as downloads from an IP address downloading more than 50 unique firms' filings daily. The number of machine downloads or total downloads for each filing are recorded as the respective downloads within seven days after the filing becomes available on EDGAR.

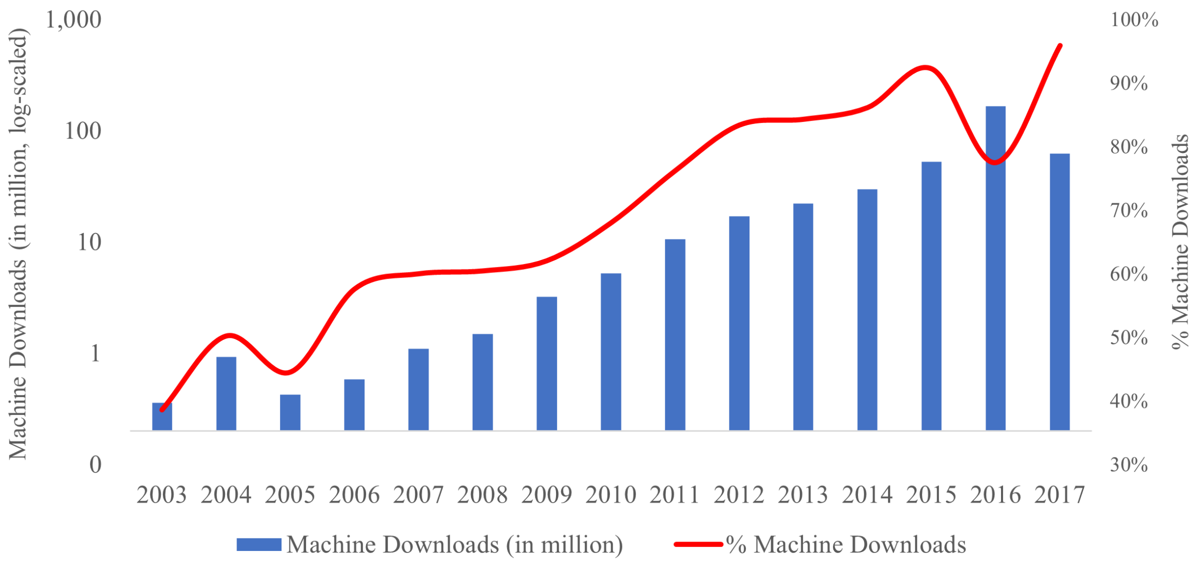

Figure 2 Trend of Machine Readability

This figure plots the annual *Machine Readability* across all 10-K and 10-Q filings from 2004 to 2015. *Machine Readability* is the average of five standardized filing attributes: *Table Extraction*, *Number Extraction*, *Table Format*, *Self-Containedness*, and *Standard Characters*. All attributes are defined in Appendix A.

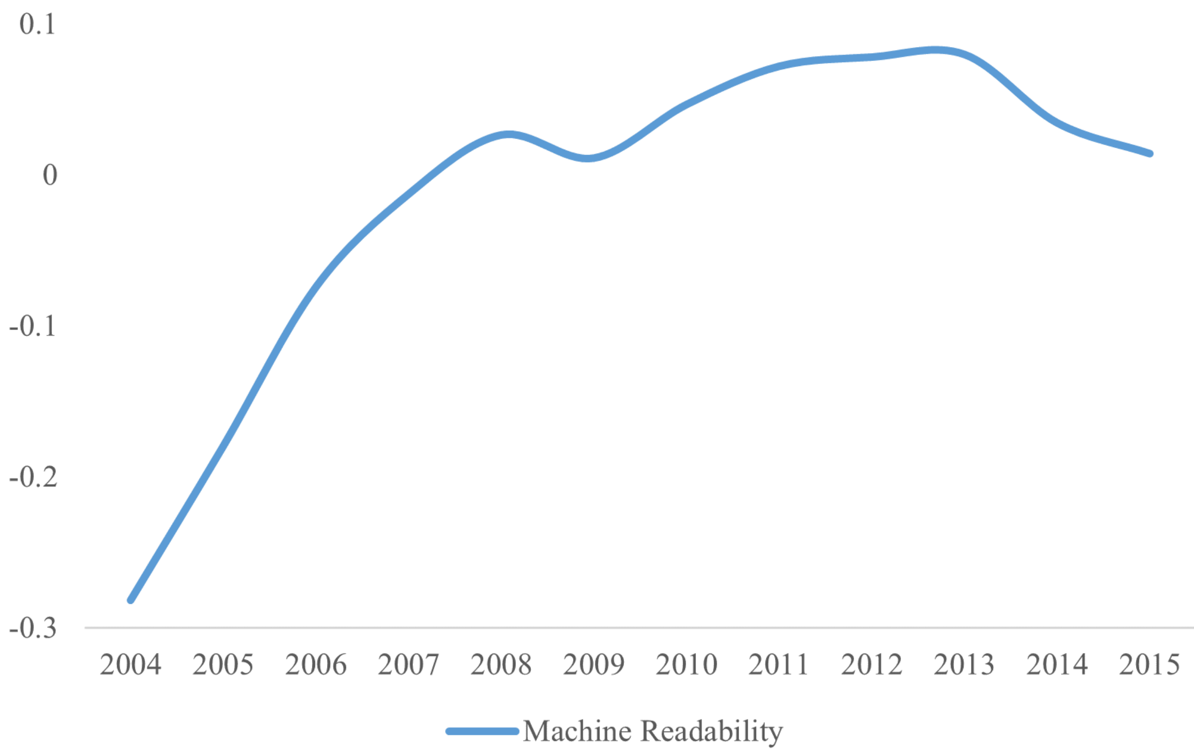

Figure 3 Sentiment Trend and Machine Downloads

This figure plots *LM – Harvard Sentiment* of 10-K and 10-Q filings and compares sentiment of firms with high machine downloads with that of the low group. *LM – Harvard Sentiment* is the difference of *LM Sentiment* and *Harvard Sentiment*. *LM Sentiment* is defined as the number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing. *Harvard Sentiment* is defined as the number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing. Filings are sorted into top tercile or bottom tercile based on *Machine Downloads*, defined in Appendix A. *LM Sentiment* and *Harvard Sentiment* are normalized to one, respectively, in 2010 within each group, one year before the publication of Loughran and McDonald (2011). The dotted lines represent the 95% confidence limits.

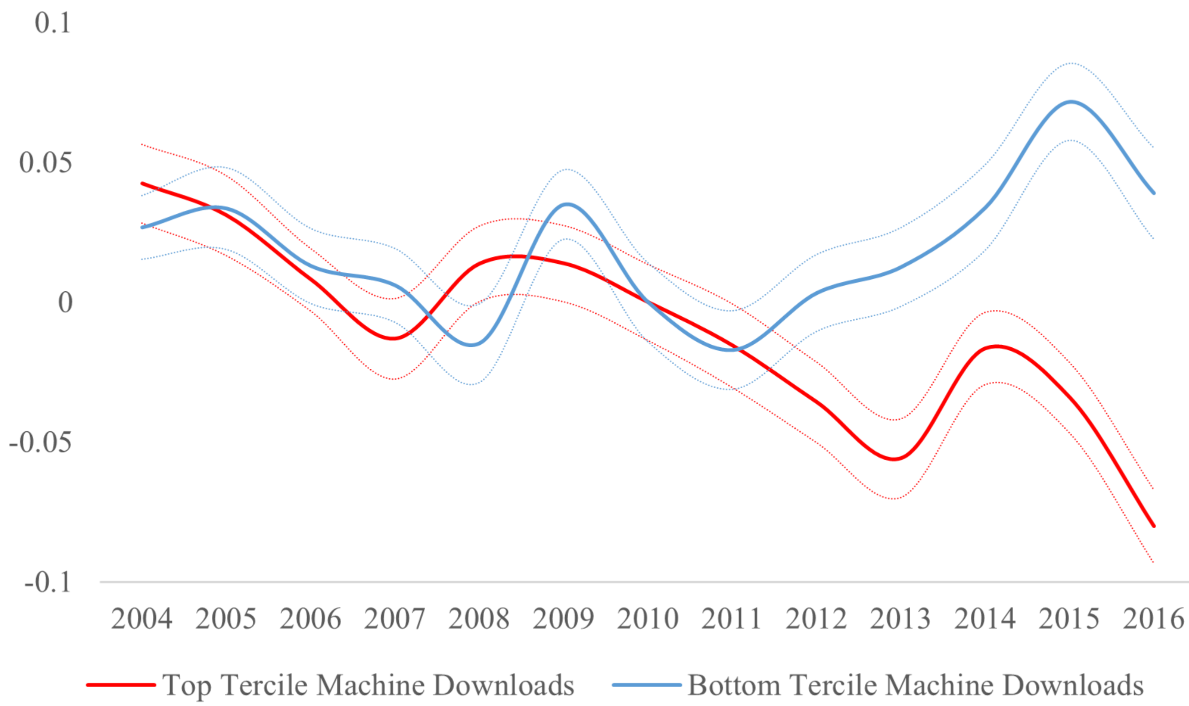

Table 1 Summary Statistics

This table provides summary statistics. Filing-level variables are based on the sample of SEC EDGAR 10-K and 10-Q filings from 2004 to 2016. Conference-call-level variables are based on the sample of the audios of corporate conference calls from 2010 to 2016. Firm-year-level control variables are calculated annually using information available at the previous year-end. Variables are defined in Appendix A.

Variables	Mean	Median	Std	P25	P75	N
Filing level						
Machine Downloads	4.729	4.508	1.763	3.296	6.377	324,607
Other Downloads	3.448	3.474	1.378	2.615	4.363	324,607
Total Downloads	5.09	4.915	1.609	3.829	6.535	324,607
% Machine Downloads	0.742	0.775	0.179	0.623	0.892	324,231
Machine Readability	-0.020	0.125	0.584	-0.224	0.359	199,421
LM – Harvard Sentiment	-2.413	-2.385	0.544	-2.747	-2.047	324,589
LM Sentiment	1.625	1.543	0.599	1.185	1.982	324,589
Harvard Sentiment	4.038	4.021	0.697	3.561	4.492	324,589
Litigious	0.965	0.82	0.537	0.593	1.177	324,589
Uncertainty	1.425	1.377	0.398	1.146	1.652	324,589
Weak Modal	0.521	0.427	0.304	0.314	0.634	324,589
Strong Modal	0.295	0.271	0.133	0.202	0.359	324,589
Conference call level						
Emotion Valence	0.331	0.375	0.261	0.227	0.498	43,462
Emotion Arousal	0.647	0.650	0.138	0.557	0.740	43,462
Firm-year-level control variables						
Size	6.238	6.22	2.022	4.804	7.617	43,764
Tobin's Q	0.672	0.557	0.718	0.178	1.064	43,764
ROA	0.0491	0.101	0.271	0.028	0.163	43,764
Leverage	0.221	0.16	0.244	0.008	0.337	43,764
Growth	0.152	0.0736	0.42	-0.005	0.191	43,764
IndAdjRet	0.000	-0.001	0.039	-0.021	0.019	43,764
InstOwnership	0.482	0.528	0.359	0.080	0.816	43,764
Log(#analyst)	1.498	1.609	1.193	0	2.485	43,764
IdioVol	0.463	0.386	0.289	0.263	0.576	43,764
Turnover	2.150	1.619	1.960	0.826	2.791	43,764
Segment	5.323	5	3.564	2	7	43,764

Table 2 Top Machine Downloaders

This table lists the 20 13F-filing institutional investors with the highest number of machine downloads (#MD) during our sample period of 2004 to 2016.

Rank	Name of institution	#MD	Type of institution
1	Renaissance Technologies	536,753	Quantitative hedge fund
2	Two Sigma Investments	515,255	Quantitative hedge fund
3	Barclays Capital	377,280	Financial conglomerate with asset management
4	JPMorgan Chase	154,475	Financial conglomerate with asset management
5	Point72 Asset Management	104,337	Quantitative hedge fund
6	Wells Fargo	94,261	Financial conglomerate with asset management
7	Morgan Stanley	91,522	Investment bank with asset management
8	Citadel LLC	82,375	Quantitative hedge fund
9	RBC Capital Markets	79,469	Financial conglomerate with asset management
10	D. E. Shaw Co.	67,838	Quantitative hedge fund
11	UBS AG	64,029	Financial conglomerate with asset management
12	Deutsche Bank AG	55,825	Investment bank with asset management
13	Union Bank of California	50,938	Full-service bank with private wealth management
14	Squarepoint Ops	48,678	Quantitative hedge fund
15	Jefferies Group	47,926	Investment bank with asset management
16	Stifel, Nicolaus Company	24,759	Investment bank with asset management
17	Piper Jaffray	18,604	Investment bank with asset management
18	Lazard	18,290	Investment bank with asset management
19	Oppenheimer Co.	15,203	Investment bank with asset management
20	Northern Trust Corporation	11,916	Financial conglomerate with asset management

Table 3 Machine Downloads and Machine Readability

This table examines the relation between the machine readability of a firm's filing and the machine downloads of the firm's past filings. *Machine Downloads* measures the expected machine readership of a filing. Panel A reports a single-index *Machine Readability* score that measures the ease at which a filing can be processed by an automated program. *MR Upgrade* indicates an upgrade event, i.e., when a filing incurs a one standard deviation increase over the previous-year average *Machine Readability*. Δ*Machine Downloads* measures the change of machine readership. Panel B reports the underlying components of *Machine Readability*: *Table Extraction* (the ease of separating tables from text), *Number Extraction* (the ease of extracting numbers from text), *Table Format* (the ease of identifying the information contained in the table), *Self-Containedness* (whether a filing includes all needed information), and *Standard Characters* (the proportion of characters that are standard ASCII characters). Panel C reports alternative machine-readership measures. *AI Ownership* is the aggregate ownership of a firm by AI-equipped investment company shareholders. *AI Talent Supply* measures the local talent supplies to a firm's institutional shareholders, weighted by their ownership; the local talent supply is the available workforce with IT degrees in the state where an investor is headquartered. Control variables include *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. Variables are defined in Appendix A. In all panels, the t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: *Machine Readability*

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable		*Machine Readability*			*MR Upgrade*	
Machine Downloads	0.076***	0.075***	0.060***	0.078***		
	(13.89)	(17.45)	(10.33)	(15.93)		
Δ*Machine Downloads*					0.005***	0.006***
					(2.90)	(3.40)
Other Downloads	0.005	0.002	-0.007	-0.006	0.000	-0.001
	(1.15)	(0.47)	(-1.44)	(-1.33)	(0.20)	(-0.44)
Size			0.004	0.021***	-0.002	-0.001
			(1.05)	(2.66)	(-1.27)	(-0.27)
Tobin's Q			-0.006	-0.008	-0.002	-0.000
			(-0.92)	(-1.00)	(-0.94)	(-0.03)
ROA			0.056***	0.009	0.006	0.026**
			(3.15)	(0.49)	(1.15)	(2.52)
Leverage			-0.087***	-0.037*	0.017***	0.016*
			(-4.62)	(-1.67)	(3.02)	(1.66)
Growth			-0.017**	0.010	0.006**	-0.001
			(-2.34)	(1.27)	(2.29)	(-0.26)
IndAdjRet			0.033	0.013	0.024	0.004
			(0.52)	(0.20)	(0.82)	(0.13)
InstOwnership			0.050***	-0.038	-0.001	0.008
			(2.69)	(-1.50)	(-0.21)	(0.73)
Log(#analyst)			0.005	0.000	-0.003*	-0.003
			(0.79)	(0.02)	(-1.74)	(-0.76)
IdioVol			-0.072***	0.015	0.009	0.004
			(-3.81)	(0.86)	(1.36)	(0.40)

(continued)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable		*Machine Readability*			*MR Upgrade*	
Turnover			-0.002	-0.007***	-0.000	-0.001
			(-1.17)	(-3.16)	(-0.68)	(-0.69)
Segment			0.004***	-0.003	0.001*	0.001
			(3.05)	(-1.42)	(1.95)	(1.09)
Observations	198,358	199,241	150,425	150,346	135,146	135,068
R-squared	0.082	0.363	0.084	0.357	0.025	0.144
Firm FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Components of *Machine Readability*

	(1)	(2)	(3)	(4)	(5)
	Machine Readability				
Dependent Variable	*Table Extraction*	*Number Extraction*	*Table Format*	*Self-Containedness*	*Standard Characters*
Machine Downloads	0.051***	0.028***	0.026***	0.161***	0.125***
	(6.02)	(3.47)	(2.88)	(21.80)	(14.68)
Other Downloads	0.018**	-0.011	0.022**	-0.036***	-0.040***
	(2.37)	(-1.49)	(2.51)	(-6.69)	(-6.08)
Observations	149,484	150,346	149,484	150,245	140,061
R-squared	0.471	0.389	0.439	0.306	0.344
Control Variables	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Panel C: Alternative machine-readership measures

	(1)	(2)	(3)	(4)
Dependent Variable	*Machine Readability*			
AI Ownership	0.515***	0.356***		
	(8.06)	(8.29)		
AI Talent Supply			0.160***	0.192**
			(3.09)	(2.29)
Observations	50,747	50,608	70,969	70,912
R-squared	0.093	0.373	0.088	0.361
Control Variables	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Table 4 Effects of Machine Downloads

This table examines the effects of *Machine Downloads* on trading and information dissemination. *Machine Downloads* measures the expected machine readership of a filing. *Machine Readability* measures the ease at which a filing can be processed by an automated program. Panel A reports the relation between the time to the first trade after a firm's filing is publicly released and the expected machine readership of the filing, and how the machine readability of the filings affects such a relation. *Time to the First Trade* is the length of time, in seconds, between the EDGAR publication time stamp and the first trade of the issuer's stock since the publication. *Time to the First Directional Trade* is defined analogously, where the first directional trade is the first buy (sell) trade at a price below (above) the terminal value at the end of a 15-minute window. Panel B reports the relation between *Machine Downloads* and *Bid-Ask Spread*, where the sample consists of filing-minute-level observations from 15 minutes before to 15 minutes after the posting of the filings. *Bid-Ask Spread* is the difference between the ask price and the bid price scaled by the midpoint, calculated at the minute level following the NBBO rule. *After* is an indicator variable equal to one if the time is after a filing is publicly released and zero otherwise. The sorting variable *Turnover*, the ratio of trading volume to shares outstanding, separates firms into two subsamples by the median. Control variables include *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm in Panel A and by filing in Panel B. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Time to the first trade

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Time to the First Trade*				*Time to the First Directional Trade*			
Machine Downloads	-8.353**	-4.857*	-7.347**	-3.398	-12.365***	-7.540***	-12.374***	-7.258**
	(-2.56)	(-1.68)	(-2.19)	(-1.14)	(-3.94)	(-2.71)	(-3.87)	(-2.55)
Machine Downloads ×			-3.761**	-3.887***			-2.815*	-2.127*
Machine Readability			(-2.46)	(-2.84)			(-1.87)	(-1.67)
Machine Readability			-6.540	-5.980			-5.695	-8.709
			(-0.99)	(-0.92)			(-0.91)	(-1.46)
Other Downloads	15.342***	3.499	15.151***	1.304	13.961***	3.885*	13.436***	2.336
	(5.29)	(1.42)	(5.06)	(0.51)	(4.95)	(1.72)	(4.67)	(1.00)
Observations	161,749	161,664	144,281	144,193	161,749	161,664	144,281	144,193
R-squared	0.116	0.269	0.118	0.272	0.120	0.285	0.122	0.286
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Effects of machine readership: bid-ask spread

	(1)	(2)	(3)	(4)
Dependent Variable	*Bid-Ask Spread*		*Bid-Ask Spread*	
Groups	Entire sample		Low Turnover	High Turnover
Machine Downloads × *After*	0.055***	0.081***	0.080***	0.089***
	(8.46)	(10.91)	(7.18)	(8.97)
Machine Readability × *After*		0.023	0.010	0.030
		(1.15)	(0.33)	(1.10)
Observations	2,673,992	2,416,151	1,203,653	1,212,498
R-squared	0.720	0.732	0.738	0.715
Firm FE	Subsumed	Subsumed	Subsumed	Subsumed
Filing FE	Yes	Yes	Yes	Yes
Minute FE	Yes	Yes	Yes	Yes

Table 5 Machine Downloads and Sentiment: Loughran and McDonald (2011) Publication

This table reports the impact of the publication of Loughran and McDonald (2011) on the relation between the negative sentiment of a firm's filing and the machine downloads of the firm's past filings. *Machine Downloads* measures the expected machine readership of a filing. *LM Sentiment* (*Harvard Sentiment*) is the number of Loughran-McDonald finance-related (Harvard General Inquirer) negative words in a filing, scaled by the total number of words in the filing. *LM – Harvard Sentiment* is the difference between *LM Sentiment* and *Harvard Sentiment*. *Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before. Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*LM – Harvard Sentiment*		*LM Sentiment*		*Harvard Sentiment*	
Machine Downloads	-0.072***	-0.079***	-0.062***	-0.050***	0.010	0.029***
× *Post*	(-6.95)	(-8.94)	(-4.98)	(-4.99)	(0.76)	(2.65)
Machine Downloads	-0.007	-0.011**	-0.009	-0.019***	-0.002	-0.008
	(-1.17)	(-2.46)	(-1.18)	(-3.72)	(-0.23)	(-1.43)
Observations	158,578	158,515	158,578	158,515	158,578	158,515
R-squared	0.217	0.568	0.241	0.632	0.208	0.590
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Table 6 Machine Downloads and Other Tones: Loughran and McDonald (2011) Publication

This table reports the impact of the publication of Loughran and McDonald (2011) on the relation between the various tones of a firm's filing and the machine downloads of the firm's past filings. *Machine Downloads* measures the expected machine readership of a filing. *Litigious*/*Uncertainty*/*Weak Modal*/*Strong Modal* is the number of Loughran-McDonald litigation-related/uncertainty-related/weak modal/strong modal words in a filing, scaled by the total number of words in the filing. *Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before. Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Litigious*		*Uncertainty*		*Weak Modal*		*Strong Modal*	
Machine Downloads × *Post*	-0.056***	-0.057***	-0.016**	-0.021***	-0.028***	-0.034***	-0.008***	-0.007***
	(-5.38)	(-6.02)	(-2.01)	(-3.49)	(-4.85)	(-8.86)	(-4.39)	(-4.39)
Machine Downloads	0.011*	0.007	-0.006	-0.009***	-0.018***	-0.021***	-0.003**	-0.004***
	(1.71)	(1.44)	(-1.33)	(-3.05)	(-5.39)	(-10.05)	(-2.19)	(-4.98)
Observations	158,578	158,515	158,578	158,515	158,578	158,515	158,578	158,515
R-squared	0.188	0.509	0.196	0.600	0.238	0.624	0.277	0.571
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 7 Machine Readability and Sentiment: Cross-Sectional Effects in Terms of Costs and Benefits

This table explores the cross-sectional variation in the relation between machine readability (first two columns)/sentiment (last four columns) and the machine downloads of the firm's past filings. *Litigation Risk*, the machine learning-predicted probability of litigation at a firm's industry, and *External Financing Needs*, the excess net total issuance of a firm, are the sorting variables that separate the sample into the top quartile and the rest. *Machine Downloads* measures the expected machine readership of a filing. *Machine Readability* measures the ease at which a filing can be processed by an automated program. *LM – Harvard Sentiment* measures the difference in sentiments based on Loughran-McDonald finance-related negative words and Harvard General Inquirer negative words. *Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before. Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. Difference of coefficients compares the coefficients on variables of interest *Machine Downloads* (first two columns) and *Machine Downloads* × *Post* (last four columns) between the top-quartile group and the rest. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, for the regression coefficients (two-tailed) and for the difference of coefficients (one-tailed).

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*Machine Readability*		*LM – Harvard Sentiment*		*LM – Harvard Sentiment*	
	External Financing Needs				Litigation Risk	
Groups	Top-quartile	Other	Top-quartile	Other	Top-quartile	Other
Machine Downloads × *Post*			-0.102***	-0.075***	-0.054***	-0.090***
			(-6.60)	(-7.53)	(-3.46)	(-8.81)
Machine Downloads	0.107***	0.076***	-0.025***	-0.011**	-0.018**	-0.012**
	(10.29)	(13.37)	(-2.90)	(-1.96)	(-2.54)	(-2.16)
Difference of Coefficients	0.031***		-0.027*		0.036**	
p-value	0.004		0.071		0.027	
Observations	35,014	101,242	36,984	106,468	48,457	102,467
R-squared	0.439	0.365	0.634	0.572	0.598	0.591
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Table 8 Managing Sentiment in Response to Recent Technology (BERT)

This table examines the impact of the publication of BERT on the relation between the negative sentiment of a firm's 10-K filing and the machine readership on the firm's filing. *BERT Sentiment* is defined as the number of negative sentences, scaled by the total number of sentences in Columns (1) and (2), and scaled by the total number of words in Columns (3) and (4), respectively. *AI Ownership* is a firm's aggregate ownership of AI-equipped investment company shareholders. *AI Talent Supply* measures the local talent supplies to a firm's institutional shareholders, weighted by their ownership; the local talent supply is the available workforce with IT degrees in the state where an investor is headquartered. *Post-BERT* is an indicator variable equal to one for filings after 2018, and zero before 2018. Control variables include *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable	*BERT Sentiment*			
	NegSent/TotalSent		NegSent/TotalWords	
AI Ownership × *Post-BERT*	-4.276**		-0.190**	
	(-2.13)		(-2.37)	
AI Ownership	2.025		0.096	
	(1.08)		(1.27)	
AI Talent Supply × *Post-BERT*		-0.983***		-0.041***
		(-3.61)		(-3.98)
AI Talent Supply		-0.522		-0.010
		(-1.18)		(-0.65)
Observations	6,399	6,627	6,399	6,627
R-squared	0.795	0.796	0.803	0.804
Control Variables	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry FE	No	No	No	No
Year FE	Yes	Yes	Yes	Yes

Table 9 Machine Downloads and Managers' Emotion during Conference Calls

This table examines the relation between the manager's speech emotion during conference calls and the machine downloads of the firm's past filings. *Machine Downloads* measures the expected machine readership of the most recent filing before a firm's conference call. *Emotion Valence* and *Emotion Arousal* measure the positivity and excitedness, respectively, of the conference call speech emotion. Control variables include *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment* as in the previous tables. Columns (4) and (8) further include *EarningsSurprise* as an additional control. All variables are defined in Appendix A. The sample consists of audio of conference calls between January 2010 and December 2016. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable		*Emotion Valence*				*Emotion Arousal*		
Machine Downloads	0.043***	0.035***	0.042***	0.042***	0.004*	0.003	0.005**	0.007**
	(11.40)	(8.13)	(11.14)	(8.84)	(1.79)	(0.94)	(2.28)	(2.49)
Other Downloads	-0.017***	-0.014***	-0.017***	-0.012***	-0.006***	0.000	-0.006***	-0.006***
	(-5.74)	(-4.32)	(-5.67)	(-3.12)	(-3.65)	(0.19)	(-3.71)	(-2.92)
Observations	43,336	41,340	41,224	27,437	43,336	41,340	41,224	27,437
R-squared	0.389	0.189	0.383	0.388	0.395	0.132	0.395	0.469
Control Variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Firm FE	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Industry FE	No	Yes	No	No	No	Yes	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Internet Appendix for "How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI"

1. A Model of Machine Trading and Disclosure

We consider a one-period model with informed trading similar to Kyle (1985) but augmented with machine trading and firm disclosure. A single productive firm is financed all by equity. The firm's stock is a risky asset that has an initial price p_0 at the beginning of the period and a payoff $\tilde{v} = p_0 + \epsilon$, with $\epsilon \in N(0, \Sigma_0)$. The value of $\tilde{v}$ will be realized at the end of the period.

The firm makes a disclosure regarding its fundamentals in the beginning of the period. A machine reader/trader parses the firm's disclosure and obtains a signal about the firm's fundamental value.[1]

$$s = \tilde{v} + \eta_1 + \eta_2. \tag{1}$$

where $\eta_1 \sim N(a_\eta, \Sigma_1)$ and $\eta_2 \sim N(0, \Sigma_\eta)$ are independent random variables. η_1 represents the tone of the firm's disclosure (with a given variance Σ_1) and η_2 represents the noise of the disclosure (with mean 0) when parsed by the machine reader. The firm can modify its disclosure (with costs) to affect the values of a_η and Σ_η, respectively, thereby influencing the level and precision of the signal received by the machine trader.

Σ_η is meant to capture machine readability since a lower Σ_η implies that the machine reader can obtain more precise information from the firm's disclosure. The parameter a_η reflects the average tone of the disclosure. To the extent that machine readers use a well-known algorithm (e.g., the Loughran-McDonald dictionary) to compute the tone or sentiment of the disclosure, the firm can revise the tone of its disclosure in a specific way (e.g., by

[1]We first consider the case with one machine trader to better derive the intuition of the model. Later, we extend the model to having multiple machine traders and human traders.

reducing the use of LM-negative words) to influence the value of a_η derived by the machine.[2] The neutral values of these parameters are $(a_{\eta,0} = 0, \Sigma_{\eta,0})$. In other words, if the firm does not "manage" its disclosure policy to cater to the machine readers, then $a_\eta = 0$ and $\Sigma_\eta = \Sigma_{\eta,0}$, which is the upper bound. That is, firm actions, e.g., technology upgrades, are meant to increase machine readability from the neutral state.

The machine trader submits a trade $\tilde{x}$ after observing the signal s. There is a group of noise traders who make an aggregate trade $\tilde{u} \sim N(0, \sigma_u^2)$. A market maker observes the aggregate order $\tilde{y} = \tilde{x} + \tilde{u}$ and sets the price p. The trades by the machine and noise traders are then executed at the price p. At the end of the period, the payoff $\tilde{v}$ realizes for all holders of the asset and the game ends.

The firm has a risk-neutral utility function

$$U(a_\eta, \Sigma_\eta) = E_f[kp - c_1 a_\eta^2 - c_2(\Sigma_\eta - \Sigma_{\eta,0})^2], \tag{2}$$

with $k, c_1, c_2 > 0$, where E_f is expectation with respect to the firm's information set and belief. The first term kp is increasing in the stock price p, reflecting the reality that managerial payoffs are increasing in stock prices. (Because the terminal value of the firm is exogenous, it does not affect utility optimization.) The second term refers to the cost of manipulating tones in disclosure, which can be related to reputation and litigation risk. The third term reflects the cost of increasing machine readability. We note that with higher machine readability or more precise machine signals, the machine trader responds more to the signal, which in turn increases the impact of tones on prices. Therefore, under such an objective function, the firm desires, from an initial level, to adopt more positive tones and higher machine readability but is eventually constrained by the costs in mispricing and technology upgrades.

The machine trader chooses a trade size $\tilde{x}$ to optimize its expected profit conditional on

[2] We assume that the machine trader and the market maker do not fully undo the potential bias introduced by tone management in a one-period model. Over the long run, the machine reader could potentially modify its algorithm and thus improve the estimate of the "real" tone of the document, which is out of the scope of the current model.

its information set, taking into account the price impact of the trades, i.e.,

$$\tilde{x} = \max_{x} E\left[x(\tilde{v} - p)|\tilde{v} + \eta_1 + \eta_2\right]. \tag{3}$$

To break even, the market maker sets the price to be the expected value of the stock given its information set,

$$p = E[\tilde{v}|\tilde{y}]. \tag{4}$$

Following the literature, we consider equilibria in which the machine trader and market maker adopt linear strategies. The machine trader chooses a linear strategy that maximizes its profit,

$$\tilde{x} = \beta s + \mu = \beta(\tilde{v} + \eta_1 + \eta_2) + \mu, \tag{5}$$

where β and μ are parameters resulting from the optimization. β is the sensitivity of the machine's trade to the signal. The market maker selects a linear pricing scheme that breaks even,

$$p = \lambda \tilde{y} + p_0, \tag{6}$$

where λ is the resulting parameter that indicates the inverse depth of the market, that is, a higher value of λ indicating less liquidity in the market.

Proposition 1. *Given a disclosure strategy (a_η, Σ_η) by the firm, there exists a linear equilibrium such that the strategies (β, λ) of the machine trader and the market maker are given as follows,*

$$\beta = \sqrt{\frac{\sigma_u^2}{\Sigma_0 + \Sigma_1 + \Sigma_\eta}}, \tag{7}$$

$$\lambda = \frac{1}{2} \frac{\Sigma_0}{\sqrt{\Sigma_0 + \Sigma_1 + \Sigma_\eta} \sigma_u}. \tag{8}$$

Proofs of all propositions are appended at the end of this document. Importantly, the above proposition provides the following intuition: When machine readability improves (lower Σ_η), the machine trades more aggressively (higher β) and the stock becomes less liquid (higher λ). This is because a more precise signal for the machine trader increases the information asymmetry between the market maker and the machine trader and thus decreases liquidity. Without loss of generality, we assume $\Sigma_1 = 0$ below.[3]

The following proposition shows that in the presence of the machine trader, the firm has strictly positive incentives to increase machine readability and decrease the negative tone of its disclosure. Intuitively, increasing the tone of firm disclosure as captured by the machine algorithm will feed into the machine reader's trades and thus stock prices, positively affecting the firm management's utility. Increasing machine readability facilitates the incorporation of disclosure information into prices and thus benefits the firm more. Furthermore, the lower the costs of adjustment of tone/readability, the greater the firm will change its disclosure along these dimensions.

Proposition 2. *In the model with the machine trader, the firm has an optimal disclosure policy* (a_η, Σ_η) *with* $a^*_\eta > 0, \Sigma^*_\eta < \Sigma_{\eta,0}$. *Furthermore,*

$$\frac{\partial a^*_\eta}{\partial c_1} < 0,$$
$$\frac{\partial \Sigma^*_\eta}{\partial c_2} > 0.$$

In the following, we consider an extension of the basic model to the case with multiple human and machine readers. This allows us to derive quantitative implications of an increasing presence of machine traders.

There are $N = N_H + N_M$ traders, of which N_H are human readers and N_M are machine readers. The total number of traders, N, is fixed for simplicity as the focus of the model is on the variations in the presence of machine traders. The human traders observe a signal $v + \eta_H$, with $\eta_H \sim N(0, \Sigma_H)$. The machine traders obtain a signal $v + \eta_{1,M} + \eta_{2,M}$ from the firm's disclosure,[4] with $\eta_{1,M} \sim N(a_M, \Sigma_1), \eta_{2,M} \sim N(0, \Sigma_M)$. $\eta_H, \eta_{1,M}, \eta_{2,M}$ are mutually independent. The assumptions about noise traders, the market marker, and the firm remain the same as in the basic model.

As before, the firm can manage tone and machine readability of its disclosure by changing the values of a_M and Σ_M. Such activities do not affect the signal received by human traders.

[3]In the general case $\Sigma_1 > 0$, results and formulas remain the same by replacing Σ_η by $\Sigma_\eta + \Sigma_1$ throughout.
[4]For simplicity, we assume all machine (human) traders observe the same signal. Allowing different signals for the traders does not change the results qualitatively.

Note that such an assumption is not to rule away the possibility that firms may also want to manage tone and readability with human readers; instead, we simply assume that such an interaction had reached an equilibrium before the beginning of the period or before the appearance of machine traders. In other words, this study focuses on incremental disclosure management catering to machine readers. We again consider symmetric linear equilibria in which each machine (human) trader adopts a linear strategy with trade intensity β_M (β_H) and the market maker employs a linear strategy with parameter λ.

Proposition 3. *There exists a unique linear symmetric trading equilibrium such that the strategies* $(\beta_M, \beta_H, \lambda)$ *of the machine traders, human traders, and market maker are given as follows,*

$$\lambda = \frac{\Sigma_0}{\sigma_u} \frac{\sqrt{N_H(\Sigma_H + \Sigma_0) f_M^2 + N_M(\Sigma_M + \Sigma_0) f_H^2}}{f_M f_H + f_H N_M \Sigma_0 + f_M N_H \Sigma_0}. \tag{9}$$

$$\beta_M = \frac{\sigma_u f_H}{\sqrt{N_H(\Sigma_H + \Sigma_0) f_M^2 + N_M(\Sigma_M + \Sigma_0) f_H^2}}, \tag{10}$$

$$\beta_H = \frac{\sigma_u f_M}{\sqrt{N_H(\Sigma_H + \Sigma_0) f_M^2 + N_M(\Sigma_M + \Sigma_0) f_H^2}}. \tag{11}$$

where $f_M = (N_M + 1)\Sigma_M + \Sigma_0$ *and* $f_H = (N_H + 1)\Sigma_H + \Sigma_0$.

The following proposition summarizes the effect of the number of machine traders on disclosure tone, machine readability, and liquidity, under the conditions that there is a large number of traders so that each trade is close to being atomistic, and that machine reader presence is sufficiently large.

Proposition 4. *Assume that* N *is sufficiently large,* $\frac{N_M}{N_H} \geq \sqrt{\frac{\Sigma_H + \Sigma_0}{\Sigma_M + \Sigma_0}}$, *and* $\frac{\Sigma_H}{\Sigma_0 + \Sigma_M} > 5$. *Given a linear symmetric equilibrium* $(\beta_M, \beta_H, \lambda)$ *and the corresponding optimal disclosure policy* (a_M^*, Σ_M^*) *by the firm, the following holds.*
(1) Disclosure tone a_M^* *increases with the number of machine traders,* N_M, *i.e.,*

$$\frac{\partial a_M^*}{\partial N_M} > 0.$$

(2) Machine readability increases with the number of machine readers, i.e.,

$$\frac{\partial \Sigma_M^*}{\partial N_M} < 0.$$

(3) Stock liquidity decreases with the number of machine readers, i.e.,

$$\frac{\partial \lambda}{\partial N_M} < 0.$$

First, when there are more machine traders, their trades exert a greater influence on the stock price and thus the firm has a greater incentive to positify its disclosure tone. Second, the firm also has the incentive to improve machine readability. This is because the more precise signals for machine, the greater the machine reader's trades would be, and the greater impact the tone will have on the stock price. In other words, more precise signals coupled with positive tones lead to higher expected prices. Finally, given more accurate signals for machine traders and a greater number of machine traders, information asymmetry between the market maker and the machine traders become more severe. Anticipating this, the market maker thus increases the sensitivity of price to trade to avoid being taken advantage of by the machine trader. As a result, stock liquidity decreases.

As a final remark, we note that in this equilibrium with both human and machine traders, humans will trade less when machines get more precise signals because humans will be at a disadvantage in terms of information asymmetry. Under the model assumption that machine presence is sufficiently large, the market maker faces more adverse selection on the net, hence deteriorating market liquidity.

Proofs of Propositions

Proof of Proposition 1. Assuming that the machine trader and market maker adopt the linear strategies as in Equations (5) and (6), we examine the equilibrium conditions that must be satisfied for the strategy pair (β, λ). We use the notation $\eta = \eta_1 + \eta_2$ below. First, for a trade position x by the machine trader after learning the firm's disclosure, its profits are given by

$$\begin{aligned} E\left[\Pi_x|\tilde{v}+\eta\right] &= E\left[x(\tilde{v}-p)|\tilde{v}+\eta\right] = E[x(\tilde{v}-(\lambda\tilde{y}+p_0))|\tilde{v}+\eta] \\ &= E[x(\tilde{v}-(\lambda x+\lambda\tilde{u}+p_0))|\tilde{v}+\eta] \\ &= -\lambda x^2 + (E[\tilde{v}|\tilde{v}+\eta]-p_0)x. \end{aligned}$$

The first-order condition of the above implies that the optimal trade satisfies

$$\tilde{x} = \frac{1}{2\lambda}\left(E[\tilde{v}|\tilde{v}+\eta] - p_0\right) = \frac{1}{2\lambda}\left(\frac{\Sigma_0}{\Sigma_0+\Sigma_1+\Sigma_\eta}(\tilde{v}+\eta) - p_0\right), \tag{12}$$

where we used the condition that random variables in the model such as $\tilde{v}$ and η are jointly normally distributed. Comparing equations (5) and (12), we obtain

$$\beta = \frac{1}{2\lambda}\frac{\Sigma_0}{\Sigma_0+\Sigma_1+\Sigma_\eta}. \tag{13}$$

Next, we consider the market maker's break-even condition.

$$\begin{aligned} p = E[\tilde{v}|\tilde{y}] &= E[\tilde{v}] + \frac{\mathrm{Cov}(\tilde{v},\tilde{y})(\tilde{y}-E[\tilde{y}])}{\mathrm{Var}(\tilde{y})} \\ &= p_0 + \frac{\mathrm{Cov}(\tilde{v},\beta(\tilde{v}+\eta)+\tilde{u})(\tilde{y}-E[\beta(\tilde{v}+\eta)+\tilde{u}+\mu]}{\mathrm{Var}(\beta(\tilde{v}+\eta)+\tilde{u})} \\ &= p_0 + \frac{\beta\Sigma_0}{\sigma_u^2+\beta^2(\Sigma_0+\Sigma_1+\Sigma_\eta)}(\tilde{y}-(\mu+\beta p_0)) \qquad (14) \\ &= \frac{\beta\Sigma_0}{\sigma_u^2+\beta^2(\Sigma_0+\Sigma_1+\Sigma_\eta)}\tilde{y} + \frac{p_0(\sigma_u^2+\beta^2\Sigma_\eta)-\beta\Sigma_0\mu}{\sigma_u^2+\beta^2(\Sigma_0+\Sigma_1+\Sigma_\eta)}. \qquad (15) \end{aligned}$$

Equations (6) and (15) imply that

$$\lambda = \frac{\beta\Sigma_0}{\sigma_u^2+\beta^2(\Sigma_0+\Sigma_1+\Sigma_\eta)}. \tag{16}$$

We then solve from (13) and (16) that

$$\beta = \sqrt{\frac{\sigma_u^2}{\Sigma_0+\Sigma_1+\Sigma_\eta}}, \tag{17}$$

$$\lambda = \frac{1}{2}\frac{\Sigma_0}{\sqrt{\Sigma_0+\Sigma_1+\Sigma_\eta}\sigma_u}. \tag{18}$$

□

Proof of Proposition 2. We consider the optimality conditions for the firm's disclosure policy. First, note that in the linear equilibrium, from (5), (6), (7), and (8), period-1 price can be written as

$$\begin{aligned} p &= p_0 + \lambda(\beta(\tilde{v}+\eta)+\tilde{u}) \\ &= p_0 + \frac{\Sigma_0}{2(\Sigma_0+\Sigma_\eta)}(\tilde{v}+\eta) + \frac{1}{2}\frac{\Sigma_0}{\sqrt{\Sigma_0+\Sigma_\eta}\sigma_u}\tilde{u}. \end{aligned}$$

Again note that all random variables in the model are jointly normally distributed. The firm's utility function is thus given by

$$\begin{aligned} U(a_\eta, \Sigma_\eta) &= E_f[kp - c_1 a_\eta^2 - c_2(\Sigma_{\eta,0} - \Sigma_\eta)^2] \\ &= kE_f[p] - c_1 a_\eta^2 - c_2(\Sigma_{\eta,0} - \Sigma_\eta)^2 \\ &= kp_0 + k(\lambda\beta)a_\eta - c_1 a_\eta^2 - c_2(\Sigma_{\eta,0} - \Sigma_\eta)^2 \\ &= kp_0 + k\frac{\Sigma_0}{2(\Sigma_0 + \Sigma_\eta)}a_\eta - c_1 a_\eta^2 - c_2(\Sigma_{\eta,0} - \Sigma_\eta)^2. \end{aligned} \tag{19}$$

The first-order condition for a_η in (19) implies that

$$a_\eta^* = \frac{k\Sigma_0}{4c_1(\Sigma_0 + \Sigma_\eta)}. \tag{20}$$

Substituting this back into (19), the optimization problem for Σ_η is equivalent to

$$\Sigma_\eta^* = \max_{\Sigma_\eta} \frac{k^2\Sigma_0^2}{4c_1(\Sigma_0 + \Sigma_\eta)^2} - c_2(\Sigma_{\eta,0} - \Sigma_\eta)^2. \tag{21}$$

Since the first term above is strictly decreasing in Σ_η and the second term has zero derivative at $\Sigma_\eta = \Sigma_{\eta,0}$, it is easy to see that the problem has an optimal solution $\Sigma_\eta^* < \Sigma_{\eta,0}$. Equation (21) also implies that $\frac{\partial \Sigma_\eta^*}{\partial c_2} > 0$ and $\frac{\partial \Sigma_\eta^*}{\partial c_1} < 0$. This together with Equation (20) imply that $\frac{\partial a*_\eta}{\partial c_1} < 0$. □

Proof of Proposition 3. We first analyze the optimal trading policies for machine traders and human traders. For simplicity, we use the notation $\eta_M = \eta_{1,M} + \eta_{2,M}$ below. The profit of a machine trader i given a trade x_i is:

$$\begin{aligned} E\left[\Pi_{i,x_i} | \tilde{v} + \eta_M\right] &= E\left[x_i(\tilde{v} - p) | \tilde{v} + \eta_M\right] = E[x_i(\tilde{v} - (\lambda\tilde{y} + p_0)) | \tilde{v} + \eta_M] \\ &= E[x_i(\tilde{v} - (\lambda(x_i + x_{-i} + p_0)) | \tilde{v} + \eta_M] \\ &= -\lambda x_i^2 + (E[\tilde{v} + (N_M - 1)x_M + N_H x_H | \tilde{v} + \eta_M] - p_0)x_i. \end{aligned}$$

Here we used the fact that other machine traders and human traders are deploying an optimal trade size x_M and x_H, respectively, in the symmetric equilibrium. Equating the optimal x_i above with x_M, we obtain

$$2\lambda x_M + \lambda(N_H E[x_H | \tilde{v} + \eta_M] + (N_M - 1)x_M) = E[\tilde{v} - p_0 | \tilde{v} + \eta_H]. \tag{22}$$

Similarly, we obtain the following equation from analyzing human traders' optimal policy,

$$2\lambda x_H + \lambda((N_H - 1)x_H + N_M E[x_M | \tilde{v} + \eta_H]) = E[\tilde{v} - p_0 | \tilde{v} + \eta_M]. \tag{23}$$

Since $x_H = \beta_i(\tilde{v} + \eta_i - p_0)$ for $i = H, M$ in the linear strategy equilibrium and $E[\tilde{v} | \tilde{v} + \eta_i] =$

$\frac{\Sigma_0}{\Sigma_0+\Sigma_i}$ $(i = M, H)$, Equations (22) and (23) imply

$$\lambda(N_H + 1)\beta_H(\Sigma_0 + \Sigma_H) + \lambda N_M\beta_M\Sigma_0 = \Sigma_0, \tag{24}$$

$$\lambda N_H\beta_H\Sigma_0 + \lambda(N_M + 1)\beta_M(\Sigma_0 + \Sigma_M) = \Sigma_0. \tag{25}$$

Subtracting the two equations above and simplifying, we obtain

$$\beta_H((N_H + 1)\Sigma_H + \Sigma_0) = \beta_M((N_M + 1)\Sigma_M + \Sigma_0). \tag{26}$$

Letting $f_M = (N_M + 1)\Sigma_M + \Sigma_0$ and $f_H = (N_H + 1)\Sigma_H + \Sigma_0$, there then exists a constant K such that

$$\beta_M = Kf_H, \quad \beta_H = Kf_M. \tag{27}$$

Analyzing the optimal strategy by the market maker, we obtain

$$\begin{aligned} p = E[\tilde{v}|\tilde{y}] &= E[\tilde{v}] + \frac{\text{Cov}(\tilde{v},\tilde{y})(\tilde{y} - E[\tilde{y}])}{\text{Var}(\tilde{y})} \\ &= p_0 + \frac{\text{Cov}(\tilde{v}, N_M x_M + N_H x_H + \tilde{u})(\tilde{y} - E[\tilde{y}])}{\text{Var}(N_M x_M + N_H x_H + \tilde{u})}. \end{aligned}$$

Equating the coefficient of $\tilde{y}$ above with λ, we have

$$\lambda = \frac{(N_H\beta_H + N_M\beta_M)\Sigma_0}{(N_H\beta_H + N_M\beta_M)^2\Sigma_0 + N_H^2\beta_H^2\Sigma_H + N_M^2\beta_M^2\Sigma_M + \sigma_u^2}. \tag{28}$$

Eliminating λ from Equations (25) and (28), we obtain

$$N_H(\Sigma_H + \Sigma_0)\beta_H^2 + N_M(\Sigma_M + \Sigma_0)\beta_M^2 = \sigma_u^2. \tag{29}$$

Plugging Equation (27) into the above, we solve K as

$$K^2 = \frac{\sigma_u^2}{N_H(\Sigma_H + \Sigma_0)f_M^2 + N_M(\Sigma_M + \Sigma_0)f_H^2}. \tag{30}$$

Equations (27) and the above representation of K then give the desired formulas for β_M and β_H. We can then obtain the formula for λ by plugging those into Equation (25).

□

Proof of Proposition 4. We first analyze the property of the illiquidity parameter λ. Since N_M and N_H are sufficiently large, we will just consider the leading terms containing them in various expressions. By this approach, Equation (9) implies that

$$\lambda \approx c\frac{\sqrt{N_H N_M(N_M(\Sigma_M + \Sigma_0) + N_H(\Sigma_H + \Sigma_0))}}{N_H N_M} \tag{31}$$

$$= c\sqrt{\frac{\Sigma_M + \Sigma_0}{N_H} + \frac{\Sigma_H + \Sigma_0}{N_M}}. \tag{32}$$

The derivative of the function $\frac{\Sigma_M+\Sigma_0}{x}+\frac{\Sigma_H+\Sigma_0}{N-x}$ is positive if $\frac{x^2}{(N-x)^2}>\frac{\Sigma_H+\Sigma_0}{\Sigma_M+\Sigma_0}$. Since $\frac{N_M}{N_H}>\sqrt{\frac{\Sigma_H+\Sigma_0}{\Sigma_M+\Sigma_0}}$, we conclude that λ is increasing in N_M, or $\frac{\partial\lambda}{\partial N_M}>0$.

Next, we proceed to prove $\frac{\partial\Sigma_M^*}{\partial x}<0$. Consider the optimization problem of the firm:

$$\begin{aligned}U(a_\eta,\Sigma_\eta)&=E_f[kp-c_1a_M^2-c_2(\Sigma_{M,0}-\Sigma_M)^2]\\&=kp_0+k\lambda\beta_MN_Ma_M-c_1a_M^2-c_2(\Sigma_{M,0}-\Sigma_M)^2.\end{aligned}\tag{33}$$

The optimal tone is thus given by

$$a_M^*=\frac{k}{2c_1}\lambda\beta_MN_M.\tag{34}$$

Recall that $f_M=(N_M+1)\Sigma_M+\Sigma_0$ and $f_H=(N_H+1)\Sigma_H+\Sigma_0$. From Equations (9), (10), and (11),

$$\begin{aligned}\lambda\beta_MN_M&=\frac{N_Mf_H\sigma_0^2}{(N_H+1)f_M(\Sigma_H+\Sigma_0)+N_Mf_H\sigma_0^2}\\&=\frac{1}{\frac{(N_H+1)f_M}{N_Mf_H}\frac{\Sigma_H+\Sigma_0}{\Sigma_0}+1}\\&=\frac{1}{\frac{\Sigma_M+\Sigma_0+\frac{2\Sigma_0+\Sigma_M}{N_M}}{\Sigma_H+\Sigma_0+\frac{\Sigma_0}{N_H+1}}+1}=\frac{1}{d_1\Sigma_M+d_2}.\end{aligned}\tag{35}$$

Here we use the notations

$$d_1=\frac{1+\frac{1}{N_M}}{\Sigma_H+\Sigma_0+\frac{\Sigma_0}{N_H+1}},\tag{36}$$

$$d_2=1+\frac{\Sigma_0+\frac{2\Sigma_0}{N_M}}{\Sigma_H+\Sigma_0+\frac{\Sigma_0}{N_H+1}}.\tag{37}$$

Plugging (35) into (34) and then (33), we have the firm's objective equal to

$$\frac{c}{(d_1\Sigma_M+d_2)^2}-c_2(\Sigma_{M,0}-\Sigma_M)^2+kp_0,\tag{38}$$

for some constant c. The first-order condition based on Equation (38) implies that the optimal machine readability parameter Σ_M^* satisifies

$$\Sigma_M^*=\frac{c'd_1}{(d_1\Sigma_M^*+d_2)^3}+c'',\tag{39}$$

where $c'>0$ and c'' are constants. We next take derivative with respect to $x=N_M$ on both sides of (39) and simplify:

$$\frac{\partial \Sigma_M^*}{\partial x}\left(1 + 3c' \frac{d_1^2}{(d_1\Sigma_M + d_2)^4}\right) = c' \frac{(d_2 - 2d_1\Sigma_M)\frac{\partial d_1}{\partial x} - 3d_1\frac{\partial d_2}{\partial x}}{(d_1\Sigma_M + d_2)^4}. \tag{40}$$

Next, note that $\frac{\partial d_1}{\partial x} < 0, \frac{\partial d_1}{\partial x} < 0$ and that $\left|\Sigma_0 \frac{\partial d_1}{\partial x}\right| > \frac{1}{2}\left|\frac{\partial d_2}{\partial x}\right|$. Therefore, we can conclude that $\frac{\partial \Sigma_M^*}{\partial N_M} = \frac{\partial \Sigma_M^*}{\partial x} < 0$ if $d_2 - 2d_1\Sigma_M - 6d_1\Sigma_0 > 0$. But this is true because $\frac{\Sigma_H}{\Sigma_0 + \Sigma_M} > 5$.

Finally, we analyze how a_M^* depends on the number of machine traders. Equations (34) and (35) imply that

$$a_M^* = \frac{c_3}{d_1\Sigma_M + d_2}, \tag{41}$$

for some constant $c_3 > 0$. Given that $\frac{\partial d_1}{\partial x} < 0, \frac{\partial d_1}{\partial x} < 0$ and $\frac{\partial \Sigma_M^*}{\partial x} < 0$, it follows that $\frac{\partial a_M^*}{\partial N_M} = \frac{\partial a_M^*}{\partial x} < 0$. □

2. List of Supplementary Tables and Figures

Figure IA.1 Machine Readability: Excerpts of Two 10-K Filings

This figure shows two sample fillings, one with a low *Machine Readability* score (-1.09, or 1.90 standard deviation below the mean) by APPLEBEES INTERNATIONAL INC in 2005 and one with a high *Machine Readability* score (1.37, or 2.38 standard deviation above the mean) by BANK OF HAWAII CORP in 2012. *Machine Readability* is the average of five standardized filing attributes, including (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators); (iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters.

Excerpt 1. APPLEBEES INTERNATIONAL INC, CIK: 0000853665, March 30, 2005

```
We opened 32 new company Applebee's restaurants in 2004 and anticipate opening
at least 40 new company Applebee's restaurants in 2005, excluding up to eight
restaurants that were closed in 2004 by a former franchisee which we may re-open
in Memphis, Tennessee. The following table shows the areas where our company
restaurants were located as of December 26, 2004:

                                   Area
                  -----------------------------------------------------------
                  New England (includes Maine, Massachusetts, New Hampshire,
                      New York, Rhode Island and Vermont)....................          65
                  Detroit/Southern Michigan..................................          62
                  Minneapolis/St. Paul, Minnesota............................          58
                  St. Louis, Missouri/Illinois...............................          47
                  North/Central Texas........................................          45
                  Virginia...................................................          42
                  Kansas City, Missouri/Kansas...............................          33
                  Washington, D.C. (Maryland, Virginia)......................          29
                  San Diego/Southern California..............................          20
                  Las Vegas/Reno, Nevada.....................................          15
                  Albuquerque, New Mexico....................................           8
                                                                             ------------------
                                                                                      424
                                                                             ==================
```

(omitted)

```
<TYPE>EX-10
<SEQUENCE>4
<FILENAME>form10kexhf_032905.htm
<DESCRIPTION>EXHIBIT 10.2
<TEXT>
<HTML>
<HEAD>
<TITLE>Exhibit 10.2</TITLE>
```

Excerpt 2. BANK OF HAWAII CORP, CIK: 0000046195, February 28, 2012

Text format for machine processing:

```
<DIV ALIGN="CENTER"><TABLE width="100%"  BORDER=0 CELLSPACING=0 CELLPADDING=0>
<TR><!-- TABLE COLUMN WIDTHS SET -->
<TD WIDTH="" style="font-family:times;"></TD>
<TD WIDTH="12pt" style="font-family:times;"></TD>
<TD WIDTH="6pt" ALIGN="RIGHT" style="font-family:times;"></TD>
<TD WIDTH="42pt" style="font-family:times;"></TD>
<TD WIDTH="12pt" style="font-family:times;"></TD>
<TD WIDTH="6pt" ALIGN="RIGHT" style="font-family:times;"></TD>
<TD WIDTH="42pt" style="font-family:times;"></TD>
<TD WIDTH="12pt" style="font-family:times;"></TD>
<TD WIDTH="6pt" ALIGN="RIGHT" style="font-family:times;"></TD>
<TD WIDTH="42pt" style="font-family:times;"></TD>
<TD WIDTH="12pt" style="font-family:times;"></TD>
<!-- TABLE COLUMN WIDTHS END --></TR>

<TR VALIGN="BOTTOM">
<TH COLSPAN=4 ALIGN="LEFT" style="font-family:times;"><FONT SIZE=2><B>Discount Rate Sensitivity Analysis</B></FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=2> </FONT></TH>
<TH COLSPAN=5 ALIGN="CENTER" style="font-family:times;"><FONT SIZE=2><B>Table 1</B></FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=2> </FONT></TH>
</TR>
<TR style="font-size:1.5pt;" VALIGN="BOTTOM">
<TH COLSPAN=10 ALIGN="CENTER" style="font-family:times;border-bottom:solid #000000 1.0pt;"> </TH>
<TH style="font-family:times;"> </TH>
</TR>
<TR VALIGN="BOTTOM">
<TH ALIGN="LEFT" style="font-family:times;"><FONT SIZE=1> </FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
<TH COLSPAN=8 ALIGN="CENTER" style="font-family:times;border-bottom:solid #000000 1.0pt;"><FONT SIZE=1><B>Impact of </B></FONT></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
</TR>
<TR VALIGN="BOTTOM">
<TH ALIGN="LEFT" style="font-family:times;"><FONT SIZE=1>(dollars in thousands)</FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
<TH COLSPAN=2 ALIGN="RIGHT" style="font-family:times;"><FONT SIZE=1><B>Base<BR>
Discount<BR>
Rate</B></FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
<TH COLSPAN=2 ALIGN="RIGHT" style="font-family:times;"><FONT SIZE=1><B>Discount<BR>
Rate<BR>
25 Basis<BR>
Point<BR>
Increase</B></FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
<TH COLSPAN=2 ALIGN="RIGHT" style="font-family:times;"><FONT SIZE=1><B>Discount<BR>
Rate<BR>
25 Basis<BR>
Point<BR>
Decrease</B></FONT><BR></TH>
<TH style="font-family:times;"><FONT SIZE=1> </FONT></TH>
</TR>
<TR style="font-size:1.5pt;" VALIGN="BOTTOM">
<TH COLSPAN=10 ALIGN="CENTER" style="font-family:times;border-bottom:solid #000000 1.0pt;"> </TH>
<TH style="font-family:times;"> </TH>
</TR>
```

HTML as in a web browser (for the reader's convenience, the following picture shows the contents of the above scripts if shown as an HTML in a web browser[5]):

Discount Rate Sensitivity Analysis	**Table 1**		
		Impact of	
(dollars in thousands)	**Base Discount Rate**	**Discount Rate 25 Basis Point Increase**	**Discount Rate 25 Basis Point Decrease**
2011 Net Periodic Benefit Cost	5.75%	$ (220)	$ 219
Benefit Plan Obligations as of December 31, 2011	5.04%	(3,514)	3,678
Estimated 2012 Net Periodic Benefit Cost	5.04%	(32)	16

[5]From human perspectives, Excerpt 2 in a web browser is similar to Excerpt 1; From machine perspectives, it is much easier to process the text format of Excerpt 2 than Excerpt 1

Table IA.1 Sensitivity Check: Alternative Definition

This table examines the relation between the machine readability of a firm's filing and the machine downloads of the firm's past filings using alternative definitions. *Total Downloads* is the natural logarithm of the sum of *Machine Downloads* and *Other Downloads* (before taking the natural logarithm for both variables). *% Machine Downloads* is the ratio of Machine Downloads to Total Down- loads (without taking the natural logarithm for both variables) *Machine Downloads (Alt.)* and *Other Downloads (Alt.)* are alternative definitions of *Machine Downloads* and *Other Downloads* based on a criterion to classify machine visits in Loughran and McDonald (2017). Control variables include *Lagged Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. Variables are defined in Appendix A. In all panes, the t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*PCA Machine Readability*		*Machine Readability*			
Machine Downloads	0.131***	0.162***				
	(11.18)	(16.14)				
Other Downloads	-0.047***	-0.046***				
	(-4.75)	(-5.88)				
% Machine Downloads			0.121***	0.173***		
			(3.91)	(6.39)		
Total Downloads			0.053***	0.074***		
			(10.27)	(16.26)		
Machine Downloads (Alt.)					0.052***	0.064***
					(9.51)	(13.72)
Other Downloads (Alt.)					-0.010	-0.000
					(-1.51)	(-0.05)
Observations	139,436	139,330	150,377	150,298	150,425	150,346
R-squared	0.089	0.336	0.084	0.357	0.084	0.357
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Table IA.2 Sensitivity Check: Lagged Machine Downloads

This table reproduces the main results using one-year lagged *Machine Download*, which measures the expected machine readership of a filing. *Machine Readability* measures the ease at which a filing can be processed by an automated program. *LM – Harvard Sentiment* measures the difference in sentiments based on Loughran-McDonald finance-related negative words and Harvard General Inquirer negative words. *Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before. Control variables include *Lagged Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable	*Machine Readability*		*LM – Harvard*	
Lagged Machine Downloads × *Post*			-0.064***	-0.069***
			(-7.32)	(-9.52)
Lagged Machine Downloads	0.041***	0.059***	0.004	0.005
	(7.00)	(11.91)	(0.63)	(1.08)
Observations	133,733	133,610	140,690	140,580
R-squared	0.073	0.341	0.221	0.578
Control Variables	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Table IA.3 Determinants of Machine Downloads

This table reports the determinants of *Machine Downloads*, which measures the expected machine readership of a filing. Variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			*Machine Downloads*		
Size	0.135***	0.139***	0.040***	0.139***	0.040***
	(40.29)	(45.62)	(7.05)	(45.62)	(7.05)
Tobin's Q	-0.048***	-0.066***	-0.022***	-0.066***	-0.022***
	(-9.41)	(-13.24)	(-3.38)	(-13.24)	(-3.38)
ROA	-0.011	-0.031***	-0.002	-0.031***	-0.002
	(-0.94)	(-2.68)	(-0.14)	(-2.68)	(-0.14)
Leverage	0.085***	0.122***	0.055***	0.122***	0.055***
	(6.58)	(9.39)	(3.37)	(9.39)	(3.37)
Growth	-0.078***	-0.068***	-0.024***	-0.068***	-0.024***
	(-13.69)	(-12.21)	(-3.63)	(-12.21)	(-3.63)
IndAdjRet	-0.847***	-0.729***	-0.322***	-0.729***	-0.322***
	(-15.75)	(-13.97)	(-6.00)	(-13.97)	(-6.00)
InstOwnership	-0.005	-0.024*	-0.026	-0.024*	-0.026
	(-0.32)	(-1.66)	(-1.24)	(-1.66)	(-1.24)
Log(#analyst)	-0.008	-0.008	-0.021***	-0.008	-0.021***
	(-1.52)	(-1.54)	(-2.92)	(-1.54)	(-2.92)
IdioVol	0.091***	0.060***	-0.062***	0.060***	-0.062***
	(6.07)	(4.32)	(-4.37)	(4.32)	(-4.37)
Turnover	0.022***	0.019***	0.022***	0.019***	0.022***
	(13.20)	(12.08)	(12.11)	(12.08)	(12.11)
Segment	0.007***	0.007***	0.007***	0.007***	0.007***
	(6.81)	(6.97)	(3.89)	(6.97)	(3.89)
AI Hedge Fund				0.728***	0.417**
				(4.52)	(2.54)
Observations	171,296	171,296	171,234	171,296	171,234
R-squared	0.924	0.926	0.941	0.926	0.941
Firm FE	No	No	Yes	No	Yes
Industry FE	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes

Table IA.4 Sensitivity Check: Standard Errors

This table reproduces the main results. *Machine Downloads* measures the expected machine readership of a filing. *Machine Readability* measures the ease at which a filing can be processed by an automated program. *LM – Harvard Sentiment* measures the differential sentiments between Loughran-McDonald finance-related and Harvard General Inquirer negative words. *Time to the First (Directional) Trade* is the length of time between the EDGAR publication time stamp and the first (directional) trade of the issuer's stock since the publication. Panel A and Panel C report standard errors double clustered by industry × year and Panel B reports Driscoll and Kraay (1998) standard error. Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The t-statistics are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Standard errors double clustered by Industry × Year: Baseline results

	(1)	(2)	(3)	(4)
Dependent Variable	*Machine Readability*		*LM – Harvard Sentiment*	
Machine Downloads × Post			-0.072***	-0.079***
			(-6.94)	(-10.46)
Machine Downloads	0.061***	0.078***	-0.007	-0.011**
	(8.11)	(10.62)	(-1.13)	(-2.56)
Observations	150,425	150,346	158,578	158,515
R-squared	0.084	0.357	0.217	0.568
Control Variables	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Panel B: Driscoll and Kraay's standard errors: Baseline results

	(1)	(2)	(3)	(4)
Dependent Variable	*Machine Readability*		*LM – Harvard Sentiment*	
Machine Downloads × Post			-0.072***	-0.079***
			(-6.60)	(-7.32)
Machine Downloads	0.060***	0.078***	-0.007	-0.011*
	(4.97)	(5.76)	(-1.08)	(-1.90)
Observations	150,425	150,346	158,578	158,515
R-squared	0.084	0.357	0.217	0.568
Control Variables	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Panel C: Standard errors double clustered by Industry × Year: Time to trade

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Time to the First Trade				Time to the First Directional Trade			
Machine Downloads	-8.353***	-4.857*	-7.347**	-3.398	-12.365***	-7.540***	-12.374***	-7.258***
	(-2.83)	(-1.81)	(-2.41)	(-1.23)	(-4.69)	(-3.09)	(-4.46)	(-2.84)
Machine Downloads ×			-3.761***	-3.887***			-2.815**	-2.127**
			(-3.09)	(-3.33)			(-2.35)	(-2.03)
Machine Readability			-6.540	-5.980			-5.695	-8.709*
			(-1.18)	(-1.06)			(-1.09)	(-1.74)
Other Downloads	15.342***	3.499	15.151***	1.304	13.961***	3.885*	13.436***	2.336
	(7.46)	(1.61)	(7.08)	(0.58)	(7.20)	(1.90)	(6.64)	(1.10)
Observations	161,749	161,664	144,281	144,193	161,749	161,664	144,281	144,193
R-squared	0.116	0.269	0.118	0.272	0.120	0.285	0.122	0.286
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table IA.5 Linguistic Complexity and Technical Obfuscation

This table reports the relation between linguistic complexity and technical obfuscation of filings. *MR Upgrade* indicates an upgrade event, i.e., when a filing incurs a one standard deviation increase over the previous-year average *Machine Readability*, where *Machine Readability* measures the ease at which a filing can be processed by an automated program. *FileSize* is the document size of 10-K or 10-Q filing following Loughran and McDonald (2014). *ComplexWords* is the ratio of the number of complex words (i.e., words with at least three syllables) to the total number of words in a filing following Kim, Wang, and Zhang (2019). Control variables include *Lagged Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)
Dependent Variable	*MR Upgrade*	
FileSize	-0.004**	
	(-2.50)	
ComplexWords		-0.031***
		(-22.67)
Observations	135,057	135,057
R-squared	0.144	0.149
Control Variables	Yes	Yes
Firm FE	No	No
Industry FE	Yes	Yes
Year FE	Yes	Yes

Table IA.6 Effects of Machine Downloads: Time to Directional Quote Change

This table examines the relation between the time to the first directional quote change after a firm's filing is publicly released and the machine downloads of the firm's past filings, and how the machine readability of the filings affects such a relation. *Machine Downloads* measures the expected machine readership of a filing. *Machine Readability* measures the ease at which a filing can be processed by an automated program. A directional quote change is defined as an increase (decrease) in the ask (bid) price if the price at the end of the 15^{th} minute post filing is higher (lower) than the latest price prior to filing. Control variables include *Other Downloads* , *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variables	*Time to First Directional Quote Change*			
Machine Downloads	-6.267*	-3.752	-7.470**	-5.017
	(-1.92)	(-1.25)	(-2.22)	(-1.63)
Machine Downloads × *Machine Readability*			-2.111	-1.955
			(-1.35)	(-1.38)
Machine Readability			-6.941	-5.364
			(-1.01)	(-0.79)
Observations	161,119	161,030	143,689	143,597
R-squared	0.094	0.225	0.092	0.223
Control Variables	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Table IA.7 Sensitivity Check: Controlling Negative Events

This table reports the impact of the publication of Loughran and McDonald (2011) on the relation between the negative sentiment of a firm's filing and the machine downloads of the firm's past filings, after controlling for negative events. *Machine Download* measure the expected machine readership of a filing. *LM – Harvard Sentiment* measures the difference in sentiments based on Loughran-McDonald finance-related negative words and Harvard General Inquirer negative words. *Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before. *Misstatement* is an indicator variable equal to one if a firm has a misstatement identified by SEC AAERs (Accounting and Auditing Enforcement Releases) in a year, and zero otherwise. *Bankruptcy* is an indicator variable equal to one if a firm files a Chapter 7 or Chapter 11 bankruptcy in a year, and zero otherwise. *Forced CEO Turnover* is an indicator variable equal to one if a firm has a forced CEO turnover, as defined in Peters and Wagner (2014). Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The t-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)
Dependent Variable	*LM – Harvard Sentiment*	
Machine Downloads × Post	-0.072***	-0.078***
	(-6.92)	(-8.94)
Machine Download	-0.007	-0.011**
	(-1.12)	(-2.44)
Misstatement	0.168***	0.053
	(3.67)	(1.39)
Bankruptcy	0.177**	0.080
	(2.49)	(1.26)
Forced CEO Turnover	0.004	0.012
	(0.19)	(0.86)
Observations	158,578	158,515
R-squared	0.217	0.568
Control Variables	Yes	Yes
Firm FE	No	Yes
Industry FE	Yes	No
Year FE	Yes	Yes

Table IA.8 Sensitivity Check: Managing Sentiment in Response to BERT

This table examines the impact of the publication of BERT on the relation between the negative sentiment of a firm's 10-K filing, including Item 1 and Item 7, and the firm's ownership of AI-equipped investment company. The dependent variable *BERT Sentiment* is defined as the number of negative sentences, scaled by the total number of sentences in Columns (1) and (2), and scaled by the total number of words in Columns (3) and (4), respectively. *AI Ownership* is a firm's aggregate ownership of AI-equipped investment company shareholders. *AI Talent Supply* measures the local talent supplies to a firm's institutional shareholders, weighted by their ownership; the local talent supply is the available workforce with IT degrees in the state where an investor is headquartered. Control variables include *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment*. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable	*BERT Sentiment* (Item 1&Item 7)			
	NegSent/TotalSent		NegSent/TotalWords	
AI Ownership × *Post-BERT*	-3.310*		-0.153**	
	(-1.92)		(-2.29)	
AI Talent Supply × *Post-BERT*		-0.576**		-0.025***
		(-2.34)		(-2.75)
Observations	6,465	6,696	6,465	6,696
R-squared	0.784	0.784	0.799	0.799
Control Variables	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes