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

Sean Cao[a] Wei Jiang[b] Baozhong Yang[a] Alan Zhang[a]

[a]J. Mack Robinson College of Business, Georgia State University

[b]Columbia Business School, ECGI, and NBER

NBER Economics of Artificial Intelligence Conference
September 24-25, 2020

Motivation

- Corporate disclosure communicates financial health, promotes the culture and brand, and engages a full spectrum of stakeholders.
- Warren Buffet's annual letters to shareholders of Berkshire Hathaway showcase Corporate American writing at its best – for human readers.
  - "Be fearful when others are greedy and greedy when others are fearful."
  - "When it's raining gold, reach for a bucket, not a thimble."

Motivation

- Corporate disclosure communicates financial health, promotes the culture and brand, and engages a full spectrum of stakeholders.
- Warren Buffet's annual letters to shareholders of Berkshire Hathaway showcase Corporate American writing at its best – for human readers.
  - "Be fearful when others are greedy and greedy when others are fearful."
  - "When it's raining gold, reach for a bucket, not a thimble."

The changing readership of disclosure

- A substantial amount of buying and selling of shares are triggered by assessment and recommendations made by robots and algorithms.
- Technology makes it feasible: Machine learning and natural language processing kits.
- The sheer volume of regulatory filings makes it inevitable.
  - EDGAR hosts 11 million filings by over 600,000 reporting entities using 478 unique form types. There were 1.5 billion unique requests via SEC.gov in 2016 alone (Bauguess, 2018).
  - The length of 10k increases by five times from 2005 to 2017, and the number of textual changes over previous filings increases by over 12 times (Cohen, Malloy, and Nguyen, 2020).
- The SEC estimates that "as much as 85% of the documents visited are by internet bot" (Bauguess, 2018).
- Corporate disclosure needs to resonate with both human and machine readers.

Objectives of the study

- Research question: Whether and how public companies adjust the way they talk knowing that machines are listening.
  - Quantify and connect expected AI reader base and machine-friendliness of disclosure documents.
  - Identify changes in writing patterns affecting "sentiment" and "tone" after the availability of new algorithms, notably Loughran and McDonald (2011).
  - An "out-of-the-sample" test on the machine-assessed voice emotional quality of conference calls.
- Connect and contribute to the growing literature on:
  - Information acquisition and dissemination via downloads of SEC filings. Bernard, Blackburne, and Thornock (2020), Cao, Du, Yang, and Zhang (2020), Chen, Cohen, Gurun, Lou, and Malloy (2020), and Crane, Crotty, and Umar (2020).
  - Assessing qualitative information using textual analyses and machine learning. Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), and Hanley and Hoberg (2010), Loughran and McDonald (2011), Da, Engelberg, and Gao (2011), Garcia (2013), Jegadeesh and Wu (2013), Ahern and Sosyura (2014), Jiang, Lee, Martin, and Zhou (2019), Huang, Tan, and Wermers (2020); a survey by Loughran and McDonald (2016).

A novel "feedback effect"

- While financial markets reflect firm fundamentals, the market perception also influences manager's information set and decision making (survey by Bond, Edmans, and Goldstein, 2012).
- A novel "feedback effect" from machine learning about firm fundamentals to corporate decisions: Encoded rules are at least partially transparent, observable, or reverse-engineerable, agents who are impacted by the decisions thus have the incentive to manipulate the inputs to machine-learning.
    - Reminiscent of Goodhart's (1975) Law and Lucas (1976) Critique on macro intervention.
    - Particularly pertinent to machine learning: The feedback effect is, by construction, absence in training samples.
    - Recently formalized as "strategic classification" in the machine learning theory (Hardt, Megiddo, Papadimittriou, and Wootters, 2016; Dong, Roth, Schutzman, and Waggoner, 2018; Milli, Miller, Dragan, and Hardt, 2019).
- Highlight the challenge on machine-learning to be "manipulation proof," i.e. anticipating the strategic behavior of informed agents without observing it in the training samples (Bjorkegren, Blumenstock, and Knight, 2020).

Data sample

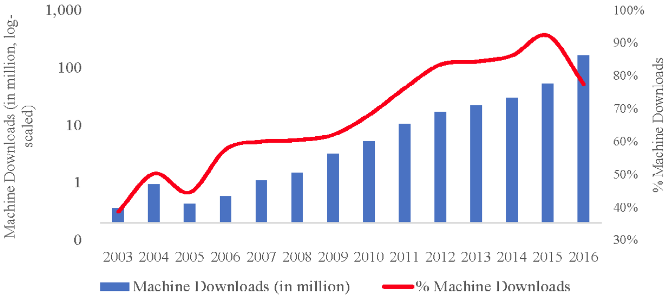
- The primary data source is the SEC EDGAR (regulatory filing archive) and the associated Log File Data Set (tracking requests and downloads). We focus on 10-K and 10-Q only for which firms have little discretion on the audience. Sample period spans 2003-2016.
- Firm characteristics are retrieved or constructed using information from standard WRDS databases: CRSP/Computser; Thomson Reuters, IBES.
- A total of 438,752 filings (119,135 10-K and 319,617 10-Q). After matching to CRSP/Compustat, 359,819 filings (90,437 10-K and 269,382 10-Q) filed by 13,763 unique CIKs.
- An additional sample of 43,462 conference call audios by 3,290 unique firms from EarningsCast between 2010 and 2016.

KEY VARIABLE: *Machine Downloads*

- The frequency of *Machine Downloads* of corporate filings as an upper bound as well as a proxy for the presence of "machine readers."
- Identify an IP address downloading more than 50 unique firms' filings on a given day, and requests that are attributed to web crawlers in the SEC Log File Data, as a machine (i.e., robot) visitor (Lee, Ma, and Wang, 2015). All remaining requests are labeled as "Other" requests.
- Aggregate machine requests and other requests, respectively, for each filing within seven days after its EDGAR posting.

Key variable: *Machine Downloads*

- The frequency of *Machine Downloads* of corporate filings as an upper bound as well as a proxy for the presence of "machine readers."
- Identify an IP address downloading more than 50 unique firms' filings on a given day, and requests that are attributed to web crawlers in the SEC Log File Data, as a machine (i.e., robot) visitor (Lee, Ma, and Wang, 2015). All remaining requests are labeled as "Other" requests.
- Aggregate machine requests and other requests, respectively, for each filing within seven days after its EDGAR posting.



Machine Downloads: Who's who in 2016

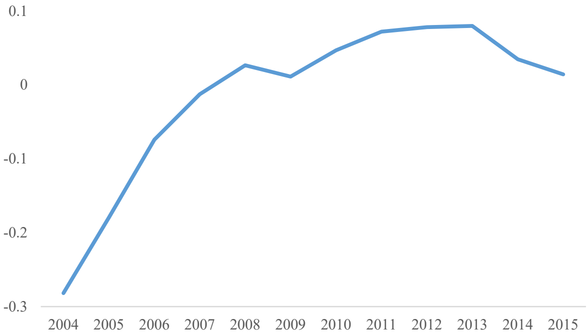
- Firms with the highest and lowest *Machine Downloads*:
  - Highest: CAMPBELL SOUP CO, TYSON FOODS INC, CISCO SYSTEMS INC, CLOROX CO, WESTERN DIGITAL CORP.
  - Lowest: GEORGIA POWER, COCA-COLA EUROPEAN PARTNERS, QVC INC, ALLSTATE LIFE INSURANCE CO, ARRIS INTERNATIONAL PLC.
- Firms with the highest and lowest *% Machine Downloads*:
  - Highest: MARATHON PETROLEUM CORP, GEORGIA POWER, CLOROX CO, DISCOVERY INC, DOLLAR TREE INC, MOTOROLA SOLUTIONS INC.
  - Lowest: CHINA AUTOMOTIVE SYSTEMS INC, FIFTH THIRD BANCORP, APPLE INC, ALPHABET INC, INTL BUSINESS MACHINES CORP, FACEBOOK INC, COTY INC, CARNIVAL CORPORATION.

Key variable: *Machine Readability*

- Measures the ease at which a filing can be "understood," i.e., processed and parsed, by an automated program.
- Summary of literature: The ease of (i) separating tables from text; (ii) extracting numbers from text; (iii) identifying the information contained in the table; (iv) inclusion of all needed information without relying on external exhibits; and (iv) proportion of characters that are standard ASCII characters.
- Take the average (or PCA) of the standardized component statistics.

Key variable: *Machine Readability*

- Measures the ease at which a filing can be "understood," i.e., processed and parsed, by an automated program.
- Summary of literature: The ease of (i) separating tables from text; (ii) extracting numbers from text; (iii) identifying the information contained in the table; (iv) inclusion of all needed information without relying on external exhibits; and (iv) proportion of characters that are standard ASCII characters.
- Take the average (or PCA) of the standardized component statistics.



EXAMPLE OF HIGH *Machine Readability*

VIASAT INC, CIK 0000797721, May 25, 2012

```
<TABLE CELLSPACING="0" CELLPADDING="0" WIDTH="68%" BORDER="0" STYLE="BORDER-COLLAPSE:COLLAPSE" ALIGN="center">


<TR>
<TD WIDTH="84%"></TD>
<TD VALIGN="bottom" WIDTH="10%"></TD>
<TD></TD>
<TD></TD>
<TD></TD></TR>
<TR>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom" COLSPAN="2" ALIGN="center" STYLE="border-bottom:1px solid #000000"><FONT STYLE="font-family:Times New Roman" SIZE="1"><B>Amortization</B></FONT><
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD></TR>
<TR>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom" COLSPAN="2" ALIGN="center"><FONT STYLE="font-family:Times New Roman" SIZE="1"><B>(In thousands)</B></FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD></TR>


<TR BGCOLOR="#cceeff">
<TD VALIGN="top"> <P STYLE="margin-left:1.00em; text-indent:-1.00em"><FONT STYLE="font-family:Times New Roman" SIZE="2">Expected for fiscal year 2013</FONT></P></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom"><FONT STYLE="font-family:Times New Roman" SIZE="2">$</FONT></TD>
<TD VALIGN="bottom" ALIGN="right"><FONT STYLE="font-family:Times New Roman" SIZE="2">15,592</FONT></TD>
<TD NOWRAP VALIGN="bottom"><FONT STYLE="font-family:Times New Roman" SIZE="2">  </FONT></TD></TR>
```

The filing has tags for table, row and data cell. It does not refer to any external exhibit.

Example of low *Machine Readability*

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
                                                           ===================
```

The "table" is surrounded by text rather than enclosed with the "<Table>... </Table>" tags.

(omitted)

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

It refers to an external exhibit (i.e., "form10kexhf_032905.htm"), which is not included in the filing.

Machine Readability & Machine Downloads POSITIVELY RELATED

The positive relation holds for both the volume and representation of machine downloads; and for various specifications of machine readability.

Dependent Variable	*Machine Readability*					*Residual MR*		*LM (2017) MD*	
Machine Downloads	0.075***	0.060***	0.078***			0.060***	0.078***	0.052***	0.064***
	(17.45)	(10.33)	(15.93)			(10.33)	(15.93)	(9.51)	(13.72)
Other Downloads	0.002	-0.007	-0.006			-0.007	-0.006	-0.010	0.000
	(0.47)	(-1.44)	(-1.33)			(-1.44)	(-1.33)	(-1.51)	(-0.05)
% Machine Downloads				0.121***	0.173***				
				(3.91)	(6.39)				
Total Downloads				0.053***	0.074***				
				(10.27)	(16.26)				
Observations	199,241	150,425	150,346	150,377	150,298	150,425	150,346	150,425	150,346
R-squared	0.363	0.084	0.357	0.084	0.357	0.002	0.005	0.084	0.357
Controls included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	No	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Principal and individual components of *Machine Readability*

All submetrics of *Machine Readability*, individually or the first principal component of all, bear a significant and positive relation to *Machine Downloads*.

Dependent Variable	*PCA MR*		*Table Extraction*	*Number Extraction*	*Table Format*	*Self-Containedness*	*Standard Characters*
Machine Downloads	0.132***	0.162***	0.051***	0.028***	0.026***	0.161***	0.125***
	(11.20)	(16.15)	(6.02)	(3.47)	(2.88)	(21.80)	(14.68)
Other Downloads	-0.046***	-0.046***	0.018**	-0.011	0.022**	-0.036***	-0.040***
	(-4.66)	(-5.85)	(2.37)	(-1.49)	(2.51)	(-6.69)	(-6.08)
Observations	139,436	139,330	149,484	150,346	149,484	150,245	140,061
R-squared	0.089	0.336	0.471	0.389	0.439	0.306	0.344
Controls included	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	No	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Cross validation of *Machine Readability & Machine Downloads*

- Do *Machine Downloads & Machine Readability* measure the extent of machine readership and machine readability?
- If they do, high *Machine Downloads* be would be associated with faster trades after filings postings, and more so when *Machine Readability* is high.
- Conduct a duration analysis using TAQ high-frequency data in which "time to trade" is the dependent variable (Bolandnazar, et al., 2020).
  - *Time to the First Trade* is defined as the number of seconds between the EDGAR publication time stamp and the first trade of the issuer's stock.
  - *Time to the First Directional Trade* is defined as the number of seconds between the EDGAR publication time stamp and the first profitable trade based on the price change during the 15-minute window afterwards.
  - Control variables include firm fundamental characteristics (e.g., size and Tobin's q) and those related to the trading environment (e.g., analyst coverage and share turnover).

Speed to trade and machine readers

Machine Downloads is significantly associated with faster trades post filing, and significantly more so when *Machine Readability* is high.

Dependent Variabile	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)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
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
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
Adjusted R-squared	0.114	0.241	0.116	0.242	0.119	0.257	0.121	0.257

Manage sentiment with human and machine readers

- Corporate disclosure is usually positively biased (Loughran and McDonald, 2011; Kothari, Shu, and Wysocki, 2009).
- Representation of "positive" and especially "negative" based on the Harvard Psychosociological Dictionary (Harvard-IV-4 TagNeg (H4N) file) provides additional information about firm outcomes and stock returns.
  - It serves as an objective standard to analyze the sources and consequences of tones and sentiments in corporate disclosures and new media as perceived by the general, mostly human readership (Tetlock, 2007, Tetlock, Saar-Tsechansky, and Macskassy, 2008, Hanley and Hoberg, 2010).
- Loughran and McDonald (2011) presented a specialized dictionary of positive/negative and tone words that fits the unique text of financial situations, which has been feeding into algorithms.
- "Sentiment" is defined as the representation of "negative" words in the documents.
- Differential tone management with respect to the two dictionaries is informative about "writing for machines."

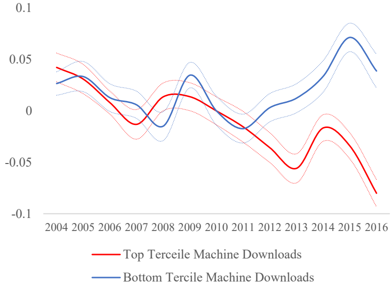
Harvard and LM (2011) Sentiments

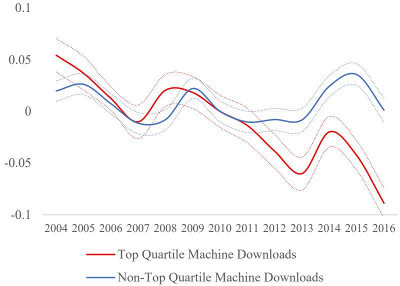
The Loughran and McDonald (2011), but not the Harvard, sentiment adapts to potential machine readership after 2011.

Dependent Variable	*LM Sentiment*		*Harvard Sentiment*		*LM – Harvard Sentiment*	
Machine Downloads × Post	-0.062***	-0.050***	0.01	0.029***	-0.072***	-0.079***
	(-4.98)	(-4.99)	(0.76)	(2.65)	(-6.95)	(-8.94)
Machine Downloads	-0.009	-0.019***	-0.002	-0.008	-0.007	-0.011**
	(-1.18)	(-3.72)	(-0.23)	(-1.43)	(-1.17)	(-2.46)
Observations	158,578	158,515	158,578	158,515	158,578	158,515
R-squared	0.241	0.632	0.208	0.59	0.217	0.568
Controls included	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Parallel pre-trends of *LM - Harvard*

Firms with high expected machine downloads differentially avoid LM-negative words relative to the Harvard-negative words, only after the publication of LM (2011), the exact timing of which is quasi-random.





Other tones developed in LM (2011)

- *Litigious* words (such as "claimant" and "tort") reflect a litigious environment.
- *Uncertain* words (such as "approximate" and "contingency") capture a general notion of imprecision.
- *Weak Modal* (such as "possibly" and "could") and *Strong Modal* (such as "always" and "must") words convey levels of confidence.
- Measured as the ratio of each category of words to the length of the filing.
- A high level of each of the four tones predicts one or more of negative outcomes: More "material weakness," fraud, and law suits; and is met with lower short-term stock return.
- Do managers avoid these words after the dictionary became publicly known?

Tone for machines

Firms avoid all four categories of tone words significantly more after the public knowledge of their impact.

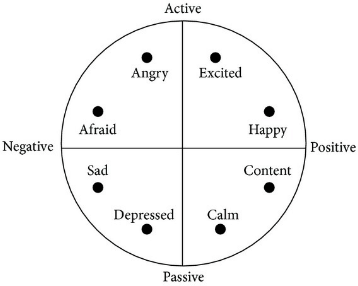
Dependent Variable	*Litigious*		*Uncertainty*		*Weak Modal*		*Strong Modal*	
Machine Downloads $\times$ *Post*	-0.056***	-0.057***	-0.016**	-0.021***	-0.028***	-0.034***	-0.008***	-0.007***
	(-5.38)	(-6.02)	(-2.01)	(-3.49)	(-4.85)	(-8.86)	(-4.39)	(-4.39)
Machine Downloads	0.011*	0.007	-0.006	-0.009***	-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	0.6	0.238	0.624	0.277	0.571
Controls included	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

Managing audio tones: An out-of-the-sample test

- Starting around 2008, voice analytic software (e.g., Layered Voice Analysis (LVA)) has gained popularity among investors looking for an edge in information processing.
- Such software has enabled researchers to study the vocal expressions of managers and their implications on capital markets (Mayew and Ventakachalam, 2012; and Hu and Ma, 2020).
- Is there a feedback effect to how managers talk?
- Explore a sample of 43,462 conference call speeches from 3,290 unique companies during 2010–2016.
- Two key measures based on the existent literature: Emotional *Valence* and *Arousal* correspond to positivity and excitedness of voices.
- Hu and Song (2020) showed that venture capitalists are more likely to invest in start-ups whose founders give pitches that are rated high in both.

EXAMPLE OF HIGH AND LOW *Emotion Valence/Arousal*

Valence and Arousal in a 2D Cartesian Coordinates system



Play high valence

Play high arousal

Play low valence

Play low arousal

How to talk to machines

Managers talk with higher valence and, to a lesser degree, higher arousal, when there are more machines expected in the audience.

Dependent Variable	*Emotion Valence*			*Emotion Arousal*		
Machine Downloads	0.043***	0.035***	0.042***	0.004*	0.003	0.005**
	(11.40)	(8.13)	(11.14)	(1.79)	(0.94)	(2.28)
Other Downloads	-0.017***	-0.014***	-0.017***	-0.006***	0.000	-0.006***
	(-5.74)	(-4.32)	(-5.67)	(-3.65)	(0.19)	(-3.71)
Observations	43,336	41,340	41,224	43,336	41,340	41,224
R-squared	0.389	0.189	0.383	0.395	0.132	0.395
Controls included	No	Yes	Yes	No	Yes	Yes
Company FE	Yes	No	Yes	Yes	No	Yes
Industry FE	No	Yes	No	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Conclusion

- Corporate disclosure in writing and speaking has been reshaped by machine readership employed by algorithmic traders and quantitative analysts.
- Increasing AI readership motivates firms to prepare filings that are more friendly to machine parsing and processing.
- Firms adapt sentiment and tone management to evolving algorithms.
- The feedback effect from technology calls for more studies to understand the induced behavior by AI in financial economics in order to have manipulation-proof algorithms.