



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

---

Year : 2023

## THREE ESSAYS ON DYNAMIC MEDIA EXPOSURE AND THE RELATED METHODOLOGICAL ISSUES OF CONTROL VARIABLE SELECTION AND AUTOMATED SENTIMENT ANALYSIS

Mändli Fabian

Mändli Fabian, 2023, THREE ESSAYS ON DYNAMIC MEDIA EXPOSURE AND THE RELATED METHODOLOGICAL ISSUES OF CONTROL VARIABLE SELECTION AND AUTOMATED SENTIMENT ANALYSIS

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB\_8421FBB3F1701

### **Droits d'auteur**

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

### **Copyright**

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

---

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES  
DÉPARTEMENT STRATÉGIE, GLOBALISATION ET SOCIÉTÉ

**THREE ESSAYS ON DYNAMIC MEDIA EXPOSURE  
AND THE RELATED METHODOLOGICAL ISSUES OF  
CONTROL VARIABLE SELECTION AND AUTOMATED  
SENTIMENT ANALYSIS**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales  
de l'Université de Lausanne

pour l'obtention du grade de  
Doctorat en Management

par

Fabian Benjamin MÄNDLI

Directeur de thèse  
Prof. Jean-Philippe Bonardi

Co-directrice de thèse  
Dr. Estefania Amer Maistriau

Jury

Prof. Boris Nikolov, Président  
Prof. Patrick Haack, expert interne  
Prof. Michalis Vlachos, expert interne  
Prof. Michael Etter, expert externe

LAUSANNE  
2023





UNIL | Université de Lausanne

---

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES  
DÉPARTEMENT STRATÉGIE, GLOBALISATION ET SOCIÉTÉ

**THREE ESSAYS ON DYNAMIC MEDIA EXPOSURE  
AND THE RELATED METHODOLOGICAL ISSUES OF  
CONTROL VARIABLE SELECTION AND AUTOMATED  
SENTIMENT ANALYSIS**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales  
de l'Université de Lausanne

pour l'obtention du grade de  
Doctorat en Management

par

Fabian Benjamin MÄNDLI

Directeur de thèse  
Prof. Jean-Philippe Bonardi

Co-directrice de thèse  
Dr. Estefania Amer Maistriau

Jury

Prof. Boris Nikolov, Président  
Prof. Patrick Haack, expert interne  
Prof. Michalis Vlachos, expert interne  
Prof. Michael Etter, expert externe

LAUSANNE  
2023

# IMPRIMATUR

La Faculté des hautes études commerciales de l'Université de Lausanne autorise l'impression de la thèse de doctorat rédigée par

**Fabian Mändli**

intitulée

*Three Essays on Dynamic Media Exposure and the Related Methodological Issues of Control Variable Selection and Automated Sentiment Analysis*

sans se prononcer sur les opinions exprimées dans cette thèse.

Lausanne, le 11.09.2023



Professeure Marianne Schmid Mast, Doyenne



## Thesis Committee

Prof. Jean-Philippe Bonardi  
Full Professor, University of Lausanne (Switzerland)  
Thesis supervisor

Dr. Estefania Amer Maistriau  
Senior Lecturer, University of Lausanne (Switzerland)  
Thesis co-supervisor

Prof. Patrick Haack  
Full Professor, University of Lausanne (Switzerland)  
Internal expert

Prof. Michalis Vlachos  
Full Professor, University of Lausanne (Switzerland)  
Internal expert

Prof. Michael Etter  
Associate Professor, King's College London (United Kingdom)  
External expert



University of Lausanne  
Faculty of Business and Economics

Ph.D. in Management

I hereby certify that I have examined the doctoral thesis of

**Fabian Benjamin MÄNDLI**

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members  
made during the doctoral colloquium  
have been addressed to my entire satisfaction.

Signature:



Date: 28.08.2023

Prof. Jean-Philippe BONARDI  
Thesis supervisor





University of Lausanne  
Faculty of Business and Economics

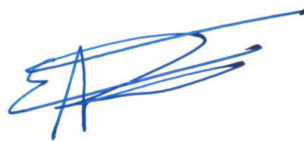
Ph.D. in Management

I hereby certify that I have examined the doctoral thesis of

**Fabian Benjamin MÄNDLI**

and have found it to meet the requirements for a doctoral thesis.  
All revisions that I or committee members  
made during the doctoral colloquium  
have been addressed to my entire satisfaction.

Signature:

A handwritten signature in blue ink, consisting of several overlapping loops and lines, positioned to the right of the 'Signature:' label.

Date: August 25, 2023

Dr. Estefania AMER MAISTRIAU  
Thesis co-supervisor



University of Lausanne  
Faculty of Business and Economics

Ph.D. in Management

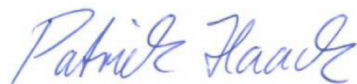
I hereby certify that I have examined the doctoral thesis of

**Fabian Benjamin MÄNDLI**

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members  
made during the doctoral colloquium  
have been addressed to my entire satisfaction.

Signature:



Date: August 24, 2023

Prof. Patrick HAACK  
Internal expert



University of Lausanne  
Faculty of Business and Economics

Ph.D. in Management

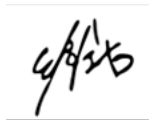
I hereby certify that I have examined the doctoral thesis of

**Fabian Benjamin MÄNDLI**

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members  
made during the doctoral colloquium  
have been addressed to my entire satisfaction.

Signature :



Date : 25 Aug 2023

Prof. Michalis VLACHOS  
Internal expert



University of Lausanne  
Faculty of Business and Economics

Ph.D. in Management


I hereby certify that I have examined the doctoral thesis of

**Fabian Benjamin MÄNDLI**

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members  
made during the doctoral colloquium  
have been addressed to my entire satisfaction.

Signature:

A handwritten signature in black ink, appearing to read 'Michael Etter', is written over a light blue rectangular stamp.

Date: 24.8.2023

Prof. Michael ETTER  
External expert





## Acknowledgements

I would like to express my inmost gratitude to my co-supervisors, Dr. Estefania Amer and Prof. Jean-Philippe Bonardi, for their professional and personal guidance during my doctoral studies. I am indebted for their advice regarding the world of academia, for helping me grasp and develop my strengths and for our many stimulating discussions we had over the years. I am additionally particularly thankful for the experience in the role of teaching assistant for Estefania Amer, where I was able to obtain a plethora of valuable skills related to teaching in an academic context, as well as to gain some first experience on my own.

A sincere thank you goes to my thesis committee, Prof. Michael Etter, Prof. Patrick Haack, and Prof. Michalis Vlachos. I appreciate that they took the time to evaluate my work and provided constructive and valuable comments.

I am also grateful to my co-author Prof. Mikko Rönkkö. Through his guidance and our collaboration, I have acquired skills to conduct methodological research and went through my first publication process.

I further thank the members of the SGS department as well as HEC, professors and colleagues. They have helped me to learn new skills, develop original ideas and refine my research over the years. A special thanks also goes to Pavla Le Moing and Bénédicte Moreira for their valuable support on endless occasions.

Finally, I am indebted to my family and friends. My partner, Fabienne Fend and my daughter Ruba have always ensured that I don't lose my head over research problems and showed me what is most important in life, which is our time together. I also thank my parents, Bernhard and Edith Mändli, who have provided moral support and wisdom. A last note goes to my valued friend Wouter van Minnen, who has helped me stay motivated in many interesting conversations over the years.

I am thankful for the memories of my brother, Dominik Mändli, who is no longer with us.

Fabian Mändli  
Lausanne, September 2023



## Table of Contents

Introduction	3
Chapter 1: <i>Dynamics of Negative Media Exposure and Firm Strategy</i>	13
Chapter 2: <i>To Omit or to Include? The Frugal and Prolific Perspectives on Control Variable Use</i>	63
Chapter 3: <i>Automated Sentiment Analysis vs. Manual Coding: Examining the Tone of News on Companies' Environmental Issues</i>	103
Conclusion	169



## **Introduction**

This thesis consists of three chapters. Chapter 1 is an empirical examination of dynamic negative media exposure in the context of firms' environmental issues. Two important methodological challenges encountered in this first chapter inspired the subsequent chapters: Chapter 2 is an examination and integration of the opposing few and many perspectives on control variable use in empirical research. Chapter 3 presents an investigation of different automated sentiment analysis techniques commonly used in management science and studies their performance compared to manual coding. In what follows I briefly present the motivation of each chapter as well as the respective research questions and address how the three chapters are related. Towards the end of this thesis document, I outline the main contributions and limitations of each chapter and present an overall conclusion.

### **Chapter 1: Dynamics of Negative Media Exposure and Firm Strategy**

The relationship of firms and the media is an important topic of investigation in management. Media coverage plays an important role in what the public and stakeholders learn about firms (Carroll & McCombs, 2003), as it represents a major legitimate source about their inner workings. Favorable media coverage consequently also constitutes an important strategic asset for firms (Deephouse, 2000). Media coverage can significantly affect performance and evaluations of firms (Graf-Vlachy et al., 2020), in particular dimensions of social evaluation such as reputation (e.g., Brammer & Pavelin, 2004; Wartick, 1992) or legitimacy (e.g., Brown & Deegan, 1998; Vergne, 2011). As social and environmental issues of firms are usually not directly observable for stakeholders and the public, the news media are also a central source of information regarding these topics. Finally, media reporting regarding these issues are also crucial to ideas such as the reputational halo effect (Godfrey et al., 2009), and the liability effect

(King & McDonnell, 2015).

We discuss that media exposure should be considered as a dynamic phenomenon, that does not only depend on what happens in the world (i.e., what firms do), or what is interesting for the audience of news outlets, but also on what has been reported in the press in the past. Others have pointed out the limited understanding of such dynamic media effects in management science (Graf-Vlachy et al., 2020). We extend the current conceptualization of relationship between firms and the media: By integrating insights from Communication Studies, we theorize which factors lead to time-dynamic reporting in the news regarding firms' environmental issues. Specifically, we argue that explicit strategic considerations as well as implicit factors regarding newsworthiness and storytelling of media organizations and journalists lead to dynamic effects in media coverage. We test this theory on a large hand-coded dataset of environmental news articles collected over a 16-year period from the 30 largest US firms.

This paper faced two important methodological challenges: First, the choice of control variables for the identification strategy is critical. While time-invariant characteristics at the firm and industry level are partialled out by using a fixed-effects estimator, we still needed to control for time-variant firm-level characteristics that influence the likelihood and the tone of media coverage of environmental issues. An important control variable we use is the environmental pillar score of MSCI ESG, which has been shown to be a solid indicator for environmental performance of firms (Semenova & Hassel, 2015). Further, we used several financial firm-level control variables that have been identified in the existing literature as important drivers of media exposure. For example, the capacity of firms to implement environmental measures, or to react to activist attacks in the media, are dependent upon financial resources (Eesley & Lenox, 2006), size (Brammer & Millington, 2006), and debt ratio (Adams & Hardwick, 1998). Even though we employed the control variables that we could identify as important based on prior theory, it was far from evident if we had included the right set of

controls or if we had for example included unnecessary ones that we could have dropped. On the one hand, it was essential for us to include the relevant alternative explanations for our findings. On the other hand, if we introduced too many control variables that might be redundant or irrelevant, we would sacrifice estimation power and potentially introduce additional variance into the model. Driven by these issues, chapter 2 investigates the few or many control variables perspectives proposed in the literature and tests their specific propositions with Monte Carlo simulations.

A second challenge we faced is related to the coding of the tone of the news articles. Since the database is at the core of the paper and it is used for the dependent as well as several independent variables, its quality is essential for chapter 1. We chose a manual coding approach and employed an extensive codebook, meticulous training with cross-validation. We assessed intercoder reliability between all three coders (Krippendorff's  $\alpha = .82$ ), which is appropriate according to the proposed standards (Krippendorff, 2004, p. 241). Yet, hand-coding the database of almost 18,000 news articles and press releases engendered a substantial investment in terms of resources, foremost it was time-consuming. We could have chosen an automated approach to process the news articles, which would have done the task within seconds compared to manual coding that took several years, yet it was unclear to us whether an automated approach would deliver results of sufficient quality. While automated approaches are attractive because they allow to process large amounts of data fast and at low cost, but also because in theory they should render studies more replicable, it is questionable if their results are reliable enough for research (Barberá et al., 2021; Boukes et al., 2020). We thus wanted to understand whether we could have used automated approaches for this task, also given that there has been rapid evolution in machine learning in recent years. Chapter 3 identifies the most commonly used approaches in management science, applies them to the news articles from chapter 1 and compares their results to the manual coding.



## **Chapter 2: To Omit or to Include? The Frugal and Prolific Perspectives on Control Variable Use**

In quantitative research, control variables allow to account for alternative explanations and consequently estimate causal effects. But how many and which control variables should be chosen? There are two opposing perspectives that answer this question differently in management research: Advocates of the frugal perspective argue that control variables should be used sparsely. Specific recommendations of this perspective are to omit control variables that are unrelated to the dependent variable and to avoid proxies, because there is an increased risk of introducing additional problems into a model (e.g., endogenous controls or noise) (Becker et al., 2016). The prolific perspective on the other hand argues the opposite, that control variables should be employed generously, to prevent omitted variable bias (Antonakis et al., 2010). While the prior bases its verdict on intuitive examples, the latter is grounded in econometric proofs. Despite their differences and the centrality of the topic, to date there has not been a systematic examination of both perspectives'. We study the impact of both perspectives in the literature and test their specific recommendations using Monte Carlo simulations, to understand the merits and drawbacks of the specific recommendations of each perspective. Based on these results, we provide an integration of the two perspectives that combines the merits of each.

## **Chapter 3: Automated Sentiment Analysis vs. Manual Coding: Examining the Tone of News on Companies' Environmental Issues.**

While we chose to code the news articles and press releases in chapter 1 manually, we could have also resorted to automated approaches. Such approaches are substantially faster, cheaper and could also yield more replicable results (Duriiau et al., 2007). Yet, there is disagreement on whether automated approaches produce results of sufficient quality. While some

have claimed that they are unreliable (Barberá et al., 2021; Boukes et al., 2020), and argue that human coding remains the “gold standard”, others have advocated for their increased use (Boumans & Trilling, 2018), and there is evidence that some approaches reach human-level accuracy (Hutto & Gilbert, 2014; Pang et al., 2002). Automated approaches have found their way into the research methods used in management science and have been also used in high-profile articles (e.g., Kölbel et al., 2017; Pfarrer et al., 2010), yet there is to date no systematic mapping and comparison of popular approaches used in management.

In chapter 3, I intend to close this gap. To understand which tools are being used in management research, I conduct a literature analysis in the highest ranked journals. I then test the performance of the most popular approaches for sentiment analysis by applying them to the hand-coded news articles used in the study of dynamic media effects (chapter 1). I also test two large language models, as they might be more capable to determine sentiment in complex texts such as news articles.

## References

- Adams, M., & Hardwick, P. (1998). An Analysis of Corporate Donations: United Kingdom Evidence. *Journal of Management Studies*, 35(5), 641–654.  
<https://doi.org/10.1111/1467-6486.00113>
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6), 1086–1120.  
<https://doi.org/10.1016/j.leaqua.2010.10.010>
- Barberá, P., Boydston, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19–42.  
<https://doi.org/10.1017/pan.2020.8>
- Becker, T. E., Atinc, G., Breugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers: Statistical Control in Correlational Studies. *Journal of Organizational Behavior*, 37(2), 157–167. <https://doi.org/10.1002/job.2053>
- Boukes, M., van de Velde, B., Araujo, T., & Vliegthart, R. (2020). What’s the tone? Easy doesn’t do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2), 83–104.  
<https://doi.org/10.1080/19312458.2019.1671966>
- Boumans, J. W., & Trilling, D. (2018). Taking stock of the toolkit. In M. Karlsson & H. Sjøvaag (Eds.), *Rethinking Research Methods in an Age of Digital Journalism* (1<sup>st</sup> ed., pp. 8–23). Routledge. <https://doi.org/10.4324/9781315115047-2>
- Brammer, S. J., & Millington, A. (2006). Firm size, organizational visibility and corporate philanthropy: An empirical analysis. *Business Ethics: A European Review*, 15(1), 6–18. <https://doi.org/10.1111/j.1467-8608.2006.00424.x>
- Brammer, S. J., & Pavelin, S. (2004). Building a Good Reputation. *European Management*

- Journal, 22(6), 704–713. <https://doi.org/10.1016/j.emj.2004.09.033>
- Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—A dual test of media agenda setting theory and legitimacy theory. *Accounting and Business Research*, 29(1), 21–41.  
<https://doi.org/10.1080/00014788.1998.9729564>
- Carroll, C. E., & McCombs, M. (2003). Agenda-setting effects of business news on the public's images and opinions about major corporations. *Corporate Reputation Review*, 6(1), 36–46. <https://doi.org/10.1057/palgrave.crr.1540188>
- Deephouse, D. L. (2000). Media Reputation as a Strategic Resource: An Integration of Mass Communication and Resource-Based Theories. *JOURNAL OF MANAGEMENT*, 26(6).
- Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods*, 10(1), 5–34.  
<https://doi.org/10.1177/1094428106289252>
- Eesley, C., & Lenox, M. J. (2006). Firm responses to secondary stakeholder action. *Strategic Management Journal*, 27(8), 765–781. <https://doi.org/10.1002/smj.536>
- Godfrey, P. C., Merrill, C. B., & Hansen, J. M. (2009). The relationship between corporate social responsibility and shareholder value: An empirical test of the risk management hypothesis. *Strategic Management Journal*, 30(4), 425–445.  
<https://doi.org/10.1002/smj.750>
- Graf-Vlachy, L., Oliver, A. G., Banfield, R., König, A., & Bundy, J. (2020). Media coverage of firms: Background, integration, and directions for future research. *Journal of Management*, 46(1), 36–69. <https://doi.org/10.1177/01492063198641>
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment

- analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- King, B. G., & McDonnell, M.-H. (2015). Good firms, good targets: The relationship among corporate social responsibility, reputation, and activist targeting. In K. Tsutsui & A. Lim (Eds.), *Corporate Social Responsibility in a Globalizing World* (pp. 430–454). Cambridge University Press. <https://doi.org/10.1017/CBO9781316162354.013>
- Kölbel, J. F., Busch, T., & Jancso, L. M. (2017). How Media Coverage of Corporate Social Irresponsibility Increases Financial Risk: Media Coverage of Corporate Social Irresponsibility. *Strategic Management Journal*, 38(11), 2266–2284. <https://doi.org/10.1002/smj.2647>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2<sup>nd</sup> ed). Sage. <https://lccn.loc.gov/2003014200>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques (arXiv:cs/0205070). arXiv. <http://arxiv.org/abs/cs/0205070>
- Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, 53(5), 1131–1152. <https://doi.org/10.5465/amj.2010.54533222>
- Semenova, N., & Hassel, L. G. (2015). On the validity of environmental performance metrics. *Journal of Business Ethics*, 132(2), 249–258. <https://doi.org/10.1007/s10551-014-2323-4>
- Vergne, J.-P. (2011). Toward a new measure of organizational legitimacy: Method, validation, and illustration. *Organizational Research Methods*, 14(3), 484–502. <https://doi.org/10.1177/1094428109359811>

Wartick, S. L. (1992). The relationship between intense media exposure and change in corporate reputation. *Business & Society*, 31(1), 33–49.

<https://doi.org/10.1177/000765039203100104>



# **Dynamics of Negative Media Exposure and Firm Strategy**

Fabian Mändli, Estefania Amer, and Jean-Philippe Bonardi, HEC Lausanne

## **Abstract**

Negative media exposure and the damaging impact it can have for firms is a core topic in the relationship between firms and the media. However, it is generally overlooked that critical media reports could have cumulative effects over time, leading to negative spirals that trigger even worse and long-lasting consequences, but also that firms might be able to prevent these spirals. In this paper, we develop a theory of dynamic media exposure, suggesting that negative news regarding a firm's environmental record increases the likelihood of subsequent negative exposure of the same type, threatening a firm's social evaluations beyond the short-run and potentially turning this firm into an environmental villain. However, we also show that firms can prevent such negative spirals by nurturing positive environmental news and emitting proactive press releases about their environmental progress. We test our hypotheses with a unique 16-year panel-dataset of media articles on the 30 largest US firms and find support for them. We discuss how these results broaden our understanding of firm-media interactions, and what they also imply regarding sustainability concerns.

**Keywords:** *social evaluations, media coverage, environmental issues*



## Introduction

The relationship of firms and the media is a central topic in management research (Graf-Vlachy et al., 2020). On the one hand, media coverage affects the publics' view of firms' or individual members of the latter and how these react to such coverage, thus media coverage influences evaluative constructs such as reputation (Brammer & Pavelin, 2004; Eisenecker & Schranz, 2011) and/or legitimacy (e.g., Bansal & Clelland, 2004; Brown & Deegan, 1998; Vergne, 2011). On the other hand, favorable media reporting constitutes an important strategic asset for firms (Deepphouse, 2000; Pollock & Rindova, 2003). Given its significance, firms as well as individuals associated with the firm strive to shape media coverage. While after negative reports, these aims are to limit damages to reputation (Amer & Bonardi, 2023) or to maintain legitimacy (Lamin & Zaheer, 2012), in calmer times firms actively work towards obtaining favorable media coverage (e.g., Rindova et al., 2006).

Reporting in the news media is found as being negatively biased against firms (Baron, 2005; Jonkman, Trilling, et al., 2020), which is consequential as this bias also affects audience perceptions (Soroka & McAdams, 2015). This bias is not solely explained by journalists that take on a watchdog stance and critically discuss societal actors such as firms (Johnson et al., 2005): Research in communication studies describes that factors embedded at the organizational, system and individual level in the news production contribute to this negative bias (Soroka & Carbone, 2016).

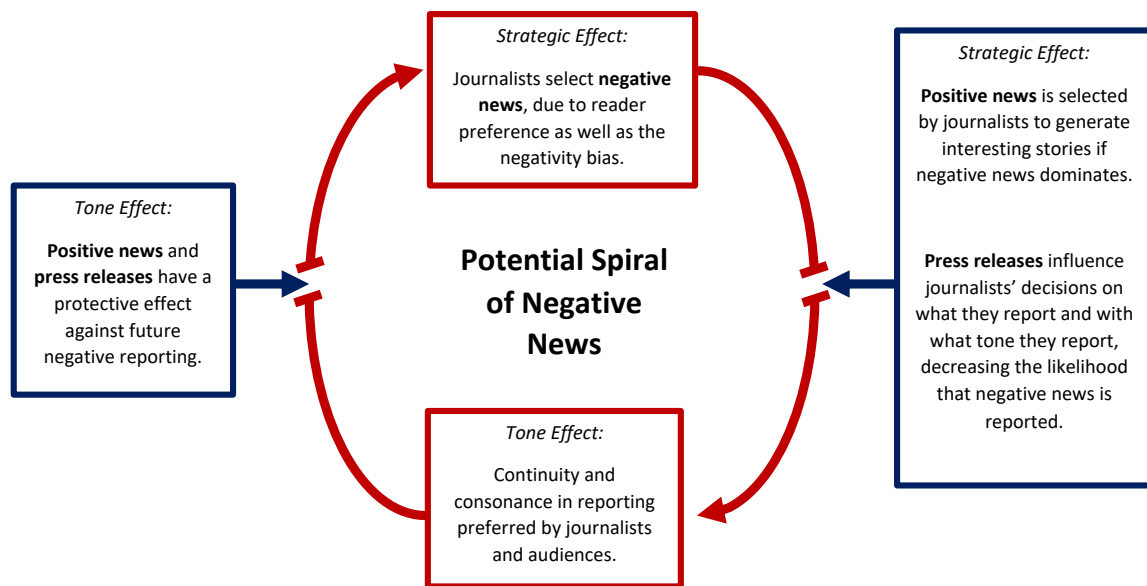
Reporting of environmental and social issues in the media is particularly critical for firms. First, much of what the public learns about these issues originates in the media, as direct observation is usually not possible (Thøgersen, 2006). This is consequential for example for social evaluations (Aerts & Cormier, 2009; Bansal & Clelland, 2004; Brammer & Pavelin, 2006), as it is the news media who determine to some extent what the public learns about which firms. Second, environmental and social issues may end up being reported in dramatic

narratives to attract scarce reader attention (Lovelace et al., 2022; Zavyalova et al., 2017), thus firms are at risk to be portrayed as villains in the news (Diermeier, 2011; Rindova et al., 2006). Fourth, the reputational threat media exposure of environmental and/or social issues poses to firms, serves as a strategic mechanism for activists and social movements to elicit concessions from the latter (McDonnell & King, 2013). Finally, media reporting is also regarded as a catalyst for change, by shaping the public sentiment that exerts pressure on firms (Baron & Diermeier, 2007).

Media exposure in the context of social and environmental issues is consequently likely to suffer from a manifest negativity bias, as activists and social movements target firms via the news media (Dyck et al., 2008), while at the same time journalists cover topics in dramatic stories where firms are often the villains (Diermeier, 2011). Moreover, vivid reports about environmental or social deeds of firms translate towards increased newsworthiness, also because the public is more conscious regarding these matters (Flammer, 2013). As news organizations aim to maximize market potential for their products, and journalists embed new reports in the context of the old, this heightens the risk of repeated negative coverage of firms. Negative media exposure on environmental and social issues could consequently become dynamic. If such negative coverage is sustained over time, firms end up being portrayed as villains: Through a combination of the *strategy-related effect*, linked to how media organizations strategically behave to secure audience interest, and the *tone-related effect* linked to social and cognitive factors of journalists, firms might end up in a spiral of negative news reporting, threatening social evaluation constructs such as reputation or legitimacy beyond the short run (cf. Figure 1 below).

**Figure 1**

*Spiral of Negative News and Disrupting Mechanisms.*



Given the importance of favorable media reporting, firms attempt to prevent such emerging negative dynamics, making negative spirals more fragile. The literature reports that after initial transgressions and related negative coverage, firm reaction reaches from countering statements such as press releases (e.g., Wickman, 2014; Zavyalova et al., 2012), strategic change (e.g., Bednar et al., 2013; Durand & Vergne, 2015; Piazza & Perretti, 2015), to retaliation against critical journalists (Westphal & Deephouse, 2011), to prevent additional criticism in the media. Yet, firms might actually also proactively work towards prevention of such negative spirals. We identify two different proactive mechanisms within the firm-media setting that have the potential to prevent negative dynamic exposure and consequently also diminish the risk of negative spirals.

First, while positive media reports are predominantly viewed as sustaining and/or repairing social approval of the firm (Deephouse, 2000; Pollock & Rindova, 2003), they might also provide a more direct buffer effect against negative news. With rising number of negative reports about a firm, publishing positive news about the same firm becomes increasingly

attractive for news outlets to sustain audience interest (Oliver et al., 2023; Soroka & Krupnikov, 2021). If positive news are reported, other journalists observing this might subsequently become more hesitant to publish negative news (Pollock et al., 2008). Indeed, favorable media reporting might also cumulatively translate to a protective effect (Zavyalova et al., 2016), as firms are given the “benefit of the doubt” in case of wrongdoing (Godfrey et al., 2009, p. 428). This should dampen the negative dynamics and consequently also prevent potential spirals of negative news; positive news should therefore have a protective effect against negative news spirals.

Second, firms usually engage in impression management tactics after transgression (Bansal & Clelland, 2004; McDonnell & King, 2013), even when media exposure has targeted their industry peers or their industry (Desai, 2011). Press releases, one tool firms use to manage impressions, have been mostly understood as working in a reactive manner (e.g., Pfarrer et al., 2008; Zavyalova et al., 2012). However, they could also proactively diminish the likelihood of subsequent negative media exposure, as press releases have the potential to shape media coverage (Carroll & McCombs, 2003; Maat, 2007). By making journalists and editors more reluctant to publish negative reports that would contradict information received through firm press releases, press releases might provide a similar protective effect against negative dynamics as positive news do. Due to both of these mechanisms, negative media spirals might not develop as easily and as often against firms, and even be rare, if these two mechanisms are able to counteract those that would plunge the firm into a downwards spiral.

In the remainder of the paper, we examine this process of dynamic negative media exposure, a topic where management research currently only has limited understanding (Graf-Vlachy et al., 2020). Our work is among the first to provide evidence for dynamics in news reporting about firms. This has important implications for management research: We contribute to an understanding of how firms could end up in spirals of negative news coverage, potentially

suffering much stronger and long-lasting implications than generally assumed in the literature. Moreover, we also show that firms proactively emit press releases and/or foster positive media reports to protect themselves from such negative media spirals. We thus also add an additional, proactive function of press releases to the literature, one where these statements from firms have a protective effect against subsequent negative exposure.

We explore these mechanisms empirically by using a database covering the 30 largest US-firms over a 16-year time period. Contrary to common practice, our database contains only a limited number of firms, this allowed us to gather detailed and precise information on each of them over a long period of time. Due to our manual pre-scanning process and the level of agreement between coders in the coding procedure of the articles, this database thus possesses high overall quality (e.g., it contains all the articles related to environmental issues and none that are unrelated to them) and allows us to observe a firm's environmental exposure over a long period of time, while simultaneously providing fine-grained monthly data.

Overall, we find empirical support for the idea that negative media spirals might be put in motion by dynamic negative news reporting, but also for the idea that positive news and press releases can break these negative spirals and proactively decrease additional negative exposure, which should be able to protect firms. We discuss the implications of our findings for the relationship of firms and the media and anchor our findings within the context of firms' media reputations.

## **Theory and Hypotheses**

### **Media Exposure and Newsworthiness**

The news media are an important source of information that influences the public's perceptions and therefore is a central component for example for the reputation of firms (Fombrun & Shanley, 1990; Pollock et al., 2019), or firm legitimacy (Deephouse & Carter, 2005; Vergne, 2011). First, they serve as disseminators of information from third parties such

as activists, governments, and organizations (Bonardi & Keim, 2005; Donsbach, 2012). Second, through their selection and evaluation of information, the news media themselves are autonomous actors (Eisenegger & Schranz, 2011). Indeed, much of what stakeholders and the public learn about firms comes from the news media (Carroll & McCombs, 2003). Issues that are not covered are absent from people's worldviews (Shoemaker & Vos, 2009). This is especially true for environmental issues, which by their nature often are not directly observable and thus remain undetected by the public (Thøgersen, 2006). Therefore, to learn about firms' environmental issues, the public relies to a large extent on news media reports.

Media exposure is described as “[...] the aggregated news reports relating to a specific firm within a prescribed period.” (Wartick, 1992, p. 34). Media exposure for example influences corporate governance policies towards environmental issues (Dyck & Zyngales, 2002) and, more importantly, has a substantial impact on firms' social evaluations (Breitinger & Bonardi, 2019; Deephouse & Suchman, 2008; Zavyalova et al., 2017), ultimately also affecting financial performance (Flammer, 2013; Vasi & King, 2012). Moreover, negative media coverage might also damage reputations of firm executives, because reports might discuss issues that are the results of the executives' decisions (Bednar, 2012; Bednar et al., 2013; Shipilov et al., 2019).

The more intense a firm's exposure in the news media is, and the greater the number of media outlets that are involved, the stronger the impact of this exposure on a firm's social evaluations (Eisenegger, 2005). This cumulative idea is related to the concept of media reputation, which is the “overall evaluation of a firm presented in the media”, as a valuable resource for firms (Deephouse, 2000, p. 1091). Moreover, present media attention on a firm depends on cumulative past media attention, regardless of the tone of the news (Pollock et al., 2008). Dramatic negative coverage of firm-related environmental issues in the media could attract more coverage of the same type, ultimately leading to a negative spiral where the public ends up viewing the firm as an environmental villain (Diermeier, 2011; Rindova et al., 2006).

However, the existing literature has not explored this type of cumulative phenomenon, which requires adopting a temporally dynamic and more fine-grained approach to the interaction of firms and the media.

Communication studies have explored how journalists, editors, and media organizations select the news that their outlets report on (Harcup & O'Neill, 2017; O'Neill & Harcup, 2009; Shoemaker & Vos, 2009). The assessment of the potential of a piece of information - its newsworthiness (Shoemaker, 2006) - is based on various factors including personal perception, economic constraints, characteristics of media organizations, the modalities of the event itself, social norms, and values of the public (Andrews & Caren, 2010; Caple & Bednarek, 2013). Consequently, a piece of information has to be judged as newsworthy at different levels of the media production system to become news (Shoemaker et al., 2001; Soroka, 2012). The more newsworthy a given piece of information is, the more likely it is going to end up being widely broadcast by the media and, as a result, the more likely it is to influence the information processing of the audience (Eilders, 2006).

According to Diermeier (2011), the likelihood that an issue is covered in the news (i.e. its newsworthiness) is largely driven by two forces: audience interest and societal importance, which depend on societal norms and values. Similarly, in Communication studies, the amount of deviance of a reported event and the amount of social significance determine newsworthiness (Shoemaker & Cohen, 2012). Therefore, the higher the audience interest and societal importance of a piece of news about a firm's environmental record, the more newsworthy it is, and the more likely it is to be reported in the news media.

## **The Consequences of Negative Media Exposure on Future Coverage**

The literature in Communication studies reports that negative news reports are more prevalent than positive ones (Niven, 2001; Soroka, 2006). This negativity bias is not explained by a higher frequency of negative events in the world, but by three mechanisms at different levels of news production and dissemination. Specifically, individual-level strategic selection by journalists or editors, organizational-level gatekeeping by media outlets, and system-level factors all lead to greater prominence of negative news in the overall news reporting (Soroka, 2006). Next, we discuss why each level of the news system favors negativity and also why some aspects of the media system lead to temporal dynamics of negative news.

### ***The strategy-related effect of negative news***

Negativism, damage, and failure are important characteristics of information that have been related to newsworthiness (Galtung & Ruge, 1965; Staab, 1990). First, the human brain is biologically hardwired to favor such negative traits (Rozin & Royzman, 2001). Indeed, bad information has been found to be processed more thoroughly in the brain than good information (Baumeister et al., 2001). Second, negative information is perceived as being more truthful than positive information (Hilbig, 2009). While these cognitive mechanisms affect journalists and editors (Lovelace et al., 2022), they crucially affect the media's audiences, who tend to pay more attention to negative news stories than to positive ones. This makes negative news particularly newsworthy. As Peterson argues, "[...] negative, or conflictual, events fulfill several requirements [of newsworthiness]" (1979, p. 120). In other words, journalists and editors who wish their pieces of news to capture the audience's interest have an incentive to select negative news.

From an economic point of view, news stories can themselves be considered a commodity, that are bought and sold, with news outlets as producers and audiences as



consumers (Shoemaker, 2006). Publication decisions by the media are thus, among other factors, shaped by economic aspects such as public reach and marketing considerations (Beale, 2006; Donsbach, 2004). This in turn determines whether and how a piece of information should appear as news in a given outlet (Richardson, 2017). The idea of media organizations competing for attention in a market for news (Fengler & Ruß-Mohl, 2008) is closely related to Diermeier's (2011) idea of newsworthiness being dependent on audience interest, since the latter corresponds to the demand side in the market for news. Additionally, as Thøgersen (2006) suggests, negative information on environmental issues is generally more likely to be newsworthy, due to its linkages to the damage caused and/or failure of an individual or organization to prevent it.

Finally, the news media in their professional understanding and cultural role as a watchdog are committed to their mission of holding governments, firms, and individuals accountable for what they say or (Donohue et al., 1995; Johnson et al., 2005). This increases the likelihood of a critical stance and/or prioritization of negative news stories about these actors over positive ones (Kalogeropoulos et al., 2015).

In sum, cognitive biases and strategic incentives at the individual level, market incentives at the organizational level, as well as professional norms of journalists as public watchdogs, favor the publication of negative news. This is what we call the *strategy-related effect* of negative news (cf. Figure 1).

### ***The tone-related effect of negative news.***

At the level of an individual journalist or editor, the likelihood that a piece of information becomes news depends on this person's evaluation of newsworthiness, as explained in the previous section. Moreover, journalists are reporting new events in the context of prior related developments (Zelizer, 2008) and incorporating new stories into the reporting

they have helped to create in the past (Zandberg et al., 2012). Journalists might even be inclined to report similar events that fit the pattern of the initial event to validate their initial decision of newsworthiness (Donsbach, 2004). Additionally, past reporting might be leveraged not only to generate more interesting narratives but also to save costs (Tan, 2016). Consequently, “[...] negativity may motivate journalists to use past media coverage as a sort of ‘memory database’ of negative events” (Jonkman, Trilling, et al., 2020, p. 6).

Negative news about environmental issues should also be dependent on past reporting on the same type of issues at the level of news organizations (Gentzkow & Shapiro, 2006). Eilders (2006) finds that continuity in reporting influences not only journalistic selection but also what audiences pay attention to. Similarly, consonance with audiences’ expectations and stereotypes has been found to be a criterion of newsworthiness (Bednarek & Caple, 2012). In short, audiences value consistency of what is reported about and how. Therefore, news organizations have incentives to consider these characteristics when trying to capture their audiences’ interest and maximize their returns (Dahlstrom, 2014; O’neill & Harcup, 2009).

Because journalists use past negative reporting as a memory database and audiences prefer consistency and negativity in reporting, the more a firm has been negatively exposed in relation to environmental issues in the past, the more likely journalists and news organizations are to subsequently report additional negative news about this firm’s environmental record. This is what we call the *tone-related effect* of negative news reporting (cf. Figure 1).

In sum, because (1) journalists, editors, media organizations, and the media system are all more likely to report negative news about a firm, (2) journalists and editors, in their professional roles, tend to embed news within a storyline determined by former reporting, and (3) media organizations have a financial interest in catering to the public’s preference for both negativity and consistency, a firm that has been exposed in the past for its environmental record is more likely to subsequently receive negative coverage in relation to environmental issues.

This could potentially have devastating consequences: If a firm is negatively exposed at some point in time, it is more likely to attract additional negative news, which subsequently raises once again the likelihood of being negatively exposed, through a combination of the *strategy-related effect* and the *tone-related effect* of negative news (cf. Figure 1). This means that it could end up being trapped in a spiral of negative exposure and, as a result, be at risk of becoming an environmental villain (Diermeier, 2011). We thus hypothesize that:

*Hypothesis 1. An increase in a firm's negative media exposure for environmental issues increases the likelihood that the firm subsequently receives additional exposure of this type.*

### **The Protective Effect of Positive Media Coverage**

Media exposure on firms is not merely negative; firms also receive positive coverage in the news (e.g., Lorraine et al., 2004). For example, positive reports might be driven by announcements of higher-than-expected earnings (Oliver et al., 2023). Some positive news can also result from firms' disclosures regarding their environmental records (Deegan & Rankin, 1996), such as increased environmental commitment (Bansal & Clelland, 2004) or from a reduction of emissions (Flammer, 2013).

The existence of positive news in a negatively biased news landscape has been attributed to (1) the systemic role of the media in providing a more or less accurate view of a world in which good things also happen sometimes, and (2) an audience potential for positive news (Soroka & Krupnikov, 2021). Positive messages deviating from consistent streams of negative news can be of higher news value (Leung & Lee, 2015), and in a context where negative news prevails, positive news messages can generate higher audience interest than negative ones (Knobloch-Westerwick et al., 2005). This is mostly because they are unexpected in a predominantly negative landscape (Oliver et al., 2023; Soroka & Krupnikov, 2021). Research on the effects of positive and negative news on audiences has thus proposed a silver lining

approach, where the news media honor their surveillance function by remaining critical and, at the same time, embed positive messages in a stream of negative news on a given issue to reap the commercial benefits of positive news (McIntyre & Gibson, 2016). Therefore, even in a situation in which a firm is repeatedly negatively targeted by the media, positive news can have a place in the coverage of this firm. Moreover, due to space and resource constraints, more positive news being reported at a given time also implies less room for negative news. As a result, the increased value of reporting positive news stories amidst negative ones can mitigate the risk of a negative spiral (cf. the *strategy-related effect* of positive news in Figure 1).

Pollock, Rindova, and Maggitti (2008) argue that the higher the amount of positive news reports about a firm that have been published in the past, the greater the availability of these positive evaluations in the journalists' and editors' minds, and the lower their hesitation in publishing positive news on this firm will be. Similarly, observing past positive coverage is likely to make journalists and editors more hesitant to publish negative news about this firm. Indeed, a good reputation acquired through media reports can provide a firm with the "benefit of the doubt" in case of wrongdoing (Godfrey et al., 2009, p. 428; Zavyalova et al., 2016, p. 254): When journalists become aware of negative news about a firm in relation to a particular subject (in our case an environmental issue), they might be more reluctant to report about it if the firm has previously received positive coverage on this subject, which corresponds to the *tone-related effect* of positive news (cf. Figure 1). The fact that this past positive coverage provides this firm with a protective effect is consistent with the observation that a positive reputation provides firms with a buffer against negative events and insurance against crises (Coombs & Holladay, 2006; Pfarrer et al., 2010; Schnietz & Epstein, 2005; Wei et al., 2017).

Research on social movements suggests the opposite, namely that firms publicizing their good environmental deeds and so attempting to create a reputational buffer are more likely to be attacked by activists, because they become good targets (King & McDonnell, 2015). Media

attention has even been reported to amplify the threat that activists' activities pose to firms' reputations (McDonnell & King, 2013). However, the fact that a firm, whose good environmental deeds have been covered by the media, may become a more attractive target to activists does not automatically mean that positive coverage of a firm's environmental record is going to increase the subsequent amount of negative news received by the firm in relation to environmental issues. Our theorization suggests that positive coverage by the media actually has a protective effect.

In sum, due to (1) the increased newsworthiness of positive news in a news media landscape dominated by negative reports, and (2) the way past positive coverage of a firm's environmental trace affects journalists' and editors' decisions on whether to publish a piece of negative environmental news, we expect that, on average, a firm that has received positive coverage in relation to environmental issues is less likely to be subsequently negatively exposed for this type of issue:

*Hypothesis 2. An increase in a firm's positive media exposure for environmental issues reduces the likelihood that the firm subsequently receives additional negative exposure of this type.*

### **The Protective Effect of Firms' Communications**

Research on public relations and the news media shows that public relations efforts can influence the news media discourse on non-financial topics such as social and environmental issues (Kioussis et al., 2007). For example, journalists may integrate information from a firm's press release into a news piece (Maat & De Jong, 2013). One reason why media producers rely on press releases is that this allows them to reduce information gathering costs (Larsson, 2009). A second reason is that some types of information on the firms' activities are only disclosed in these press releases (Bushee et al., 2010).

A firm's efforts to influence media reporting via press releases can directly aim at countering attacks, securing or rebuilding the firm's social standing or more generally educating the public (Desai, 2011; Hiatt et al., 2015). Research has particularly explored that firms engage in such impression management tactics predominantly after initial transgression, often accompanied by negative media exposure (Bansal & Clelland, 2004; McDonnell & King, 2013). For example, Wickman (2014) discusses how BP used press releases to influence the public narrative of its involvement in the Gulf Oil Spill era. Also, Zavyalova et al. (2012) discuss what kinds of announcements firms make in press releases after product recalls. If journalists observe such communication from the firm, they might integrate this information into their reporting or even choose to report more positive news and/or less negative about said firm. This represents the *strategy-related effect* for press releases (cf. Figure 1).

However, press releases might also have a proactive protective effect: They should not have an immediate impact on media reports by attempting to shape the reporting during a crisis, but also affect subsequent media coverage. For example, highly visible firms reportedly employ press releases as a defense mechanism (Jonkman, Boukes, et al., 2020). Firms might thus use press releases on environmental issues also in a proactive manner to protect themselves from criticism in the media. Journalists and editors, when deciding about whether to publicize a piece of negative information about a given firm's environmental issues might be more hesitant to do so not only when there are past positive media reports as seen in the previous section, but also when the firm has published press releases on its environmental actions. Indeed, Bansal and Clelland (2004) found that press releases could potentially enhance the legitimacy of firms in relation to their environmental records, which means that press releases can have a protective effect. Just as we argued in the case of positive coverage, the firm's press releases may be taken as a signal that the firm cares about the environment, and when journalists become aware of a negative piece of information about the firm's environmental record, they should be more likely

to grant the firm the “benefit of the doubt” and more reluctant to report this negative piece of information. This is what we call the *tone-related effect* of press releases (cf. Figure 1).

We thus hypothesize that the publication of press releases related to its environmental issues reduces the likelihood of a firm subsequently being the object of negative media coverage in relation to environmental issues:

*Hypothesis 3. An increase in a firm’s press releases for environmental issues reduces the likelihood that the firm subsequently receives additional negative media exposure of this type.*

## **Data and Methodology**

### **Data**

To test our hypotheses, we use a novel database that contains all the articles on environmental issues published by major English-speaking written media outlets on the 30 largest U.S. firms in terms of market capitalization, which were listed in the Standard & Poor’s 500 index on December 31, 2014. The data, which covers the entire 1999-2014 period, was retrieved from LexisNexis for each one of the 30 firms, by selecting the options “Major World Publications” (source type) and “Environment & Natural Resources” (index terms). The articles that covered environmental issues in relation to the firm were then coded as positive, negative, or neutral, based on a coding protocol, by the two first authors and an assistant. If there was any criticism regarding environmental issues against the firm or its industrial sector, the article was coded as negative. An article was coded as positive if there was no criticism in relation to environmental issues, and it mentioned an improvement of the firm’s or industry’s environmental practices and/or the firm or industry was praised for its environmental record. Finally, if an article covered one or more environmental issues in relation to the firm or its

industry and it could neither be classified as positive nor negative, it was considered as neutral<sup>1</sup>. To assess intercoder reliability, a random sample of 230 news articles was coded by all three coders. In 88% of the cases, all three coders agreed on the tonality of the article. Krippendorff's Alpha, which takes into account the amount of agreement that could be obtained by chance, is equal to .82, which is above the threshold of .80 (Krippendorff, 2018, p. 241). In total, 15,038 articles that mentioned environmental issues related to at least one of the 30 firms were coded. 7,783 (51.8%) of those articles were labelled as "negative", 6,000 (39.9%) as "positive", and 1,255 (8.3%) as "neutral".

Similarly, we retrieved a total of 1,166 press releases emitted in the same 1999-2014 period by the 30 firms in our sample from the Factiva database. When we applied the codebook used for the news articles to these press releases, we found that only 31 (2.7%) of them contained negative information about the firm, while the vast majority, 1,135 (97.3%), were positive or neutral. This is not surprising, as a firm will generally avoid portraying itself in a negative way.

To get a 'feel' for our data, and as an example, Figure 2 below maps the cumulated monthly articles and press releases for the firm General Electric. It shows how the tonality of the media coverage related to environmental issues of this firm evolves over the 16-year window we consider, and it also shows how the firm emits press releases on environmental issues. Interestingly, we can see that from 2004 on there is a substantial increase in the frequency of positive news in the media, as well as an increased frequency of press releases. This may be one of the reasons why the huge peak of negative news in early 2011, associated with the Fukushima disaster and the criticism towards nuclear energy and General Electric's nuclear reactors, did not lead to an increase in the frequency of negative coverage and the

---

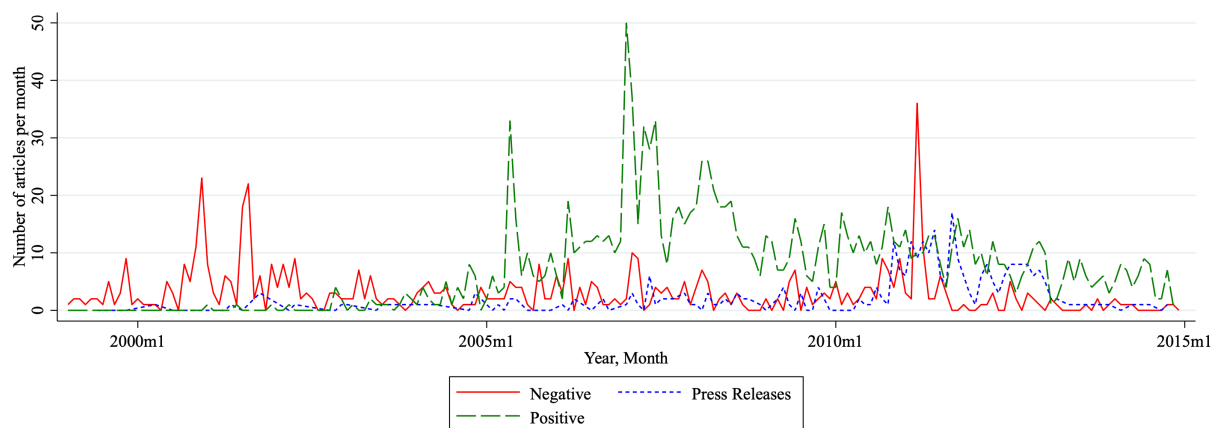
<sup>1</sup> An article containing the same amount of criticism and praise thus was considered as 'negative'. As we wanted to capture any criticism that was voiced regarding environmental issues, criticism overruled praise. This mirrors the idea that negativity is perceived stronger by the audience (Diermeier, 2011).



triggering of a negative spiral. However, in order to determine whether our hypotheses are supported, we need to rely on all the data in our sample and an adequate identification strategy.

## Figure 2

*Example of Monthly Number of Negative and Positive Environmental Articles and Monthly Press Releases of Firm General Electric.*



## Dependent Variable

Our dependent variable, *Present Negative Environmental Exposure*, is a dummy variable that is equal to 1 if a firm has at least one article labeled as “negative” within a given month and 0 otherwise. Because our data spans a 16-year period, there are 192 individual months per firm. Given that the sample contains 30 firms, in theory there should be 5,760 firm-months as units of observation. However, four of the 30 firms were founded or went public during or after 1999 (Amazon.com, Facebook Inc., Google Inc., and Visa Inc.). Therefore, the actual number of firm-months in our dataset is 5,439.

## Independent Variables

To test Hypothesis 1, according to which past negative coverage increases the likelihood of present negative exposure, we build the variable *Past Negative Environmental Exposure*, which is equal to the number of articles in which the firm or its industry have been negatively

exposed in relation to environmental issues during a year. This variable is calculated for each of the three years preceding the month of the dependent variable.<sup>2</sup> The reason why we use a monthly dependent variable and annual values for the independent variable Past Negative Environmental Exposure is that using monthly values for Past Negative Environmental Exposure would result in 36 lags of this variable on the right-hand side of the equation, resulting in unnecessary complexity.

To test Hypotheses 2 and 3, we built two additional variables, *Past Positive Environmental Exposure* and *Past Firm Environmental Press Releases*, respectively. The first accounts for the number of positive environmental articles related to a given firm within the three annual periods mentioned preceding the month of the dependent variable. The second is equal to the number of environmental press releases emitted by the firm that do not contain negative information about the firm, also within these three annual periods. The omission of press releases with negative information is due to two reasons. First, as mentioned earlier, the proportion of these press releases is extremely small. Second, Hypothesis 3 is about the protective effect of press releases, which can only happen with press releases that do not disclose negative information about the firm.

We also constructed Past Neutral Environmental Exposure, calculated using the same procedure as Past Negative Environmental Exposure and Past Positive Environmental Exposure, but using the “neutral” articles instead of the “negative” and “positive” ones, respectively. Although we have not hypothesized whether *Past Neutral Environmental Exposure* has an effect on the dependent variable, and what the effect could be, we nonetheless

---

<sup>2</sup> We examined a temporality of two, three, and four past annual periods for our models. When we compared the models including two and three annual periods, the annual period three years in the past contained important information on the effects of past exposure. Therefore, including only the two previous annual periods is insufficient. When we used four annual periods, the estimators were unable to provide accurate values for the coefficients. This is why we decided to rely on the three previous annual periods.

include three annual periods of Past Neutral Environmental Exposure to prevent an omitted variable bias.

### **Control Variables**

We include several controls to account for potential omitted variable bias as well as to rule out alternative explanations. Panel data estimators allow controlling for firm- or industry-specific time-invariant characteristics by introducing individual effects in the regression models, either in the form of random effects or fixed effects. However, we still need to control for the potential bias that could arise from omitted time-varying variables (Wooldridge, 2010).

In particular, the past environmental performance of a firm could be a confounding factor and bias our estimates. This is because firms being perceived as good or bad environmental performers are treated differently in media reporting (e.g., Aerts & Cormier, 2009). If, for example, two years before, a firm had voluntarily implemented an environmental policy such as a wastewater treatment system at its production plant, and the media reported about it, this would have raised the value of Past Positive Environmental Exposure. At the same time, this treatment system would reduce the risk of a water pollution incident that could subsequently be reported in the media, lowering the value of the dependent variable Present Negative Environmental Exposure. Therefore, if we observed a negative relationship between Past Positive Environmental Exposure and Present Negative Environmental Exposure, it could be due to a protective effect of past positive environmental news (Hypothesis 2) or, alternatively, to the implementation of an environmental policy in the past. Therefore, we must control for this source of bias by introducing the control variable *Environmental Policy*, which relies on the firm's environmental score in the MSCI ESG dataset. This score, whose values are between 0 and 10, captures the extent to which a firm has adopted policies and initiatives to reduce its environmental impact or risks. It can be considered as a valid proxy for environmental performance (Semenova & Hassel, 2015). As with all our other control variables, we used the

MSCI ESG environmental score, which is annual, to construct three Environmental Performance control variables for the three previous annual periods before the focal month.

We employ additional firm-level controls. First, large firms are more likely to be scrutinized in the media in relation to environmental issues (Brammer & Pavelin, 2006). Additionally, firm size is positively correlated with environmental performance (Ioannou & Serafeim, 2012; Jackson & Apostolakou, 2010). We therefore use a firm's *Assets* to control for firm size (Brammer & Millington, 2006; Lenox & Eesley, 2009). The control variable *Net Sales* is introduced as an additional proxy of size that accounts for the firm's size in the product market (Dang et al., 2018) and is also correlated with a firm's performance. Furthermore, a firm's availability of funds determines whether and how much it can invest in environmental and social measures (Waddock & Graves, 1997) and counteract to activist attacks in the media (Eesley & Lenox, 2006; King, 2008). To account for these financial resources, we control for *Cashflow* (Eesley & Lenox, 2006).

Moreover, profitable firms are more likely to have financial resources they can dedicate to environmentally friendly measures (Adams & Hardwick, 1998; Waddock & Graves, 1997) and they may also be more publicly visible and thus more susceptible to scrutiny. Therefore, we introduce return on assets to control for firm *Profitability*. Conversely, high levels of debt may render a firm more visible in the media and reduce the resources available for investments in corporate social responsibility measures (Adams & Hardwick, 1998). Thus, we also control for a firm's *Leverage* as the ratio between total amount of debt and total amount of assets. The data required for all five financial control variables mentioned were retrieved from Compustat (Standard & Poor's) and Worldscope database (Thomson-Reuters) in million USD or percent on a quarterly basis.

Given that our independent variables are constructed on an annual basis, as explained above, our control variables are also annual, and their temporality matches exactly that of the

independent variables. To calculate the value of these annual control variables exactly for each value of the dependent variable, which is a monthly value, we would need monthly data. To address this issue, we first used quarterly data to estimate a monthly value of the control variable. For example, for the dependent variable's firm-month observation April 2013 we would calculate, the value of the control variable for the annual period of April 2012-March 2013, using all the monthly values within this period. The same applies to the other two previous annual periods, that is, April 2011-March 2012 and April 2010-March 2011.

The procedure used to calculate the control variable's monthly values depends on whether the control variable is a variable of stock or flux. To obtain the monthly values of the variables of stock (Profitability, Assets, and Leverage) we rely on a linear interpolation between the values of one specific quarter and its temporal neighbor. In other words, if we have the monthly values for  $C_{i,t}$  and  $C_{i,t-3}$  and we need to estimate  $C_{i,t-1}$  and  $C_{i,t-2}$  by interpolation:

$$C_{i,t-1} = \frac{2 * C_{i,t}}{3} + \frac{C_{i,t-3}}{3} ; C_{i,t-2} = \frac{C_{i,t}}{3} + \frac{2 * C_{i,t-3}}{3}$$

For variables of flux (*Net Sales* and *Cash Flow*), we approximate the average monthly value by first dividing each fiscal quarter value by three to obtain an average monthly measure.

### **Identification Strategy**

We rely on a panel data approach, in which we conduct an analysis of the effect of past media coverage, as well as of past firm press releases on environmental issues, on present negative exposure in environmental issues. To that end, we regress the likelihood that a firm experiences negative exposure in environmental issues in the present on our independent variables (Past Negative Environmental Exposure for Hypothesis 1, Past Positive Environmental Exposure for Hypothesis 2, and Past Firm Environmental Press Releases for Hypothesis 3). We use both panel probit and panel logit models with individual effects that are either random or fixed at the firm level.

## Results

Descriptive statistics for our dependent, independent, and control variables are reported in Table 1 below. All the control variables are either significantly correlated with the dependent variable Present Negative Environmental Exposure, or at least one of our independent variables, which are Past Negative Environmental Exposure, Past Positive Environmental Exposure, Past Neutral Environmental Exposure, and Past Firm Environmental Press Releases.

We ran six different models, which are reported in Columns (1) to (6) in Table 2 below. First, we ran probit and logit models with random effects (Columns 1 and 2). Probit and logit models differ in terms of their underlying assumption on the cumulative density function. While probit is based on a normal distribution, logit relies on a logistic distribution (Wooldridge, 2010). As we have no theoretical reason to prefer one over the other, we run both models and compare the results. We find support for Hypothesis 1, as both models yield highly significant ( $p < .01$ ) and positive coefficient estimates for the three annual periods of Past Negative Environmental Exposure.

Therefore, an increase in a firm's negative coverage in relation to environmental issues in the past raises the probability of being subsequently negatively exposed in relation to an environmental issue. For example, if at some point in time a firm such as Apple gets more negative media coverage because of environmental degradation associated with the sourcing of rare metals employed in its smartphone models, this should lead, according to our estimates, to a higher risk of receiving negative media coverage in relation to environmental issues at a later point in time.

**Table 1***Correlations and Summary Statistics (N = 5,439).*

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Present Negative Env. Exposure	1										
2. Past Negative Env. Exposure	.48***	1									
3. Past Positive Env. Exposure	.32***	.22***	1								
4. Past Neutral Env. Exposure	.33***	.59***	.17***	1							
5. Past Firm Env. Press Releases	.17***	.16***	.56***	.10***	1						
6. Assets <sup>a</sup>	.00	-.01	.16***	.26***	.11***	1					
7. Profitability <sup>b</sup>	.01	.06***	-.05***	-.01	-.08**	-.26***	1				
8. Leverage <sup>b</sup>	.10***	-.08***	.27***	-.03**	.22***	.16***	-.47***	1			
9. Net Sales <sup>a</sup>	.34***	.55***	.35***	.33***	.22***	.25***	.00	-.02	1		
10. Cashflow <sup>a</sup>	.16***	.40***	.07***	.22***	.22***	.12***	-.06***	-.01	.51***	1	
11. Environmental Policy <sup>c</sup>	.13***	.00	.20***	-.01	.17***	-.14***	.18***	-.12***	.05***	-.02	1
Mean	.26	13.95	10.00	2.26	1.90	232.70	1.93	22.14	68.44	3.42	5.53
Standard Deviation	.44	38.54	24.54	6.81	6.36	453.30	3.14	18.14	80.71	4.15	2.55
Min	.00	.00	.00	.00	.00	.00	-38.36	.00	.00	.00	.00
Max	1.00	400.00	282.00	76.00	118.00	2,511	15.63	153.40	484.10	21.59	10.00

*Note.* The table contains the summary statistics and correlations for all variables. For ease of understanding, only the first of the previous years is presented for all variables that have several yearly temporal periods.

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

<sup>a</sup> Summary statistics in million USD.

<sup>b</sup> Summary statistics in %.

<sup>c</sup> Summary statistics of score in environmental pillar of MSCI ESG.

**Table 2***Results of Panel Probit and Logit Regressions on Present Negative Environmental Exposure.*

Variables / Model	1	2	3	4	5	6
Past Negative Env. Exposure Previous year	.02*** (.00)	.04*** (.01)	.03*** (.01)	.03*** (.01)	.03*** (.01)	.03*** (.01)
Past Negative Env. Exposure Year 2 before	.01*** (.00)	.02*** (.01)	.01** (.01)	.01 (.01)	.01** (.01)	.01 (.01)
Past Negative Env. Exposure Year 3 before	.01*** (.00)	.02*** (.00)	.01*** (.01)	.01** (.01)	.02*** (.01)	.02*** (.01)
Past Positive Env. Exposure Previous year	.01*** (.00)	.01*** (.00)	.01** (.00)	.01 (.00)	.01*** (.00)	.01** (.00)
Past Positive Env. Exposure Year 2 before	-.01** (.00)	-.01** (.01)	-.01* (.01)	-.01** (.00)	-.01* (.01)	-.01* (.00)
Past Positive Env. Exposure Year 3 before	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.01 (.00)	.01 (.00)
Past Neutral Env. Exposure Previous year	.03** (.01)	.05** (.02)	.04 (.03)	.04 (.03)	.06** (.03)	.06* (.03)
Past Neutral Env. Exposure Year 2 before	-.01 (.02)	-.02 (.03)	-.03 (.04)	.00 (.03)	-.01 (.04)	.02 (.03)
Past Neutral Env. Exposure Year 3 before	-.01 (.01)	-.02 (.01)	-.03 (.02)	-.02 (.03)	-.02 (.03)	-.02 (.03)
Past Firm Env. Press Releases Previous year					-.04*** (.01)	-.04*** (.01)
Past Firm Env. Press Releases Year 2 before					.01 (.01)	.01 (.01)
Past Firm Env. Press Releases Year 3 before					-.03*** (.00)	-.03*** (.01)
Constant	-1.57*** (.34)	-2.77*** (.63)				
Panel Estimator	Probit	Logit	Logit	Logit	Logit	Logit
Individual Random Effects	Yes	Yes	No	No	No	No
Individual Fixed Effects	No	No	Yes	Yes	Yes	Yes
Monthly Time Dummies	No	No	No	Yes	No	Yes
Environmental Policy Control	Yes	Yes	Yes	Yes	Yes	Yes
Financial Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	3,663	3,663	3,052	3,052	3,052	3,052
Number of Firms	30	30	25	25	25	25

*Note.* Robust standard errors in parentheses.\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$



However, the result in the first annual period (1 to 12 months prior to the focal month) could be due to ongoing media coverage on one particular issue over successive months (e.g., an environmental scandal that starts appearing in the media towards the end of one month and continues to be reported in the following month).

To determine whether this could explain the effect we observe in Columns (1) and (2), we split the first annual period (1 to 12 months prior to the focal month) into two different subperiods. The first one includes the first three months of the annual period and the second one the nine other months.<sup>3</sup>

The coefficient estimates for both subperiods are positive and significantly different from zero ( $p < .05$ ) showing that the effects observed in Columns (1) and (2) for the first annual period of *Past Negative Environmental Exposure* could be, at least in part, due to events spread over a few months, but that they are also due to the effect described in Hypothesis 1.

We also find support for Hypothesis 2, as there is a negative and significant ( $p < .05$ ) effect of Past Positive Environmental Exposure on Present Negative Environmental Exposure in the second annual period (13 to 24 months prior). Therefore, Past Positive Environmental Exposure leads to a lower Present Negative Exposure in the future, even if this seems to take some time. For example, if at some point in time a firm such as Coca-Cola gets praised in the media for a recycling initiative it supports, this will lead to a lower risk of receiving negative media coverage for environmental issues later in time (between 13 and 24 months afterwards).

As with the first annual period of Past Negative Environmental Exposure, we split the first annual period of Past Positive Environmental Exposure to assess whether there is an

---

<sup>3</sup> The rationale is as follows: Ongoing coverage of a given event (e.g., a pollution incident) might spread over successive months, including the month  $t$  of the dependent variable. If we did find an effect, or an effect in the opposite direction, in the second subperiod further away from  $t$  (9 to 12 months before) and only an effect in the first subperiod closest to  $t$  (1 to 3 months before), our findings about the whole period (1 to 12 months) could be exclusively driven by such an ongoing coverage. This is not the case here. The results of this analysis can be obtained from the first author upon request.

overlap of ongoing coverage.<sup>4</sup> We find that there is indeed an overlap that explains the positive and significant ( $p < .01$ ) coefficient of the first annual period (1 to 12 months prior) of Past Positive Environmental Exposure. However, the effect observed for the second lag of Past Positive Environmental Exposure, which supports Hypothesis 2, is not due to an overlap of this type.

Since the random-effects panel data estimator assumes that the individual effects are uncorrelated with the independent variables, it does not control for firm- or industry-specific time-invariant characteristics that could be omitted variables and bias the estimates. The fixed effects panel data estimator controls for this source of endogeneity, but it is less efficient (Wooldridge, 2010). Therefore, in Column (3) we run a fixed effects logit model and compare the results with Column (2) using a Hausman test, which allows assessing if the fixed effects' estimates are more trustworthy than the random effects'. The Hausman test does not allow concluding that the fixed effects model is more reliable than the random effects one ( $p = 1.00$ ). Therefore, we take into account both models (Columns 2 and 3). Results in Column (3) thus confirm the support for Hypotheses 1 and 2 provided by Columns (1) and (2). Because the probit estimator does not allow for fixed effects, we could not run the Hausman test for this model. However, the Hausman test for the logit model shows that we can rely on the random effects models' results. Additionally, there might be events at any point in time that could affect all the firms in the sample simultaneously and thus bias the estimates (Wooldridge, 2010). To account for this, we add monthly time-fixed effects to the logit fixed effects model of Column (3) and report them in Column (4). Again, this model's coefficient estimates support Hypotheses 1 and 2.

To test Hypothesis 3, we add three annual periods of Past Firm Environmental Press Releases to Columns (3) and (4)'s models and report these results in Columns (5) and (6),

---

<sup>4</sup> The results of this analysis can be obtained from the first author upon request.

respectively. In the first and third annual periods (1 to 12 months and 25 to 36 months prior) of Past Firm Environmental Press Releases, we find a negative and significant ( $p < .01$ ) effect on Present Negative Environmental Exposure, providing support for Hypothesis 3. This means that if, for example, a firm such as IBM issued a press release about the start of a recycling initiative, this press release would lead to a lower risk of receiving negative media coverage for environmental issues later in time.

As we did previously, we split the first annual period of Past Firm Environmental Press Releases into two different periods, to assess if the effect found in the first annual period (1 to 12 months prior) is due to an overlap of ongoing coverage and press releases a firm emits related to this ongoing coverage.<sup>5</sup> This analysis reveals that the effect is not due to ongoing media coverage associated with related press releases.

### **Robustness Tests**

To validate our findings, we conducted numerous robustness tests, which in essence support our results. We examined whether the effects we find could be driven by political bias of news outlets, as, for example, left-leaning media could be reporting systematically more negative news about firms (e.g., Entman, 2010; Roulet & Clemente, 2018). We were further concerned that ownership and/or advertisement patterns of firms in media outlets could bias reporting in favor of firms (e.g., Gurun & Butler, 2012). Additionally, the effect of past negative exposure on the present one could actually be due to the direct effects of past activist attacks on present negative exposure, independently of the fact that these past attacks have been reported by the media and thus contributed to the value of Past Negative Environmental Exposure. It would also be possible that our results are essentially driven by firms in controversial industries, as such industries are more contested (e.g., Palazzo & Scherer, 2006),

---

<sup>5</sup> The results of this analysis can be obtained from the first author upon request.

while absent from other industries. Our robustness tests provided no support for any of these alternative explanations to our hypotheses.

While our study does not distinguish between types of media outlets, industry-specific outlets such as trade magazines (Wilkinson & Merle, 2013), which tend to publish favorable content on their respective industry members (Edwards & Pieczka, 2013), could drive the protective effect of positive news and press releases we have observed. Indeed, field media play a role in the legitimation of industry practices, even if the reporting in the public media is hostile (Clemente & Roulet, 2015). In a post-hoc analysis, the removal of industry publications from our sample did not alter our main results and conclusions, which means that, while industry outlets could contribute to the protective effects we observe, they are not their sole drivers.<sup>6</sup>

Additionally, we examined whether there could be positive spirals in news reporting and found no empirical evidence of their existence.<sup>7</sup>

The results of our main analysis also show that the control variable we used, Environmental Performance, did not seem to have a significant effect on the dynamic of exposure we observe and hence the risk of negative spirals. This suggests that the dynamic effects that explain a firm's likelihood of future negative exposure in the environmental dimension are not directly attributable to firm's environmental practices, but rather to whether these are reported or not.

## Discussion

In this paper, we examine the dynamic effects of media exposure in the context of firms' coverage regarding environmental issues. Compared to existing research on the relationship between firms and the media – we provide a theoretical approach that relaxes two important assumptions regarding the impact of the media on firms: (1) media exposure is not considered

---

<sup>6</sup> Results of this analysis can be obtained from the first author upon request.

<sup>7</sup> Results of this analysis can be obtained from the first author upon request.

at a few specific moments but rather over a longer period in time, with a dynamic approach and (2) press releases are conceptualized to have a protective effect which firms can exploit to proactively communicate, before any negative coverage occurs. This allows us to theorize on the possibility of negative media spirals over time, in which negative news regarding a firm's environmental practices raises the likelihood of subsequent negative exposure over the same type of issues. From this perspective, the impact of the media's reporting on a firm in the context of an environmental issue might be even more damaging – for instance to its reputation– than currently described. We observe this mechanism in our unique 16-year panel dataset, which allows us to test our hypotheses regarding dynamic effects by exploiting the intertemporal variation of the data: negative news about a firm issued at a certain moment in time increases the probability that more negative media exposure will be created in subsequent periods. This mechanism could lead to the emergence of negative news spirals over time, potentially turning firms into environmental villains (Diermeier, 2011).

At the same time, our dynamic approach, allows us to theorize about how these potential negative media spirals might also be quite fragile, similar to the observation that virtuous cycles of media attention “eventually die out” (Seguin, 2016, p. 1001). Indeed, our study shows that positive coverage about the firms' environmental practices and firm's press releases has a preventive effect against these spirals. Therefore, by nurturing positive coverage and emitting press releases about its good environmental deeds, a firm can benefit from protective reputation insurance (Coombs & Holladay, 2006) against future negative coverage on environmental issues. In sum, positive media reports and proactive press releases are a source of firm resilience in the event of future crises.

We cannot rule out that doing good and publicizing ones' efforts may also present a liability, as firms whose environmental record is more visible have been reported to be better targets for activist attacks (King & McDonnell, 2015). However, our results clearly show that

positive coverage and firm statements in press releases both have, on average, a protective effect in terms of future negative exposure in the media. Further, if environmental record visibility increased the likelihood of subsequent criticism of the firm (King, 2008), and this raised the future amount of negative coverage, we would expect neutral articles on the firm's environmental record to subsequently increase the likelihood of the firm's present exposure in the environmental dimension, which is not what our data show. Additionally, the firms' ability to reduce the likelihood of future negative environmental exposure that we observe points towards a less critical stance of the media towards business than the literature generally would suggest.

Taken together, the two mechanisms that are at the core of this article provide a novel understanding that complements the existing literature on the relationship between firms and the media. They highlight the importance of considering the media's strategic choices regarding newsworthiness as well as social and cognitive factors at the level of individual journalists and editors. While negative exposure at one point in time could eventually transform a firm into an environmental villain through a spiral of negative reporting that results from a combination of newsworthiness of negativity and journalists writing ongoing dramatic narratives, firms can strategically counteract such a risk by either nurturing more positive news about themselves and/or by emitting positive press releases, because both have a protective effect. Moreover, our study shows that firms are not passive recipients of media coverage, but that they can strategically influence the tone of their environmental records' coverage.

In terms of firms' media reputation management, our study shows that firms subject to negative media reports can not only resort to changes in environmental practices (Amer & Bonardi, 2023) or to the type of strategic change that helps protect the firm's and its executives' reputation in the media from further damage (Bednar, 2012; Bednar et al., 2013), but they can also promote positive reports about their environmental records in the media and communicate

about their efforts in the environmental dimension through press releases to prevent additional damage and avoid negative spirals. In fact, these proactive communication-related strategies may even be more effective in preventing future negative coverage than adopting a reactive strategy where the firm simply responds to media accusations by complying with the explicit or implicit expectations expressed in media reports. In that spirit, our research contributes to an understanding of press releases being used in a proactive manner to shape impressions in calmer times to protect against future negative media exposure.

Indeed, Lamin and Zaheer (2012) found that, while addressing a questionable practice for which a firm is accused in the media does not improve the tone of the media coverage of this firm, the tone of subsequent coverage was less negative for firms with more positive media reputations (Deephouse, 2000). In other words, positive news coverage and firm press releases about a firm's environmental practices may be able to generate more reputation insurance (Minor & Morgan, 2011) than adopting a reactive approach of simple compliance with the expectations of stakeholders that appear in negative media reports. This is also consistent with our finding that the dynamic effects that explain a firm's likelihood of future negative exposure in the environmental dimension are not directly attributable to that firm's environmental practices, but rather to whether these are reported or not. Finally, our study shows the key role of the strategic interactions between the media and firms concerning news content and tonality.

### **Limitations and Avenues for Future Research**

In this study, we did not distinguish the news' sources. The news in the media may originate from a diversity of sources such as investigative journalists, regulators, governments, press releases, or activists' reports. The magnitude of the dynamic effects we have described in this article could, for example, depend on who is criticizing or praising the firm. A negative article covering a dramatic activist attack could result in a higher likelihood of triggering a negative spiral than if the source is an investigative journalist. However, the opposite could also

be true, since an environmental activist could also be perceived as a less reliable source of information than an investigative journalist. Therefore, evaluating the effect of the source on the magnitude of the effects described in this paper would be an interesting avenue for future research in this literature.

An obvious limitation to our study is that our sample contains large US firms, and our focus on environmental issues. Further studies should determine whether these results hold for different sets of firms in terms of size and geographical orientation, as well as for other types of issues. There is no obvious reason why our dynamic effects would not apply to social or other types of issues, but this remains to be confirmed.

A final limitation is that our analysis is focused on traditional written news, and we did not consider other types of media such as radio, TV, and social media. We do not anticipate that this is an issue, for two reasons. First, there is solid evidence that sources such as radio and TV rely on content from elite printed news and that printed news has longer-lasting effects on setting the topics the media focuses on (Golan, 2006; Vliegenthart & Walgrave, 2008). Traditional written news has also been found to have a stronger influence on public opinion and is better recalled (Deephouse & Carter, 2005; DeFleur et al., 1992). Second, the power of social media remains limited in terms of shaping discourse in traditional media but rather works as an accelerator that spreads the news originating in the latter (Zhao et al., 2011). Social media usually shape the discourse after the main direction has been set by traditional news media (Sayre et al., 2010). Yet, further research is needed to confirm whether the role of social media is analogous when it comes to dynamic media effects.

## **Conclusion**

This paper adds a dynamic understanding of negative media exposure to existing research on firm-media interactions, which has important consequences for the firms' social evaluations. In a nutshell, our study shows that negative exposure of firms in relation to



environmental issues is driven partly by past negative environmental exposure, through a dynamic process where firms are at risk ending up being perceived as environmental villains.

What is the implication of our study regarding general sustainability concerns? The positive spin is that the risk of falling into a spiral of negative exposure should be an additional incentive for firms to improve environmental practices, beyond the effects of negative news on self-regulation already uncovered by previous literature (Amer & Bonardi, 2023).

But there is also a less positive spin to our findings. If press releases have a protective effect against future negative environmental exposure, it also means that firms reluctant to substantially improve environmental practices could strategically engage in symbolic environmental actions and report them in press releases, in order to protect themselves from negative coverage, while at the same time maintaining the practices that are damaging for the environment. This implies that there are circumstances, where at least some journalists end up in a role of being rather tame lapdogs instead of critical watchdogs vis-à-vis firms (Bednar, 2012).

## References

- Adams, M., & Hardwick, P. (1998). An analysis of corporate donations: United Kingdom evidence. *Journal of Management Studies*, 35(5), 641–654. <https://doi.org/10.1111/1467-6486.00113>
- Aerts, W., & Cormier, D. (2009). Media legitimacy and corporate environmental communication. *Accounting, Organizations and Society*, 34(1), 1–27. <https://doi.org/10.1016/j.aos.2008.02.005>
- Amer, E., & Bonardi, J.-P. (2023). Firms, activist attacks, and the forward-looking management of reputational risks. *Strategic Organization*, 147612702211249. <https://doi.org/10.1177/14761270221124941>
- Andrews, K. T., & Caren, N. (2010). Making the news: Movement organizations, media attention, and the public agenda. *American Sociological Review*, 75(6), 841–866. <https://doi.org/10.1177/0003122410386689>
- Bansal, P., & Clelland, I. (2004). Talking trash: Legitimacy, impression management, and unsystematic risk in the context of the natural environment. *Academy of Management Journal*, 47(1), 93–103. <https://doi.org/10.5465/20159562>
- Baron, D. P. (2005). Competing for the public through the news media. *Journal of Economics & Management Strategy*, 14(2), 339–376. <https://doi.org/10.1111/j.1530-9134.2005.00044.x>
- Baron, D. P., & Diermeier, D. (2007). Strategic activism and nonmarket strategy. *Journal of Economics & Management Strategy*, 16(3), 599–634. <https://doi.org/10.1111/j.1530-9134.2007.00152.x>
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>

- Beale, S. S. (2006). The news media's influence on criminal justice policy: How market-driven news promotes punitiveness. *William and Mary Law Review*, 48, 397–481. <https://scholarship.law.wm.edu/wmlr/vol48/iss2/2>
- Bednar, M. K. (2012). Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. *Academy of Management Journal*, 55(1), 131–150. <https://doi.org/10.5465/amj.2009.0862>
- Bednar, M. K., Boivie, S., & Prince, N. R. (2013). Burr under the saddle: How media coverage influences strategic change. *Organization Science*, 24(3), 910–925. <https://doi.org/10.1287/orsc.1120.0770>
- Bednarek, M., & Caple, H. (2012). *News Discourse (Vol. 46)*. A&C Black.
- Bonardi, J.-P., & Keim, G. D. (2005). Corporate political strategies for widely salient issues. *Academy of Management Review*, 30(3), 555–576. <https://doi.org/10.5465/amr.2005.17293705>
- Brammer, S. J., & Millington, A. (2006). Firm size, organizational visibility and corporate philanthropy: An empirical analysis. *Business Ethics: A European Review*, 15(1), 6–18. <https://doi.org/10.1111/j.1467-8608.2006.00424.x>
- Brammer, S. J., & Pavelin, S. (2004). Building a Good Reputation. *European Management Journal*, 22(6), 704–713. <https://doi.org/10.1016/j.emj.2004.09.033>
- Brammer, S. J., & Pavelin, S. (2006). Corporate reputation and social performance: The importance of fit. *Journal of Management Studies*, 43(3), 435–455. <https://doi.org/10.1111/j.1467-6486.2006.00597.x>
- Breitinger, D., & Bonardi, J.-P. (2019). Firms, breach of norms, and reputation damage. *Business & Society*, 58(6), 1143–1176. <https://doi.org/10.1177/0007650317695531>
- Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—A dual test of media agenda setting theory and legitimacy theory.

- Accounting and Business Research, 29(1), 21–41.  
<https://doi.org/10.1080/00014788.1998.9729564>
- Bushee, B. J., Core, J. E., Guay, W. R., & Hamm, S. J. W. (2010). The role of the business press as an information intermediary. *Journal of Accounting Research*, 48(1), 1–19.  
<https://doi.org/10.1111/j.1475-679X.2009.00357.x>
- Caple, H., & Bednarek, M. (2013). Delving into the discourse: Approaches to news values in journalism studies and beyond [Working Paper]. Reuters Institute for the Study of Journalism.
- Carroll, C. E., & McCombs, M. (2003). Agenda-setting effects of business news on the public's images and opinions about major corporations. *Corporate Reputation Review*, 6(1), 36–46. <https://doi.org/10.1057/palgrave.crr.1540188>
- Clemente, M., & Roulet, T. J. (2015). Public opinion as a source of deinstitutionalization: A “spiral of silence” approach. *Academy of Management Review*, 40(1), 96–114.  
<https://doi.org/10.5465/amr.2013.0279>
- Coombs, T. W., & Holladay, S. J. (2006). Unpacking the halo effect: Reputation and crisis management. *Journal of Communication Management*, 10(2), 123–137.  
<https://doi.org/10.1108/13632540610664698>
- Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111(supplement\_4), 13614–13620. <https://doi.org/10.1073/pnas.1320645111>
- Dang, C., (Frank) Li, Z., & Yang, C. (2018). Measuring firm size in empirical corporate finance. *Journal of Banking & Finance*, 86, 159–176.  
<https://doi.org/10.1016/j.jbankfin.2017.09.006>
- Deegan, C., & Rankin, M. (1996). Do Australian companies report environmental news objectively?: An analysis of environmental disclosures by firms prosecuted successfully

- by the Environmental Protection Authority. *Accounting, Auditing & Accountability Journal*, 9(2), 50–67. <https://doi.org/10.1108/09513579610116358>
- Deephouse, D. L. (2000). Media reputation as a strategic resource: An integration of mass communication and resource-based theories. *Journal of Management*, 26(6), 1091–1112. <https://doi.org/10.1177/014920630002600602>
- Deephouse, D. L., & Carter, S. M. (2005). An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, 42(2), 329–360. <https://doi.org/10.1111/j.1467-6486.2005.00499.x>
- Deephouse, D. L., & Suchman, M. (2008). Legitimacy in Organizational Institutionalism. In *The SAGE Handbook of Organizational Institutionalism* (pp. 49–77). SAGE Publications Ltd. <https://doi.org/10.4135/9781849200387.n2>
- DeFleur, M. L., Davenport, L., Cronin, M., & DeFleur, M. (1992). Audience recall of news stories presented by newspaper, computer, television and radio. *Journalism Quarterly*, 69(4), 1010–1022. <https://doi.org/10.1177/10776990920690041>
- Desai, V. M. (2011). Mass Media and Massive Failures: Determining Organizational Efforts to Defend Field Legitimacy Following Crises. *Academy of Management Journal*, 54(2), 263–278. <https://doi.org/10.5465/amj.2011.60263082>
- Diermeier, D. (2011). *Reputation Rules: Strategies for Building Your Company's Most Valuable Asset* (Vol. 9). McGraw-Hill New York.
- Donohue, G. A., Tichenor, P. J., & Olien, C. N. (1995). A guard dog perspective on the role of media. *Journal of Communication*, 45(2), 115–132. <https://doi.org/10.1111/j.1460-2466.1995.tb00732.x>
- Donsbach, W. (2004). Psychology of news decisions: Factors behind journalists' professional behavior. *Journalism*, 5(2), 131–157. <https://doi.org/10.1177/146488490452002>

- Donsbach, W. (2012). Journalists' role perception. In W. Donsbach (Ed.), *The International Encyclopedia of Communication* (pp. 1–6). John Wiley & Sons, Ltd.  
<https://doi.org/10.1002/9781405186407.wbiecj010.pub2>
- Durand, R., & Vergne, J.-P. (2015). Asset divestment as a response to media attacks in stigmatized industries: Asset Divestment as a Response to Media Attacks. *Strategic Management Journal*, 36(8), 1205–1223. <https://doi.org/10.1002/smj.2280>
- Dyck, A., Volchkova, N., & Zingales, L. (2008). The Corporate Governance Role of the Media: Evidence from Russia. *The Journal of Finance*, 63(3), 1093–1135.  
<https://doi.org/10.1111/j.1540-6261.2008.01353.x>
- Dyck, A., & Zyngales, L. (2002). The corporate governance role of the media. In R. Islam, S. Djankov, & C. McLeish (Eds.), *The right to tell. The role of mass media in economic development* (pp. 107–137). The World Bank.
- Edwards, L., & Pieczka, M. (2013). Public relations and 'its' media: Exploring the role of trade media in the enactment of public relations' professional project. *Public Relations Inquiry*, 2(1), 5–25. <https://doi.org/10.1177/2046147X12464204>
- Eesley, C., & Lenox, M. J. (2006). Firm responses to secondary stakeholder action. *Strategic Management Journal*, 27(8), 765–781. <https://doi.org/10.1002/smj.536>
- Eilders, C. (2006). News factors and news decisions. Theoretical and methodological advances in Germany. *Comm*, 31(1), 5–24. <https://doi.org/10.1515/COMMUN.2006.002>
- Eisenegger, M. (2005). *Reputation in der Mediengesellschaft*. Springer.
- Eisenegger, M., & Schranz, M. (2011). Reputation management and corporate social responsibility. In Ø. Ihlen, J. L. Bartlett, & S. May (Eds.), *The Handbook of Communication and Corporate Social Responsibility* (1st ed., pp. 128–146). Wiley.  
<https://doi.org/10.1002/9781118083246.ch7>

- Entman, R. M. (2010). Media framing biases and political power: Explaining slant in news of campaign 2008. *Journalism*, 11(4), 389–408. <https://doi.org/10.1177/1464884910367587>
- Fengler, S., & Ruß-Mohl, S. (2008). Journalists and the information-attention markets: Towards an economic theory of journalism. *Journalism*, 9(6), 667–690. <https://doi.org/10.1177/1464884908096240>
- Flammer, C. (2013). Corporate social responsibility and shareholder reaction: The environmental awareness of investors. *Academy of Management Journal*, 56(3), 758–781. <https://doi.org/10.5465/amj.2011.0744>
- Fombrun, C., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of Management Journal*, 33, 233–258. <https://doi.org/10.5465/256324>
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1), 64–90. <https://doi.org/10.1177/002234336500200104>
- Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2), 280–316. <https://doi.org/10.1086/499414>
- Godfrey, P. C., Merrill, C. B., & Hansen, J. M. (2009). The relationship between corporate social responsibility and shareholder value: An empirical test of the risk management hypothesis. *Strategic Management Journal*, 30(4), 425–445. <https://doi.org/10.1002/smj.750>
- Golan, G. (2006). Inter-media agenda setting and global news coverage: Assessing the influence of the New York Times on three network television evening news programs. *Journalism Studies*, 7(2), 323–333. <https://doi.org/10.1080/14616700500533643>

- Graf-Vlachy, L., Oliver, A. G., Banfield, R., König, A., & Bundy, J. (2020). Media coverage of firms: Background, integration, and directions for future research. *Journal of Management*, 46(1), 36–69. <https://doi.org/10.1177/01492063198641>
- Gurun, U. G., & Butler, A. W. (2012). Don't believe the hype: Local media slant, local advertising, and firm value. *The Journal of Finance*, 67(2), 561–598. <https://doi.org/10.1111/j.1540-6261.2012.01725.x>
- Harcup, T., & O'Neill, D. (2017). What is News?: News values revisited (again). *Journalism Studies*, 18(12), 1470–1488. <https://doi.org/10.1080/1461670X.2016.1150193>
- Hiatt, S. R., Grandy, J. B., & Lee, B. H. (2015). Organizational responses to public and private politics: An analysis of climate change activists and U.S. oil and gas firms. *Organization Science*, 26(6), 1553–1804. <https://doi.org/10.1287/orsc.2015.1008>
- Hilbig, B. E. (2009). Sad, thus true: Negativity bias in judgments of truth. *Journal of Experimental Social Psychology*, 45(4), 983–986. <https://doi.org/10.1016/j.jesp.2009.04.012>
- Ioannou, I., & Serafeim, G. (2012). What drives corporate social performance? The role of nation-level institutions. *Journal of International Business Studies*, 43(9), 834–864. <https://doi.org/10.1057/jibs.2012.26>
- Jackson, G., & Apostolakou, A. (2010). Corporate social responsibility in Western Europe: An institutional mirror or substitute? *Journal of Business Ethics*, 94(3), 371–394. <https://doi.org/10.1007/s10551-009-0269-8>
- Johnson, J. L., Ellstrand, A. E., Dalton, D. R., & Dalton, C. M. (2005). The influence of the financial press on stockholder wealth: The case of corporate governance. *Strategic Management Journal*, 26(5), 461–471. <https://doi.org/10.1002/smj.457>
- Jonkman, J. G. F., Boukes, M., Vliegthart, R., & Verhoeven, P. (2020). Buffering negative news: Individual-level effects of company visibility, tone, and pre-existing attitudes on



- corporate reputation. *Mass Communication and Society*, 23(2), 272–296.  
<https://doi.org/10.1080/15205436.2019.1694155>
- Jonkman, J. G. F., Trilling, D., Verhoeven, P., & Vliegenthart, R. (2020). To pass or not to pass: How corporate characteristics affect corporate visibility and tone in company news coverage. *Journalism Studies*, 21(1), 1–18.  
<https://doi.org/10.1080/1461670X.2019.1612266>
- Kalogeropoulos, A., Svensson, H. M., Van Dalen, A., De Vreese, C., & Albæk, E. (2015). Are watchdogs doing their business? Media coverage of economic news. *Journalism*, 16(8), 993–1009. <https://doi.org/10.1177/1464884914554167>
- King, B. G. (2008). A political mediation model of corporate response to social movement activism. *Administrative Science Quarterly*, 53(3), 395–421.  
<https://doi.org/10.2189/asqu.53.3.395>
- King, B. G., & McDonnell, M.-H. (2015). Good firms, good targets: The relationship among corporate social responsibility, reputation, and activist targeting. In K. Tsutsui & A. Lim (Eds.), *Corporate Social Responsibility in a Globalizing World* (pp. 430–454). Cambridge University Press. <https://doi.org/10.1017/CBO9781316162354.013>
- Kiouis, S., Popescu, C., & Mitrook, M. (2007). Understanding influence on corporate reputation: An examination of public relations efforts, media coverage, public opinion, and financial performance from an agenda-building and agenda-setting perspective. *Journal of Public Relations Research*, 19(2), 147–165.  
<https://doi.org/10.1080/10627260701290661>
- Knobloch-Westerwick, S., Carpentier, F. D., Blumhoff, A., & Nickel, N. (2005). Selective exposure effects for positive and negative news: Testing the robustness of the informational utility model. *Journalism & Mass Communication Quarterly*, 82(1), 181–195. <https://doi.org/10.1177/107769900508200112>

- Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications. <https://lccn.loc.gov/2017050739>
- Lamin, A., & Zaheer, S. (2012). Wall street vs. main street: Firm strategies for defending legitimacy and their impact on different stakeholders. *Organization Science*, 23(1), 47–66. <https://doi.org/10.1287/orsc.1100.0631>
- Larsson, L. (2009). PR and the media: A collaborative relationship? *Nordicom Review*, 30(1), 131–147. <https://doi.org/10.1515/nor-2017-0143>
- Lenox, M. J., & Eesley, C. E. (2009). Private environmental activism and the selection and response of firm targets. *Journal of Economics & Management Strategy*, 18(1), 45–73. <https://doi.org/10.1111/j.1530-9134.2009.00207.x>
- Leung, D. K. K., & Lee, F. L. F. (2015). How journalists value positive news: The influence of professional beliefs, market considerations, and political attitudes. *Journalism Studies*, 16(2), 289–304. <https://doi.org/10.1080/1461670X.2013.869062>
- Lorraine, N. H. J., Collison, D. J., & Power, D. M. (2004). An analysis of the stock market impact of environmental performance information. *Accounting Forum*, 28(1), 7–26. <https://doi.org/10.1016/j.accfor.2004.04.002>
- Lovelace, J. B., Bundy, J., Pollock, T. G., & Hambrick, D. C. (2022). The push and pull of attaining CEO celebrity: A media routines perspective. *Academy of Management Journal*, 65(4), 1169–1191. <https://doi.org/10.5465/amj.2020.0435>
- Maat, H. P. (2007). How promotional language in press releases is dealt with by journalists: Genre mixing or genre conflict? *Journal of Business Communication*, 44(1), 59–95. <https://doi.org/10.1177/0021943606295780>
- Maat, H. P., & De Jong, C. (2013). How newspaper journalists reframe product press release information. *Journalism*, 14(3), 348–371. <https://doi.org/10.1177/1464884912448914>

- McDonnell, M.-H., & King, B. (2013). Keeping up appearances: Reputational threat and impression management after social movement boycotts. *Administrative Science Quarterly*, 58(3), 387–419. <https://doi.org/10.1177/0001839213500032>
- McIntyre, K. E., & Gibson, R. (2016). Positive news makes readers feel good: A “silver-lining” approach to negative news can attract audiences. *Southern Communication Journal*, 81(5), 304–315. <https://doi.org/10.1080/1041794X.2016.1171892>
- Minor, D., & Morgan, J. (2011). CSR as reputation insurance: Primum non nocere. *California Management Review*, 53(3), 40–59. <https://doi.org/10.1525/cmr.2011.53.3.40>
- Niven, D. (2001). Bias in the news: Partisanship and negativity in media coverage of presidents George Bush and Bill Clinton. *Harvard International Journal of Press/Politics*, 6(3), 31–46. <https://doi.org/10.1177/108118001129172215>
- Oliver, A. G., Campbell, R., Graffin, S., & Bundy, J. (2023). Media coverage of earnings announcements: How newsworthiness shapes media volume and tone. *Journal of Management*, 49(4), 1213–1245. <https://doi.org/10.1177/01492063221080125>
- O’neill, D., & Harcup, T. (2009). News values and selectivity. In *The Handbook of Journalism Studies* (pp. 181–194). Routledge.
- Palazzo, G., & Scherer, A. G. (2006). Corporate legitimacy as deliberation: A communicative framework. *Journal of Business Ethics*, 66(1), 71–88. <https://doi.org/10.1007/s10551-006-9044-2>
- Peterson, S. (1979). Foreign news gatekeepers and criteria of newsworthiness. *Journalism Quarterly*, 56(1), 116–125. <https://doi.org/10.1177/107769907905600118>
- Pfarrer, M. D., Decelles, K. A., Smith, K. G., & Taylor, M. S. (2008). After the fall: Reintegrating the corrupt organization. *Academy of Management Review*, 33(3), 730–749. <https://doi.org/10.5465/amr.2008.32465757>

- Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, 53(5), 1131–1152. <https://doi.org/10.5465/amj.2010.54533222>
- Piazza, A., & Perretti, F. (2015). Categorical stigma and firm disengagement: Nuclear power generation in the United States, 1970–2000. *Organization Science*, 26(3), 724–742. <https://doi.org/10.1287/orsc.2014.0964>
- Pollock, T. G., Lashley, K., Rindova, V. P., & Han, J.-H. (2019). Which of these things are not like the others? Comparing the rational, emotional, and moral aspects of reputation, status, celebrity, and stigma. *Academy of Management Annals*, 13(2), 444–478. <https://doi.org/10.5465/annals.2017.0086>
- Pollock, T. G., & Rindova, V. P. (2003). Media legitimation effects in the market for initial public offerings. *Academy of Management Journal*, 46(5), 631–642. <https://doi.org/10.5465/30040654>
- Pollock, T. G., Rindova, V. P., & Maggitti, P. G. (2008). Market watch: Information and availability cascades among the media and investors in the U.S. IPO market. *Academy of Management Journal*, 51(2), 335–358. <https://doi.org/10.5465/amj.2008.31767275>
- Richardson, J. E. (2017). *Analysing Newspapers: An Approach from Critical Discourse Analysis*. Bloomsbury Publishing.
- Rindova, V. P., Pollock, T. G., & Hayward, M. L. A. (2006). Celebrity firms: The social construction of market popularity. *Academy of Management Review*, 31(1), 50–71. <https://doi.org/10.5465/amr.2006.19379624>
- Roulet, T. J., & Clemente, M. (2018). Let's open the media's black box: The media as a set of heterogeneous actors and not only as a homogenous ensemble. *Academy of Management Review*, 43(2), 327–329. <https://doi.org/10.5465/amr.2016.0537>

- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4), 296–320. [https://doi.org/10.1207/S15327957PSPR0504\\_2](https://doi.org/10.1207/S15327957PSPR0504_2)
- Sayre, B., Bode, L., Shah, D., Wilcox, D., & Shah, C. (2010). Agenda setting in a digital age: Tracking attention to california proposition 8 in social media, online news and conventional news. *Policy & Internet*, 2(2), 7–32. <https://doi.org/10.2202/1944-2866.1040>
- Schnietz, K. E., & Epstein, M. J. (2005). Exploring the financial value of a reputation for corporate social responsibility during a crisis. *Corporate Reputation Review*, 7(4), 327–345. <https://doi.org/10.1057/palgrave.crr.1540230>
- Seguin, C. (2016). Cascades of coverage: Dynamics of media attention to social movement organizations. *Social Forces*, 94(3), 997–1020. <https://doi.org/10.1093/sf/sov085>
- Semenova, N., & Hassel, L. G. (2015). On the validity of environmental performance metrics. *Journal of Business Ethics*, 132(2), 249–258. <https://doi.org/10.1007/s10551-014-2323-4>
- Shipilov, A. V., Greve, H. R., & Rowley, T. J. (2019). Is all publicity good publicity? The impact of direct and indirect media pressure on the adoption of governance practices. *Strategic Management Journal*, 40(9), 1368–1393. <https://doi.org/10.1002/smj.3030>
- Shoemaker, P. J. (2006). News and newsworthiness: A commentary. *Communications*, 31(1), 105–111. <https://doi.org/10.1515/COMMUN.2006.007>
- Shoemaker, P. J., & Cohen, A. A. (2012). *News Around the World: Content, Practitioners, and the Public*. Routledge.
- Shoemaker, P. J., Eichholz, M., Kim, E., & Wrigley, B. (2001). Individual and routine forces in gatekeeping. *Journalism & Mass Communication Quarterly*, 78(2), 233–246. <https://doi.org/10.1177/107769900107800202>

- Shoemaker, P. J., & Vos, T. (2009). *Gatekeeping theory*. Routledge.
- Soroka, S. N. (2006). Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics*, 68(2), 372–385. <https://doi.org/10.1111/j.1468-2508.2006.00413.x>
- Soroka, S. N. (2012). The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, 74(2), 514–528. <https://doi.org/10.1017/S002238161100171X>
- Soroka, S. N., & Carbone, M. (2016). Gatekeeping, Technology, and Polarization. In S. N. Soroka & M. Carbone, *Oxford Research Encyclopedia of Politics*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190228637.013.43>
- Soroka, S. N., & Krupnikov, Y. (2021). *The increasing viability of good news*. Cambridge University Press.
- Soroka, S. N., & McAdams, S. (2015). News, Politics, and Negativity. *Political Communication*, 32(1), 1–22. <https://doi.org/10.1080/10584609.2014.881942>
- Staab, J. F. (1990). The role of news factors in news selection: A theoretical reconsideration. *European Journal of Communication*, 5(4), 423–443. <https://doi.org/10.1177/0267323190005004003>
- Tan, D. (2016). Making the news: Heterogeneous media coverage and corporate litigation: media coverage and litigation. *Strategic Management Journal*, 37(7), 1341–1353. <https://doi.org/10.1002/smj.2390>
- Thøgersen, J. (2006). Media attention and the market for ‘green’ consumer products. *Business Strategy and the Environment*, 15(3), 145–156. <https://doi.org/10.1002/bse.521>
- Vasi, I. B., & King, B. G. (2012). Social movements, risk perceptions, and economic outcomes: The effect of primary and secondary stakeholder activism on firms’ perceived

- environmental risk and financial performance. *American Sociological Review*, 77(4), 573–596. <https://doi.org/10.1177/0003122412448796>
- Vergne, J.-P. (2011). Toward a new measure of organizational legitimacy: Method, validation, and illustration. *Organizational Research Methods*, 14(3), 484–502. <https://doi.org/10.1177/1094428109359811>
- Vliegenthart, R., & Walgrave, S. (2008). The contingency of intermedia agenda setting: A longitudinal study in Belgium. *Journalism & Mass Communication Quarterly*, 85(4), 860–877. <https://doi.org/10.1177/107769900808500409>
- Waddock, S. A., & Graves, S. B. (1997). The corporate social performance-financial performance link. *Strategic Management Journal*, 18(4), 303–319. [https://doi.org/10.1002/\(SICI\)1097-0266\(199704\)18:4<303::AID-SMJ869>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-0266(199704)18:4<303::AID-SMJ869>3.0.CO;2-G)
- Wartick, S. L. (1992). The relationship between intense media exposure and change in corporate reputation. *Business & Society*, 31(1), 33–49. <https://doi.org/10.1177/000765039203100104>
- Wei, J., Ouyang, Z., & Chen, H. A. (2017). Well known or well liked? The effects of corporate reputation on firm value at the onset of a corporate crisis: the effects of corporate reputation on firm value. *Strategic Management Journal*, 38(10), 2103–2120. <https://doi.org/10.1002/smj.2639>
- Westphal, J. D., & Deephouse, D. L. (2011). Avoiding bad press: Interpersonal influence in relations between CEOs and journalists and the consequences for press reporting about firms and their leadership. *Organization Science*, 22(4), 1061–1086. <https://doi.org/10.1287/orsc.1100.0563>

- Wickman, C. (2014). Rhetorical framing in corporate press releases: The case of British petroleum and the Gulf oil spill. *Environmental Communication*, 8(1), 3–20. <https://doi.org/10.1080/17524032.2013.816329>
- Wilkinson, K. T., & Merle, P. F. (2013). The merits and challenges of using business press and trade journal reports in academic research on media industries: Using business reports in media research. *Communication, Culture & Critique*, 6(3), 415–431. <https://doi.org/10.1111/cccr.12019>
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- Zandberg, E., Meyers, O., & Neiger, M. (2012). Past continuous: Newsworthiness and the shaping of collective memory. *Critical Studies in Media Communication*, 29(1), 65–79. <https://doi.org/10.1080/15295036.2011.647042>
- Zavyalova, A., Pfarrer, M. D., & Reger, R. K. (2017). Celebrity and Infamy? The Consequences of Media Narratives About Organizational Identity. *Academy of Management Review*, 42(3), 461–480. <https://doi.org/10.5465/amr.2014.0037>
- Zavyalova, A., Pfarrer, M. D., Reger, R. K., & Hubbard, T. D. (2016). Reputation as a Benefit and a Burden? How Stakeholders' Organizational Identification Affects the Role of Reputation Following a Negative Event. *Academy of Management Journal*, 59(1), 253–276. <https://doi.org/10.5465/amj.2013.0611>
- Zavyalova, A., Pfarrer, M. D., Reger, R. K., & Shapiro, D. L. (2012). Managing the message: The effects of firm actions and industry spillovers on media coverage following wrongdoing. *Academy of Management Journal*, 55(5), 1079–1101. <https://doi.org/10.5465/amj.2010.0608>
- Zelizer, B. (2008). Why memory's work on journalism does not reflect journalism's work on memory. *Memory Studies*, 1(1), 79–87. <https://doi.org/10.1177/1750698007083891>



Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *Advances in Information Retrieval* (Vol. 6611, pp. 338–349). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-20161-5\\_34](https://doi.org/10.1007/978-3-642-20161-5_34)

## To Omit or to Include?

### The Frugal and Prolific Perspectives on Control Variable Use.

Fabian Mändli, HEC Lausanne

Mikko Rönkkö, Jyväskylä University

#### Abstract

Over the recent years, two perspectives on control variable use have emerged in management research: The first originates largely from within the management discipline, and argues to remain *frugal*, to use control variables as sparsely as possible. The second roots in econometrics textbooks and argues to be *prolific*, to be generous in control variable inclusion to not risk omitted variable bias and because including irrelevant exogenous variables has little consequences for regression results. We present two reviews that show that the frugal perspective is getting increasingly popular in research practice, while the prolific perspective has received little explicit attention. We summarize the key arguments of both perspectives and put them to a test in three Monte Carlo simulations. Our results challenge the two recommendations of the frugal perspective of “omitting impotent controls” and “avoiding proxies” but show the detrimental effects of including endogenous controls (bad controls). We recommend considering the control variable selection problem from the perspective of endogeneity and selecting controls based on theory and by using causal graphs, instead of focusing on the many or few question.

**Keywords:** *control variables, correlation, causality, endogeneity, Monte Carlo simulation*

## Introduction

Control variables are critically important for making causal claims in non-experimental management research and can be useful for increasing the precision and statistical power of experimental studies (Deaton & Cartwright, 2018; Hernández et al., 2004). Controls should be chosen based on existing theory so that alternative explanations can be ruled out. But how should controls be chosen when this theoretical guidance is not clear? Should researchers “when in doubt, leave them out” (Carlson & Wu, 2012, p. 413) or “err on the side of caution by including more than fewer control variables” (Antonakis et al., 2010, p. 1092)? Recently, two distinct perspectives have emerged: The *frugal perspective* holds that if a researcher is not sure about whether a control variable should be included in a model, it should be left out (e.g., Atinc et al., 2012; Bernerth et al., 2018; Bernerth & Aguinis, 2016; Carlson & Wu, 2012). In contrast, the *prolific perspective* emphasizes that more controls are better than too few (Antonakis et al., 2010) as this reduces the probability of omitted variable bias.

The frugal perspective originates from Becker (2005) and is often summarized by the phrase “When in doubt, leave them out.”, coined by Carlson and Wu (Carlson & Wu, 2012). While this perspective is presented in many guideline-type articles (Becker et al., 2016; O’Neill et al., 2014; Schjoedt & Bird, 2014) within the management discipline, it seems mostly absent in the broader research methods literature. In contrast, the prolific perspective builds on the econometric concept of irrelevant regressors and the proof that including such variables will not bias regression coefficients (Wooldridge, 2013, p. 88). This perspective is also advocated in some general research methods texts. For example, Singleton and Straits (2018) recommend that “Circumstances seldom allow to control for all variables; researchers attempt to control the effects of as many as possible. The greater the number of variables that are controlled without altering a relationship, the greater the likelihood that the relationship is not spurious.” (p. 102).

The two perspectives have been noted in the literature (Bernerth et al., 2018, p. 154; Green et al., 2016, p. 422), but thus far their merits have not been analyzed. This is what we do. Both perspectives largely agree on that control variable selection is important and should be based on theory, and that reporting should be more transparent. Yet, they differ in the overall recommendation on how liberally control variables should be included. The frugal perspective also proposes empirical rules that we argue are problematic. After introducing the perspectives, we show through two systematic reviews that the frugal perspective is getting more popular and that the prolific perspective has been seldomly explicitly applied in management research. Thereafter, we assess three specific empirical rules with a set of Monte-Carlo simulations. We find that dropping “impotent controls” and “avoiding proxies” can bias estimates, whereas inclusion of irrelevant variables has little negative consequences. We conclude that control variables should be chosen solely based on theory and the empirical rules should be abandoned.

### **Control Variables in Management Research**

Management research should make causal claims as they are important for society (Antonakis et al., 2010). This is challenging because causality is unobservable (Hitchcock, 2010; Jaccard & Jacoby, 2020, pp. 153–154), and can only be inferred indirectly using appropriate research designs. To claim causality, researchers must demonstrate: 1) association between the assumed cause and effect, 2) direction of influence, and 3) elimination of alternative explanations (Antonakis et al., 2010; Singleton & Straits, 2018, Chapter 4). The third step is the hardest part. Experiments where rival explanations are eliminated by randomization are considered the gold standard (Antonakis et al., 2010; Heckman, 2008) but they are often costly or infeasible (Cameron & Trivedi, 2005, p. 96). Consequently, statistical models (e.g., regression) that use control variables to account for alternative explanations have become the dominant strategy in management research. Next, we describe the two perspectives on control variable selection.

## **The Prolific Perspective to Control Variable Inclusion**

The main idea of prolific perspective is that controls should be used liberally to prevent omitted variable bias. This is repeated in multiple econometrics books. For example, Cameron and Trivedi (2005, p. 93) state that “Too many regressors cause little harm, but too few regressors can lead to inconsistency”, Greene (2012, p. 178) says that “Omitting variables from the equation seems generally to be the worse of the two errors”, and there are many similar examples (e.g., Berry & Feldman, 1985, pp. 21–22; Schroeder et al., 2017, p. 71; Zax, 2011, p. 465). However, in the recent literature on control variables in management research, the prolific perspective has received little attention. The only explicit recommendation that we found in the management literature is to “err on the side of caution by including more than fewer control variables” by Antonakis and coauthors (2010, p. 1092).

The prolific perspective has three main recommendations: 1) omitted variable bias should be avoided by including relevant controls, 2) inclusion of irrelevant controls has little negative consequences, but 3) overcontrolling by including endogenous controls should be avoided.

The recommendation related to omitted variables is straightforward: If a control variable is a cause of the dependent variable and is correlated with at least one of the independent variables, omitting the control has been proven to create endogeneity in the model, biasing estimates (Wooldridge, 2013, p. 88). Because omitted variable bias is a serious threat for inference, control variables that are causes of the dependent variable and correlated with the independent variables should be included.

There is also little harm in including irrelevant controls, which do not have effects on the dependent variable (Basu, 2020, p. 211). More specifically, the prolific perspective states that while irrelevant variables can reduce efficiency (precision of estimates), “reduced efficiency [...] is a cheap price to pay when consistency is at stake” (Antonakis et al., 2010, p.

1092). This is supported by the proof (Cameron & Trivedi, 2005, p. 93; Wooldridge, 2013, p. 87, Theorem 3.1) that estimates remain unbiased when irrelevant variables are included in a model. Thus, if there are potentially relevant control variables, it is safer to include them in the model, at worst they turn out irrelevant. It is essential to add that in the econometrics literature, the concept of “irrelevant regressor” itself is applied to exogenous variables only (Wooldridge, 2013, p. 88).<sup>1</sup>

Notably, econometricists do not recommend a “kitchen sink” (Greene, 2012, p. 179) perspective towards control variable use, where control variables would be wildly included into a model to prevent bias at all costs (Wooldridge, 2013, p. 88). For example, if a control variable is a mediator on a causal path, then (over-)controlling for this variable biases estimates of the total causal effect (Li, 2021), because it is endogenous (Antonakis et al., 2010, p. 1090). This part overlaps with the frugal perspective’s recommendation of being cautious about controlling for potential endogenous variables, but it is featured a lot less prominently.

The prolific perspective can be summarized along the lines of Wooldridge (2013, pp. 98–99) as a trade-off between bias and variance: Control variables that are potential omitted variables should be included to prevent inconsistent and biased estimates. The consequence of such inclusion is reduced efficiency which can be mitigated by increasing sample size.

### **The Frugal Perspective to Control Variable Inclusion**

Many recent guidelines (Aguinis & Vandenberg, 2014; Becker et al., 2016; O’Neill et al., 2014; Schjoedt & Bird, 2014) warn about including too many controls in models. This advice comes in two forms: a) reasons to be cautious with including controls generally and b) specific recommendations or rules for when controls should be left out.

---

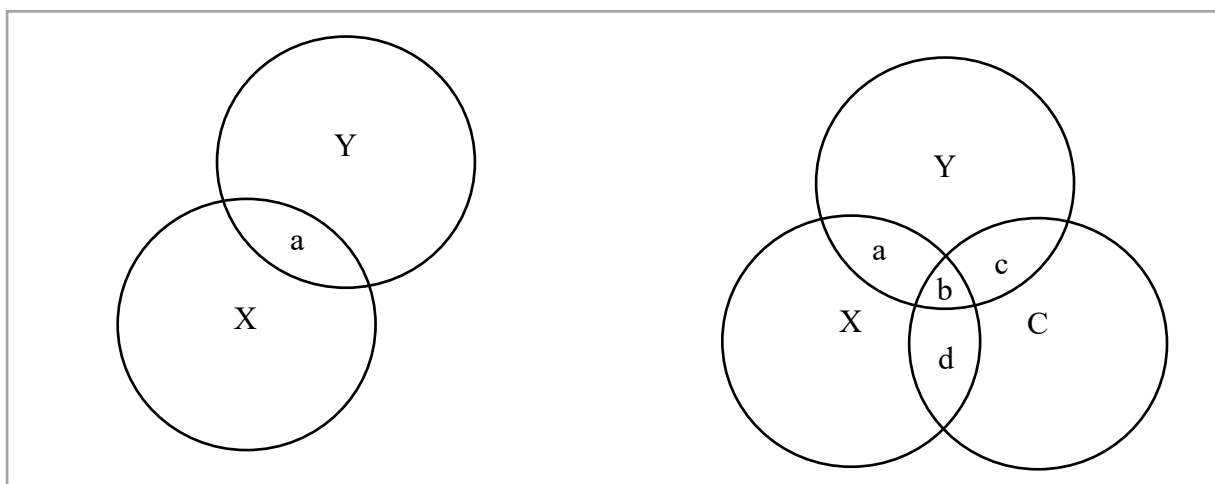
<sup>1</sup> While the importance of exogeneity is clearly stated in Antonakis and colleagues (2010), when consulting the quotes only, this is easily overlooked.

### *Reasons to be Cautious with Controls*

Perhaps the most intuitively appealing reason to be cautious with controls is the claim that “a model including control variables is no longer investigating the relationship between a predictor and a criterion, but rather the relationship between a new *residual* predictor and the criterion” (Bernerth & Aguinis, 2016, p. 231). This point can be illustrated with Venn diagrams (Breugh, 2008), where the total variance of each variable is represented by a circle and overlapping areas of the circles represent shared variance (or squared semi-partial correlations; Cohen et al., 2003, sec. 3.3.2). In the left panel of Figure 1 below, the area *a* indicates the shared variance between the dependent variable *Y* and the focal predictor *X*. In the right panel, including a control *C* eliminates the shared variance that overlaps with the predictor (*d*), the dependent variable (*c*), and what they share (*b*), thus allowing to estimate the unique shared variance (*a*) between *X* and *Y*. The key concerns in the frugal perspective are that the interpretation of *X* changes by using statistical control or that by including more controls there is no variance left to explain, as the size of area (*a*) declines.

**Figure 1**

*Venn diagrams explaining the shared variance between the independent (*X*) and dependent variable (*Y*), without and with a control variable (*C*). When controlling for *C*, the unique shared variance (*a*) between the independent (*X*) and the dependent variable (*Y*) is decreasing by (*b*).*



The residual argument is technically correct but leads to an incorrect conclusion in this case: The core idea of regression analysis is that it enables to “keep other factors fixed” (Wooldridge, 2013, p. 77), by removing their variation from the analysis (Greene, 2012, p. 76). For example, suppose our main variable of interest is CEO gender and we wanted to study its effect on firm performance. If CEO gender correlates with industry and industry also affects firm performance, this produces a spurious correlation that needs to be controlled to claim a causal effect of CEO gender on firm performance. Controlling for industry, we estimate the effect of CEO gender on firm performance as if all firms were in the same industry. That is, we eliminate all between industry variance from both CEO gender and firm performance.

To understand why reducing variance this way is not a problem for interpretation, we can compare regression against other approaches for reducing variation, such as using matched samples or sampling just from a single industry (Morgan & Winship, 2007, Chapter 4). If regression assumptions hold (i.e., the effect does not vary between industries), controlling for industry is equivalent to reducing variation by studying just one industry or doing the same through matched samples (Morgan & Winship, 2007, sec. 5.3). Thus, arguing that statistical controlling changes the meaning of the variables would logically imply that the same applies when variance is reduced by sampling, which is not the case. Indeed, regardless of how it is done, be it with control variables, matching, or sampling, reducing variance due to extraneous factors is a solid research design principle (Singleton & Straits, 2018, pp. 36–39, 89, 101–102).

A related concern is that statistical adjustments create “fictional people”. This was pointed out by Breugh (2008), who claims that this limits generalizability citing Meehl (1970). Yet, Meehl (1970) did not argue that creating “fictional people”, or more formally *counterfactuals* (p. 401), would limit generalizability, but that causal inferences rest on their proper



construction<sup>2</sup>. Counterfactuals are not a problem and in fact the current literature on causal analysis (Huntington-Klein, 2022; Morgan & Winship, 2007; Pearl, 2012) rests on them. That is, a causal effect is defined as a comparison between two potential outcomes, one in which a case received a treatment and another where the same case was not treated. Because we observe each case only as treated or untreated, one of the two potential outcomes is observed and the other remains a counterfactual that must be estimated.

In the context of the CEO gender example we discuss above, following Breugh (2008), one might ask if it makes sense to try to separate the effects of CEO gender and industry given that the two variables are correlated in practice. The answer to this is yes: firm boards would want to know the effect of hiring a female CEO holding industry constant, because firms rarely switch CEO and industry at the same time. That is, it often makes sense to assume that some variables do not change because they are not a part of the decision that a firm, policy maker, or leader typically would take. In our example, controlling for industry is a safe bet, because the CEO gender effect is a within-firm effect and firms rarely change their industry classifications making industry an exogenous variable. However, this does not mean that all counterfactuals make sense, or all controls are good, an issue that we return to later.

Another concern relates to the precision of estimates. Becker and colleagues (2016) urge to remain cautious because “including large numbers of [control variables] reduces

---

<sup>2</sup> Meehl’s (1970) article addresses three problems in ex post facto design using matched groups. The first problem is “systematic unmatched”, which means that when we match on one variable, the groups might become unmatched on another variable. However, it is not clear how common this might be and more importantly, Meehl neither shows nor proves that this would be problematic for the causal estimate of interest (Lund, 1981). The second problem is “unrepresentative subpopulation”. For example, if we are studying the effect of job promotions on job satisfaction in a sample where only men receive promotions, matching on gender would mean that we are studying only the subpopulation consisting of men and this affects generalizability. But the concern applies regardless of how the data are analyzed: if women do not receive promotions, we cannot say anything about the effect of these non-existent promotions on their job satisfaction. In this extreme example, the causal effect for women is undefined. In less extreme cases, modern matching methods can handle the “unrepresentative subpopulation problem” (Morgan & Winship, 2007, sec. 4.2.2.). The third problem is “causal-arrow ambiguity” and refers to endogenous or bad controls, which is a severe problem that we address later in the article. The “fictional people” argument is not directly related to these three arguments but is used to argue that the three concerns apply also to regression and not only matched samples.

degrees of freedom, [...] this will increase standard errors and potentially decrease the power of the test for a given independent variable.” (p. 159). This is incorrect in two different ways. First, the variance of the regression estimates depends only on the total sample variation of the independent variables, error variance, and correlation between the independent variables (Wooldridge, 2013, theorem 3.2)<sup>3</sup> and not on degrees of freedom. Second, adding controls can also decrease standard errors and increase statistical power by reducing error variance. This is the reason why controls are often used in experiments (Deaton & Cartwright, 2018; Hernández et al., 2004).

The mechanism through which control variables can make estimates less precise is multicollinearity, which is also sometimes mentioned in this context (e.g., Nielsen & Raswant, 2018). What this means is that using control variables that are highly correlated with the focal variables makes it more challenging to identify which part of the total variance is explained by the focal variables and which part is explained by the control variables, decreasing the precision of the estimates (Greene, 2012, p. 130) and thus reducing statistical power. While omitting such controls would solve this problem, it introduces omitted variable bias. Instead, if possible, researchers should increase precision by increasing sample size (Greene, 2012, p. 131; Wooldridge, 2013, pp. 94–98).<sup>4</sup>

---

<sup>3</sup> Degrees of freedom appear in the formula for estimated error variance (Wooldridge, 2013, eq. 3.56). However, reducing degrees of freedom by adding variables does not affect the expected value of the estimated error variance because the SSR (sum of squares residual) that appears in the formula is a biased estimator of the total variation of the error term and the degree of bias depends on the number of predictors. Degrees of freedom has a direct impact on statistical power because it determines which t-distribution is used for testing the regression estimates, but this effect is minor except in very small samples.

<sup>4</sup> Other alternatives include simplifying the research question and using small-sample estimation techniques. Wooldridge (2013, pp. 96–97) gives the following example of simplifying the question: Assume that we are interested in how different school expenditure categories affect student performance. If the sample size is too small to answer the question reliably, we can simply sum (make an index) of the expenditure categories and ask a simpler question of how school expenditures generally affect student performance without differentiating between the categories. If simplifying the question is not appropriate, large models can be estimated in steps, regression-type models can be estimated with shrinkage estimators, or Bayesian priors can be applied. For an overview of these techniques, see (Schoot & Miočević, 2020). If neither of these approaches is applicable and including a control is empirically impossible, the omitted control and the reasons of omission should be documented and the effect of the omission should be assessed with a sensitivity analysis (Hünernund et al., 2022).

A final argument for being cautious about control variables relates to endogenous or bad controls. Bad controls (or confounders), contrary to good controls (or deconfounders), are control variables that bring estimates further away from their true population value (thus increase bias) when included (Cinelli et al., 2022). A control variable is endogenous or bad if it depends on an independent variable of interest, the dependent variable, or shares an unobserved cause with the dependent variable (Antonakis et al., 2010; Cinelli et al., 2022). Becker and colleagues refer to this as the uncertain association between control variables and the other variables in a model (Becker et al., 2016, p. 159). Although not explicitly discussing endogenous or bad controls, they mention controls could lead to spurious associations. Indeed, as discussed by Spector and Brannick (2011), adding an endogenous control into the model would bias estimates and hence such variables should not be used.

We give examples of bad controls to illustrate the point. For instance, if we want to study the overall effect of firm environmental performance on media exposure<sup>5</sup>, we probably should not control for profitability, as more profitable firms have more funds available to improve their environmental credentials (Adams & Hardwick, 1998; Waddock & Graves, 1997), and vice-versa environmental performance can boost profitability (Russo & Fouts, 1997), while at the same time environmental performance (Aerts & Cormier, 2009) and profitability (Dai et al., 2015) both affect media exposure. Profitability would in this case represent an endogenous control. Another typical example of a bad control is controlling for a mediator (Hünermund et al., 2022; Wysocki et al., 2022). Also, if we want to study the overall effect of product innovation on profitability, we probably should not control for sales because increasing sales is one of the main mechanisms through which new products can affect profitability. Similarly, if we want to study the overall effects of leader characteristics on employee retention,

---

<sup>5</sup> As there is most likely a reverse causality between media exposure and environmental performance, a panel-data approach would be desirable.

we probably should not control for employee’s job satisfaction because this too is a likely mechanism. In these two cases, we would be asking how much product innovation affects profitability if it did not affect sales and how much leader characteristics affect retention if they did not affect satisfaction. Both questions would be illogical (see also Wooldridge, 2013, pp. 205–206). However, if we want to study a specific causal mechanism instead of overall causal effects, we need to control for mediators to rule out potential alternative mechanisms.

### ***Specific Recommendations to Leave Out Control Variables***

Beyond the general recommendations to exercise caution when including controls, the frugal perspective also provides three specific recommendations: avoiding impotent controls, avoiding proxies, and running results with and without control variables, which we discuss next.

A control is said to be “impotent” when it has “little or no relationship with the [dependent variable] (e.g.,  $|r| < .10$ )” (Becker et al., 2016, p. 160) and the specific recommendation is that such controls should be dropped. This recommendation is problematic because the correlation between two variables is a sum of a possible causal relationship and any spurious influences (Cohen et al., 2003, Chapter 12). That is, in a model with two predictors, the correlation between control  $C$  and dependent variable  $Y$  depends on the correlation between the control  $C$  and interesting variable  $X$  as well as their standardized regression coefficients  $\beta$  (e.g., Cinelli et al., 2022, eq. A.3):

$$corr_{C,Y} = \beta_C + \beta_X corr_{C,X}.$$

As shown in Table 1 below, an impotent control ( $corr_{C,Y} = 0$ ) can thus only occur in three scenarios: 1) If  $C$  is uncorrelated with  $X$  and has no effect on  $Y$ , 2) neither  $X$  or  $C$  have an effect on  $Y$ , or 3)  $C$  is correlated with  $X$  and the product of the estimated effect of  $X$  and this correlation is equal in magnitude to the estimated effect of  $C$  but in opposite directions so that they offset each other. Thus, it is entirely possible that even if a control variable is not correlated

with the dependent variable, the variables are causally related, and the control needs to be controlled for.

**Table 1**

*Comparison of Four Scenarios that Produce Impotent or Irrelevant Controls*

Scenario	Nature of control variable	Effects of dropping control
<u>1: Uncorrelated control:</u> $corr_{C,X} = 0,$ $corr_{C,Y} = 0,$ $\beta_C = 0$	Irrelevant Impotent	No effects on bias or efficiency.
<u>2: No effects:</u> $corr_{C,Y} = 0,$ $\beta_C = 0, \beta_X = 0.$	Irrelevant Impotent	No effects on bias. Efficiency can increase.
<u>3: Offsetting effects:</u> $corr_{C,X} \neq 0,$ $corr_{C,Y} = 0,$ $\beta_C \neq 0, \beta_X \neq 0.$	Not irrelevant Impotent	Bias increases. Efficiency can increase or decrease.
<u>4: Irrelevant control:</u> $corr_{C,X} \neq 0,$ $corr_{C,Y} \neq 0,$ $\beta_C = 0$	Irrelevant Not impotent	No effects on bias. Efficiency increases.

*Note.* Bivariate regression where Y is the dependent variable, X is a variable of interest and C is a control variable.

The effects of dropping impotent controls differ from dropping irrelevant controls. As explained in the section on the prolific perspective, the omission or inclusion of irrelevant variables does not affect bias of regression estimates, but it may affect their efficiency. In Scenario 1 in Table 1 above, there is no effect on efficiency, as neither the variance of error term nor the correlation between the independent variables is affected and these are the only mechanism through which efficiency can be affected (Wooldridge, 2013, theorem 3.2). In Scenario 2, efficiency will increase if X and C are correlated, but it is of little use because there is no effect to be detected. In Scenario 3, sometimes called classical suppression effect (Friedman & Wall, 2005; Lewis & Escobar, 1986; Smith et al., 1992), the causal effect of X and the spurious

correlation due to C offset each other. Because C influences Y, its omission would lead to omitted variable bias. Scenario 4 is a typical example of irrelevant controls where excluding the control can be useful to increase precision and statistical power (e.g., Wooldridge, 2013, p. 88). Yet, in this case the control is not impotent and would be kept in the model if the impotent control rule was followed. To summarize, Table 1 shows that the “dropping impotent control” rule is either useless (Scenarios 1 and 2) or harmful (Scenario 3) and further would not lead to dropping controls when it provides a benefit (Scenario 4).

We use an example of employee tardiness, conscientiousness, and distance to work (Becker et al., 2016, p. 160), to show that omitting an impotent control variable can bias regression estimates. Consider the following setup where the units are standard deviations:

- a) One unit increase in home’s distance from work increases tardiness by one unit.
- b) More conscientious workers tend to live further from work so that conscientiousness and distance from work correlate at 0.5.
- c) A one-unit increase in conscientiousness decreases tardiness by two units.

Consider that we are interested in whether employee conscientiousness affects tardiness and use distance to work as a control. In this scenario, distance from work is an impotent control because the effect of distance on tardiness (+1) is completely canceled out by the effect of more conscientious workers living further from work ( $-2 \times 0.5 = -1$ ). However, because distance to work has an effect, omitting it from the analysis would lead us to incorrectly conclude that the effect of one additional unit of consciousness decreases tardiness by one and a half units instead of two units. As this example shows, the relevant criterion is not whether a control correlates with the dependent variable, but whether it has a causal effect.

The second specific recommendation is that proxies should be avoided (Becker et al., 2016; Spector & Brannick, 2011). A proxy variable approximates a variable that researchers would like to control for, but cannot observe directly (Greene, 2012, sec. 8.5.3). For example

patent data, product launches as well surveys among managers have all been used as proxies for innovativeness (Jensen & Webster, 2009), or sales, assets, or market value as proxies for firm size (Al-Khazali & Zoubi, 2005). Some examples from textbooks include years of schooling as a proxy for education or IQ as a proxy for ability or intelligence (Greene, 2012, pp. 221, 242; Wooldridge, 2010, p. 68).

The idea that proxies can be problematic was introduced to the control variable literature by Breugh (2008), who explained that “The problem with controlling for proxy variables is that a researcher almost never knows the strength of the relationship between a proxy variable and the underlying causal variable. Thus, the researcher cannot determine to what extent he or she has controlled for the nuisance variable of interest.” (p. 291) Becker and colleagues (2016) further point out that using proxies can lead to problems also “because the proxy might relate to other variables in a way that the CV of interest does not” and thus “controlling for the proxy may control for a host of unintended variables that have substantive effects that the researcher does not wish to remove” (p.161).

We use an example from Greene (2012, p. 243) to discuss proxies. Consider estimating the effects of education on earnings but instead of education, we measure years of schooling:

$$\text{years of schooling} = \text{education} + u$$

where  $u$  is random error. The original concern by Breugh (2008) was that the association between the construct and the proxy might be weak (i.e.,  $u$  has large variance). The further concern by Becker and colleagues (2016) is that  $u$  might be related to the other variables in the model in unintended ways. If  $u$  is uncorrelated with *education* and other variables in the model, increasing variance of  $u$  means that the proxy eliminates decreasing parts of the variance of the construct it approximates (Wooldridge, 2013, pp. 320–323). Nevertheless, Aigner (1974) shows that using a proxy is still desirable because the bias caused by measurement error is smaller than the omitted variable problem. However, if  $u$  is correlated with other variables in

the model, the ignorability or redundancy assumption of proxy variables is violated (Wooldridge, 2010, pp. 67–68) producing an imperfect proxy. While imperfect proxies can reduce bias, they do not always do so (Wooldridge, 2010, pp. 69, 72) as Becker and colleagues (2016) note.

There are cases where most of the control variables in a model might be proxies. For example in the main analysis of Mändli et al. (2023), close consideration yields that essentially all control variables in the model are proxies: The MSCI ESG Environmental pillar score serves as a proxy for *environmental policy*, assets is used as proxy for firm *size*, net sales for *size* in the product market and *performance*, cashflow for *availability of funds*, ROA for *profitability*, and the ratio of total amount of debt and total amount of assets for *leverage*. While it would be desirable to use more precise measures for environmental policy for example, commonly used ESG ratings are all plagued by similar constraints (Berg et al., 2022; Semenova & Hassel, 2015) and most likely do not capture environmental performance perfectly. This on the other hand, does not mean that they should not be used as control variables. ESG performance would present an omitted variable in many models investigating ESG topics, thus these imperfect proxies should be used rather than omitted. As the above mentioned control variables have been identified using prior theory, it is highly likely that most control variables used in Mändli et al. (2023) would be omitted variables if dropped. It is also quite likely that even if they are imperfect proxies, controlling for them reduces overall bias.

The third recommendation is that regressions should be reported with and without control variables to assess the robustness of results and the impact of control variables on the results (e.g., Becker et al., 2016; Bernerth et al., 2018; Carlson & Wu, 2012), and if results are the same, report ones without control variables (Becker, 2005). While this practice might be useful in some cases, the causal effect of variables is not assessed properly in other instances if control variables are not included (Sturman et al., 2022). If a researcher trusts a control



variable should be in the model, it does not make sense to report results without it because this increases the risk of omitted variable bias.

In short, the frugal perspective claims that control variables partial out variance, potentially change the interpretation of the variables, reduce available degrees of freedom, and there is a risk of including endogenous or bad controls leading to spurious associations. Because researchers can rarely be certain that a control would not cause any problems, they should follow the guideline “When in doubt, leave them out!” (Becker et al., 2016, p. 158). This is complemented by the specific recommendations for not using control variables that are either “impotent” or that are proxies and reporting results with and without control variables.

### **The Impact of the Frugal and Prolific Perspectives on the Empirical Literature**

To understand the impact that the two perspectives have had on the management literature, we did two systematic reviews. The first review is a citation analysis that investigates the popularity of the frugal perspective over time. Doing the same for the prolific perspective was not possible because this perspective has no central source(s)<sup>6</sup>. Instead, we use a second systematic review to compare the relative impact of both perspectives in management research.

#### ***Literature Analysis 1: The Frugal Perspective Over Time***

The origin of the frugal perspective is Becker’s (2005) work, and we therefore started by reviewing all articles that cite this article or any of the other guidelines extending this work (Aguinis & Vandenberg, 2014; Becker, 2005; Becker et al., 2016; Bernerth et al., 2018; Carlson & Wu, 2012; O’Neill et al., 2014; Schjoedt & Bird, 2014). Using ISI Web of Science, we found 1,589 articles between 2006 and 2021 where at least one of these seven sources was cited.

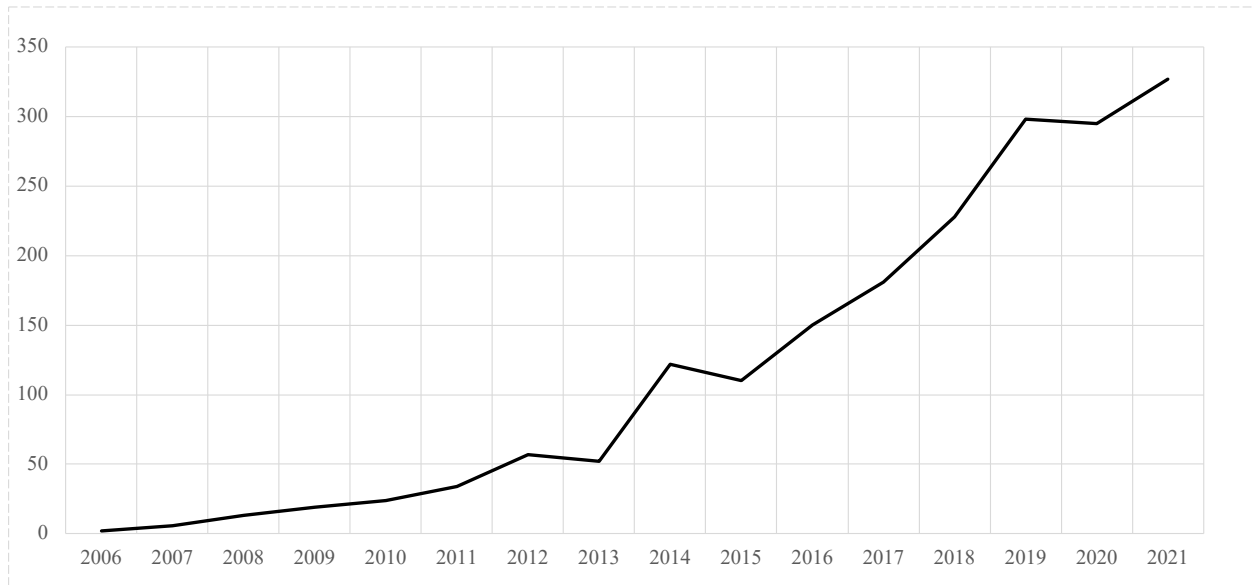
---

<sup>6</sup> We cite Antonakis et al. (2010) as an article advocating the prolific perspective, but it is more often cited for other reasons. To show this, we took a random sample of 20 from the total of 912 articles that contained the term “control variable(s)” and cited Antonakis et al. (2010) within the 2010-2021 period. None of these referred to the control variable recommendations the paper makes. Moreover, if cited correctly, one would not cite Antonakis et al. (2010) but instead the econometrics textbooks that the Antonakis paper cites as a source of the prolific perspective.

Figure 2 presents these articles over time showing the frugal perspective has gained significant traction.

## Figure 2

*Annual Number of Citations to Frugal Perspective Articles. Data from ISI Web of Science, 2006-2021.*



To better understand how the frugal perspective affects research, the articles were coded in more detail by the first author. The second author coded a random sample of 30 articles (Krippendorff's  $\alpha = 0.78$ ). Table 2 below presents an overview of this analysis. 64.1% (898) of these articles were either applying the frugal perspective's recommendations or supporting their use, while 35.9% (503) were not applying the frugal perspective but cited the guideline-articles for their recommendations on control variable selection and reporting practices. Merely one applied paper (Sudzina, 2018, p. 68) was critical of the recommendations and did not follow them. To conclude, the impact of the frugal perspective and its associated recommendations is substantial.

**Table 2***Number of Papers Citing Methodological Papers of the Frugal Perspective.*

Reference category	N (%)	Example excerpts
1) Dropping control variables due to insignificant correlation or effect.	526 (37.5)	<p>"To prevent reduction in statistical power, employee age, tenure, and support for innovation climate were not included in the final data analysis because they were not significantly related to domain-relevant skills and creativity (Becker, 2005)." (Liu et al., 2017, p. 1177).</p> <p>"We also inspected zero-order correlations to identify so-called impotent control variables; that is, variables that share variance with the predictor but not the criterion. We did so because inclusion of such impotent control variables can lead to an unnecessary reduction in statistical power (Becker, 2005; Carlson &amp; Wu, 2011), as well as to an increase in Type I errors (Becker, 2005; Spector &amp; Brannick, 2011)." (Venus et al., 2019, p. 673).</p>
2) Reporting results without control variables.	308 (22.9)	<p>"We therefore retained perceived interteam interdependence in hypotheses testing and excluded the other, non-significant covariates to avoid biased parameter estimates (Becker, 2005). Notably, results remained virtually unchanged when also excluding perceived inter-team interdependence or incorporating." (de Vries et al., 2014, p. 1344).</p> <p>"Importantly, the addition of these control variables did not qualitatively affect the results of our study—these variables did not exhibit a significant effect on whistleblowing behavior or impact the significance of the positive relationship between ostracism and whistleblowing. As such, and based on prior recommendations (e.g., Carlson and Wu 2012), we did not include these in our formal hypothesis test." (Spoelma et al., 2020, p. 349).</p>
3) Following the advice to use few control variables.	12 (0.9)	<p>"As recommended by Carlson and Wu (2012), we investigated our hypotheses while taking a conservative stance on control variables." (Clark &amp; Walsh, 2016, p. 190).</p> <p>"We were selective about which controls to use as research suggests that the inclusion of excessive controls not only reduces statistical power but may also yield biased estimates (Becker 2005)." (Sahai &amp; Frese, 2019, p. 933).</p>
4) Following other recommendations provided in the	479 (34.2)	<p>"To control for plausible alternative explanations, we controlled for several variables that are theoretically linked to the relationships of interest (Carlson &amp; Wu, 2012; Spector &amp; Brannick, 2011)." (Matta et al., 2015, p. 1693).</p>

Reference category	N (%)	Example excerpts
frugal perspective.		"Finally, to reduce concerns that spurious suppression could affect our results, given the number of control variables that we included in our analyses (Becker, 2005), we reran the analyses taking out one control variable at a time to examine the effects on the significance levels of the interactions." (McClellan et al., 2013, p. 540).
5) Critical, not applying the frugal recommendations.	1 (0.1)	"The only significant independent variable influencing intention to use deal sites is performance expectancy. Carlson and Wu (2012) suggest to exclude independent variables that are not significant. But removing the least significant independent variables one by one (like stepwise regression with backward elimination) may lead to increased significance of remaining variables [...]" (Sudzina, 2018, p. 68).
6) Theoretical, supporting a frugal perspective on control variable use.	52 (3.7)	"[...] if control variables are included [...] they may hamper the study by unnecessarily soaking up degrees of freedom or bias the findings related to the hypothesized variables (increasing either type I or type II error) (Becker, 2005). Thus, researchers should think carefully about the controls they include—being sure to include proper controls but excluding superfluous ones." (Bono & McNamara, 2011, p. 659).  "Note that one of the main sources for understanding best practices for control variable use and reporting has been Becker's (2005) article, primarily because it offers such detailed prescriptions for researchers." (Atinc et al., 2012, p. 70).
7) Theoretical, discussing the merits and drawbacks of the frugal perspective.	23 (1.6)	"In addition, identification, inclusion, and justification of control variables are critical for research using secondary data (Becker, 2005). Control variables may play important roles to rule out alternative explanations. Researchers also need to explain how they impact the relationship and why they should be included in the model (Carlson & Wu, 2011)." (Gnyawali & Song, 2016, p. 19).  "Although gender, social class, income, and occupation have been well researched by social scientists, they have often been relegated to the status of control variables in the organizational sciences, to questionable advantage (Becker et al. 2016)." (Johns, 2018, p. 35).
Total	1401	

*Note.* Excludes 188 articles that cited the frugal perspective in a context that was not related to control variable inclusion, had a citation error, or whose full text was not accessible to us.

## ***Literature Analysis 2: Impact of Both Perspectives in Management Research***

To compare the impact of both perspectives, we selected the seven journals with the most applications of the frugal perspective in the previous analysis: *Academy of Management Journal*, *Frontiers in Psychology*, *Journal of Applied Psychology*, *Journal of Management*, *Journal of Organizational Behavior*, *Leadership Quarterly* and *Personnel Psychology*. We further included *Strategic Management Journal* to get a better balance of micro- and macro-perspectives. We searched for the term “control variable” within the 2019-2021 period in these eight journals, producing a list of 1,157 articles. The first author read and coded the articles according to which perspective they applied. For example, an article was coded as applying the frugal perspective if it a) employed control variables, b) applied at least one of the recommendations the frugal perspective makes, and c) cited at least one of the methods papers we identified as belonging to the frugal perspective. The second author coded a subset of 30 articles (Krippendorff’s  $\alpha = 0.81$ ).

The coding results in Table 3 clearly show that if researchers justify their inclusion or exclusion of control variables using either perspective, the frugal perspective is more common by a wide margin. There is also a clear tendency that the frugal perspective is more common in micro-oriented journals (e.g., *Journal of Applied Psychology*) than in more macro-oriented journals (e.g., *Strategic Management Journal*). On the other hand, the few papers using the prolific perspective are exclusive to two journals that publish both micro and macro research (*Journal of Management* and *Leadership Quarterly*).

The frugal perspective has become the norm in methodological guidelines in management and particularly organizational behavior (Aguinis & Vandenberg, 2014; Becker et al., 2016; Carlson & Wu, 2012; O’Neill et al., 2014; Schjoedt & Bird, 2014) and the review results show it is increasingly followed in research practice. Yet, as explained earlier, in contrast to the prolific perspective, that builds on mathematical proofs presented in econometrics

textbooks, the methodological justification of the frugal perspective largely relies on intuitive arguments rather than proofs.

**Table 3**

*The Number of Articles Employing the Frugal / Prolific Perspective in Control Variables Use.*

Coding category	N (%)	Example excerpts	AMJ	FP	JAP	JOM	JOB	LQ	PP	SMJ
1) Applying the frugal perspective	73 (6.3)	"We checked whether we needed to control for these variables to take these possible relationships into account and avoid related potential bias in our results but retained them only if they had an impact to conserve statistical power (e.g., Becker, 2005)." (Den Hartog et al., 2020, p. 273).	4	15	22	7	17	1	6	1
2) Applying the prolific perspective	5 (0.4)	"A wide range of control variables was included in the analysis to improve the consistency of estimates (Antonakis et al., 2010)." (Bor, 2020, p. 5).	0	0	0	1	0	3	0	1
3) Not applying a specific perspective	1029 (88.9)	"In an effort to select the most relevant control variables, we consulted the literature to identify which factors affect the likelihood that activist hedge funds will target a firm."(DesJardine et al., 2021, p. 859).	113	387	163	146	76	48	46	50
4) Not using control variables	50 (4.3)	Several uses of the term "control variable" in the tables, but no empirical study (Calderwood & Mitropoulos, 2021, pp. 165–172).	2	19	2	10	4	12	1	0
Total	1157		119	421	187	164	97	64	53	52

*Note.* AMJ = Academy of Management Journal, FP = Frontiers in Psychology, JAP = Journal of Applied Psychology, JOM = Journal of Management, JOB = Journal of Organizational Behavior, LQ = Leadership Quarterly, PP = Personnel Psychology, SMJ = Strategic Management Journal.

## Monte Carlo Simulations

We present three Monte Carlo simulations. The first two test recommendations from the frugal perspective and contrast them with the prolific perspective on control variable inclusion: Dropping control variables that are not correlated with the dependent variable (“impotent control”, Simulation 1) and using proxied control variable in regressions (“proxy variable”, Simulation 2). The third simulation shows the effects of inclusion of an endogenous control variable (“bad controls”, Simulation 3). The simulations are designed to illustrate points made in the literature and the R and Stata code that we have uploaded to OSF<sup>7</sup> can be used for teaching and replication. The population models for each of the three simulations are shown in Figure 3 below.

We implemented the prolific strategy by always including the control variable(s) in all three simulations. In Simulations 1 and 3, we implemented the frugal perspective by including the control variable(s) only if it is (they are) significantly correlated with the dependent variable in a replication. In Simulation 2, we never included the proxied control variable in a frugal perspective regression. For simplicity, the coefficient  $\beta_1$  (the effect of the focal variable  $X$  on  $Y$ ) was set equal to 1 and all explanatory variables had variances of 1. Sample size was set to 250 in all three simulations, as we found that results were virtually identical across different sample sizes in a separate analysis. Other simulation specific design factors are reported below. We conducted 10,000 replications for each combination of factors.

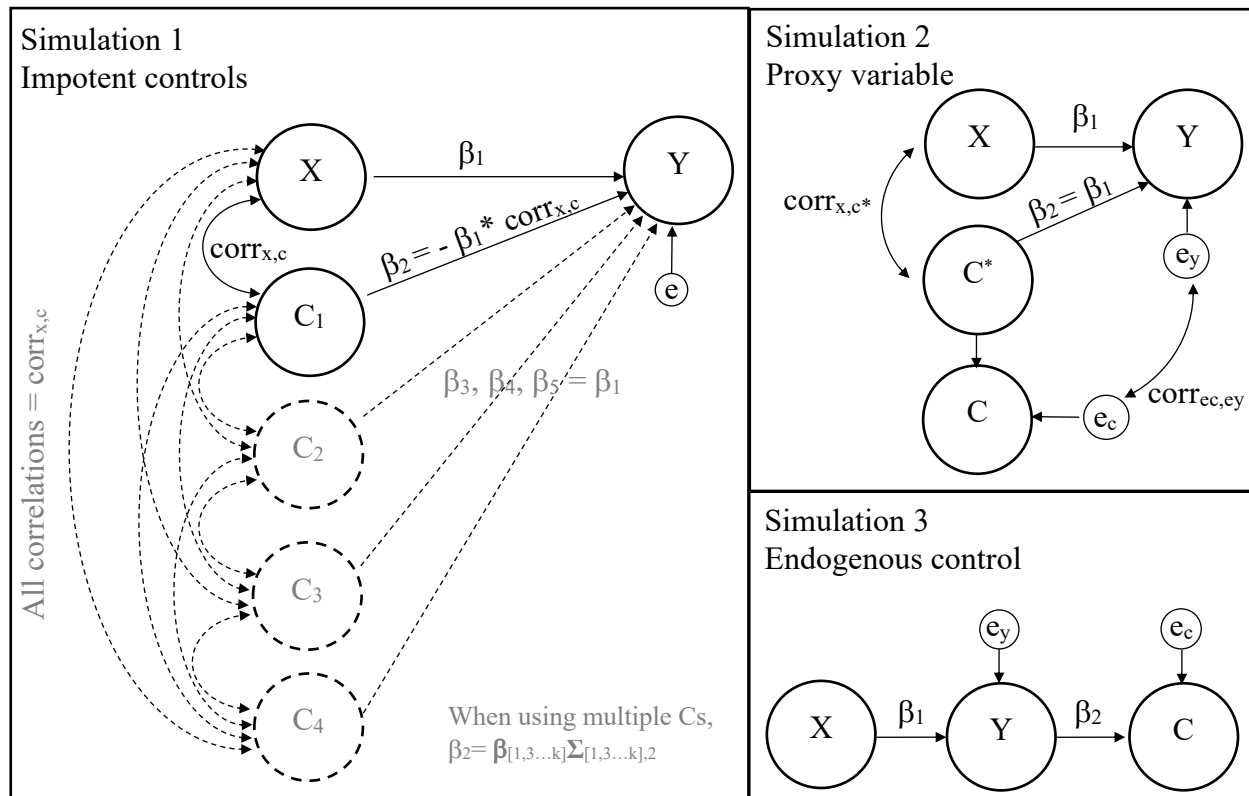
Simulation 1 tested the impotent control rule. The experimental conditions were the number of controls (1 to 4) and the correlation between explanatory variables ( $corr_{X,C}$ ) that varied from 0 to 0.8 in increments of 0.1. The first control  $C_1$  was uncorrelated with the dependent variable  $Y$  to produce an impotent control.

---

<sup>7</sup> [https://osf.io/wd3x7/?view\\_only=647b49a7557f4958a888755543bfae44](https://osf.io/wd3x7/?view_only=647b49a7557f4958a888755543bfae44)

**Figure 3**

*Depiction of the Population Models and Coefficients of the Monte Carlo Simulations. All Models are Linear and the Exogenous Variables are Standardized in the Population.*



*Note.*  $X$  = independent variable,  $Y$  = dependent variable,  $C_i$  = control variable(s) used in regressions,  $C^*$  = control of interest that is not measured directly but proxied,  $e_i$  = error term,  $corr_{k,j}$  = bivariate correlation,  $\beta_i$  = causal effect.

In the case where there is only one control variable, we did this by setting the effect ( $\beta_2$ ) of the control variable  $C_i$  on the dependent variable  $Y$  to be the negative product of its correlation with  $X$  ( $corr_{X,C}$ ) and  $\beta_1$ . In the other cases we used the corresponding matrix equation that also took the other controls into account. The error variance was scaled to produce an  $R^2 = 0.30$ , representing a substantial effect, yet one that might still be found in some organizational research.

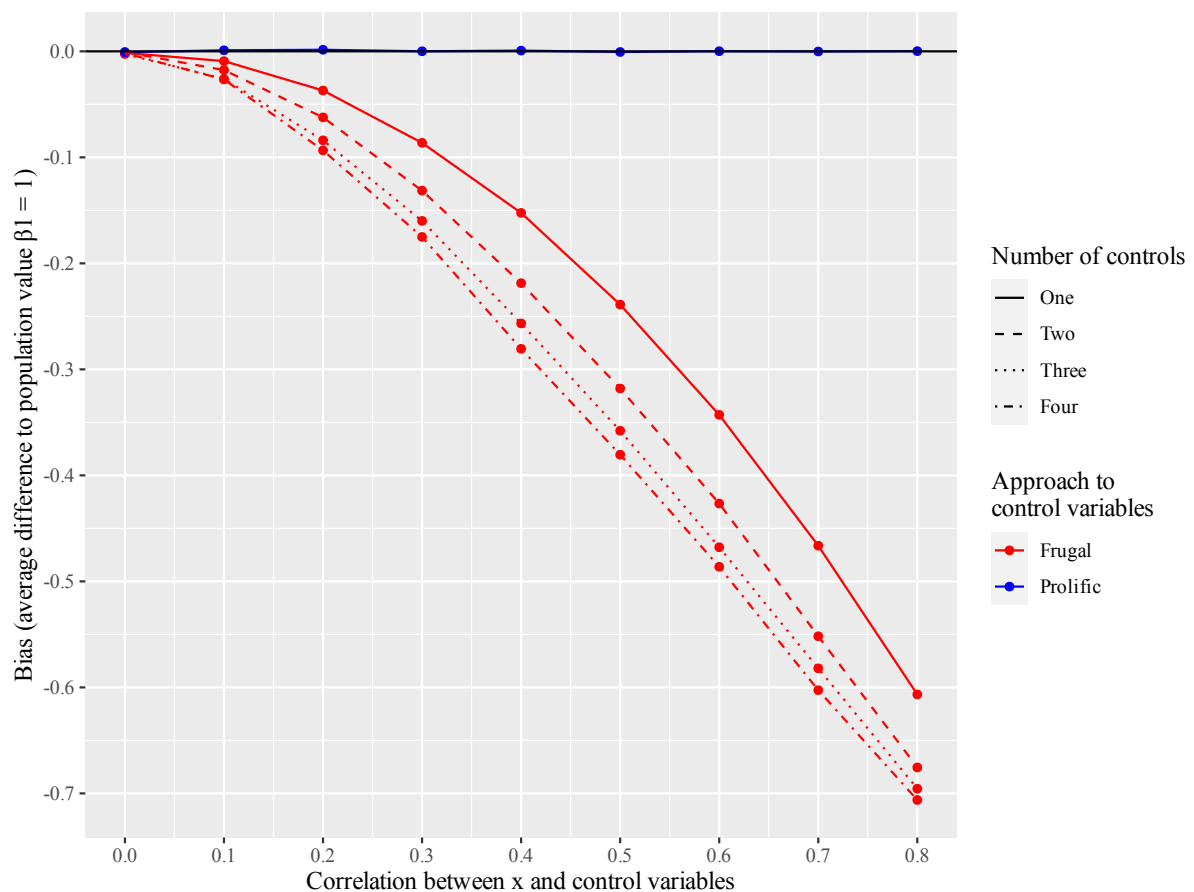
Figure 4 below shows that using the frugal perspective leads to bias in the estimate of



$\beta_1$ , and this bias increases with the increasing correlation between the control variables  $C_i$  and the independent variable  $X$  as well as the number of control variables. In contrast, when the prolific perspective is used, the estimate of  $\beta_1$  remains unbiased across all levels of correlation between  $C_i$  and  $X$ , independently of how many controls are simulated. The only case where dropping control variables does not produce bias is when the controls are uncorrelated with the focal predictor, which would be a case of irrelevant controls.

**Figure 4**

*Results From Simulation 1: Amount of Bias in Estimates of  $\beta_1$  When Applying the “Avoid Impotent Controls Rule”, With Varying Correlation Between the Independent Variable  $X$  and a Varying Amount of Controls  $C$ .*

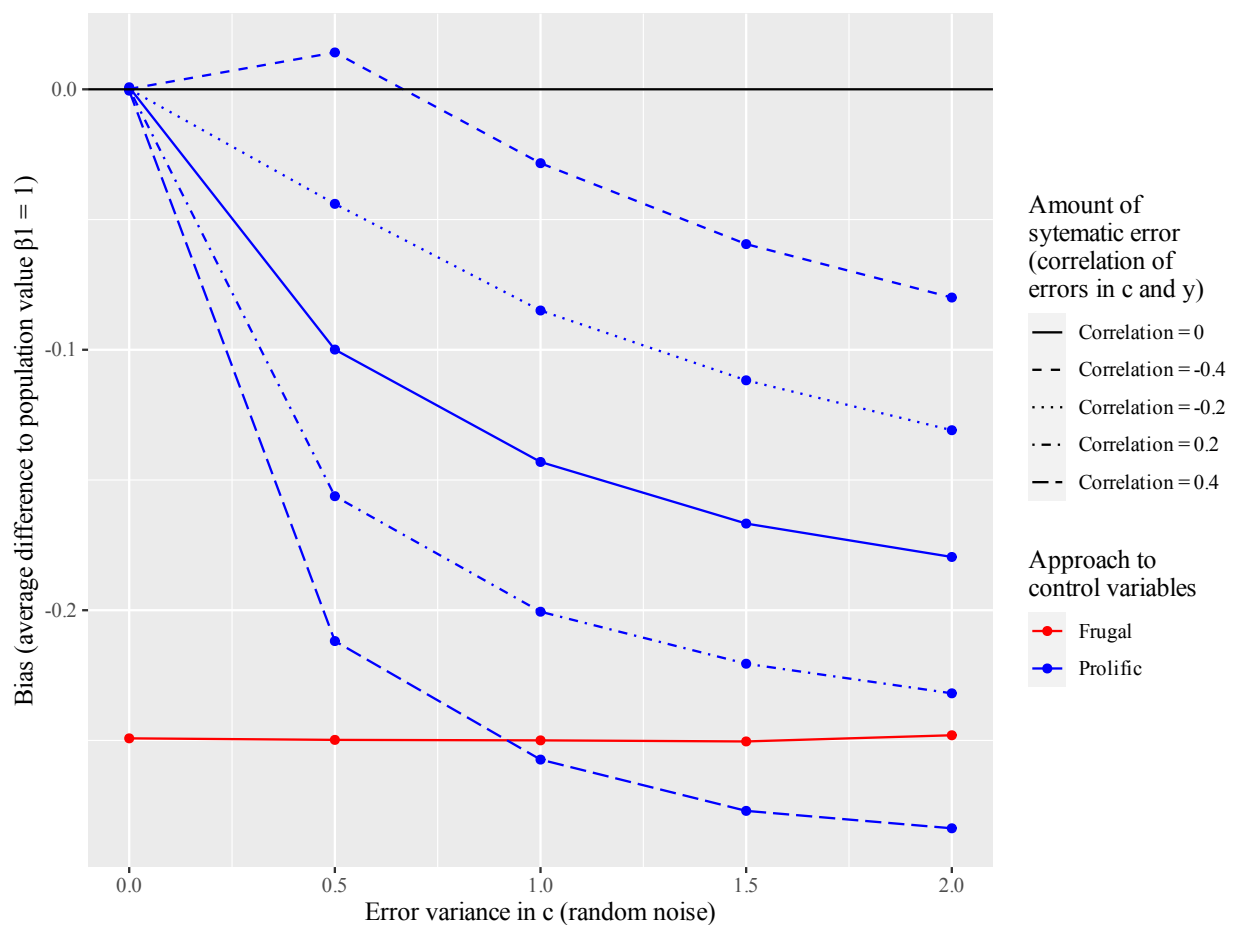


Note. Sample size = 250, SD of error in  $Y = 1$ .

In Simulation 2 we tested the avoid proxies rule. The design was identical to the single-control case in Simulation 1, except that the control variable  $C^*$  is proxied by  $C$ , which is measured with error. We varied the error variance in the proxy variable  $C$  ( $e_c$ ) from 0 to 2 in increments of 0.5 and the correlation between the error terms of the proxy variable  $C$  and the dependent variable  $Y$  ( $corr_{ec,ey}$ ) from  $-0.4$  to  $0.4$  in increments of  $0.2$ . This second experimental factor was added to model the effect of various degrees of endogeneity in the measurement error (i.e.  $corr_{ec,ey} \neq 0$ ). For simplicity, the correlation of the control variable  $C^*$  with the independent variable  $X$  ( $corr_{XC^*}$ ) is set to  $0.5$ .

**Figure 5**

*Results From Simulation 2: Amount of Bias in Estimates of  $\beta_1$  When Following the “Avoid Proxies” Rule, with Varying Error Variance of the Proxy and the Amount of Endogeneity.*

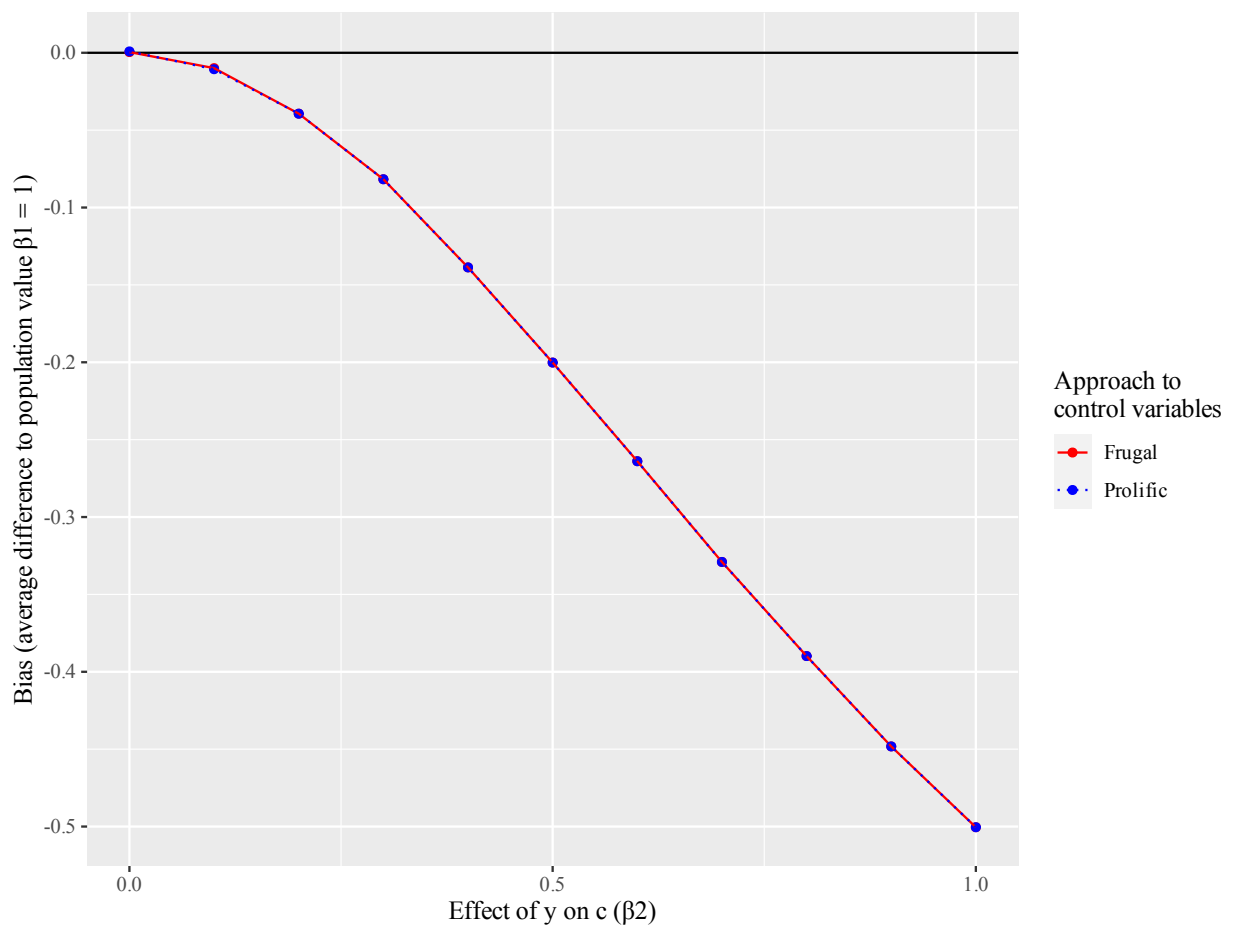


*Note.* Sample size = 250, SD of error in  $Y$  = 1, correlation  $X$  and  $C^*$  = 0.

The results from this simulation shown in Figure 5 above support the statement that random measurement error ( $corr_{ec,ey} = 0$ , no systematic error) in the predictor variables causes bias in regression estimates (Wooldridge, 2013, pp. 320–324), but this bias is always less than the bias from omitting the control variable. Further, in most cases of systematic error we tested, including the proxy in the model biases estimates less than dropping it.

**Figure 6**

*Results From Simulation 3: Amount of Bias in Estimates of  $\beta_1$  Following the Frugal or the Prolific Perspective, Varying  $\beta_2$ , the Effect of Y on C Which is the Amount of Endogeneity in C.*



*Note.* Results for both approaches are almost identical, which is why the lines overlap. Sample size = 250, SD of error in Y = 1.

In Simulation 3, the control variable  $C$  is not a real control but an outcome of the dependent variable  $Y$ , thus making it endogenous. We vary the degree of endogeneity by varying the effect of  $Y$  on  $C$  ( $\beta_2$ ) from 0 to 1 in increments of 0.1. Results in Figure 6 above demonstrate that the estimate of  $\beta_1$  is biased regardless of which perspective is applied and this bias depends on the magnitude of the causal effect of  $Y$  on  $C$  ( $\beta_2$ ). The simulations show that endogeneity (bad controls) is a serious concern, yet the specific recommendations that the two perspectives provide do not help in detecting endogenous controls.

### **Discussion**

Control variables have been discussed actively within management journals in recent years. While it is largely agreed that control variable reporting could be improved and that controls should be chosen based on relevant theory, there is some disagreement on whether many or few controls should be included. Specifically, we have the frugal perspective advocating the sparse use of controls and the prolific perspective advocating for a more generous use of control variables. We analyzed these two perspectives by comparing their recommendations against the more technical literature on regression analysis and econometrics, by conducting two reviews of their use in management research, and by running three simulations that demonstrate the effectiveness (or lack thereof) of specific empirical rules that have been proposed. Table 4 summarizes the central idea of both perspectives and their specific recommendations discussed in this paper. The table also presents an integrated perspective that we propose in this section.

**Table 4***Summary and Comparison of Both Perspectives and Proposition of Integrated Perspective.*

	Frugal perspective	Prolific perspective	Integrated perspective
Core idea	The use of control variables in management research is suboptimal. Researchers often include controls that are largely uncorrelated with the other study variables or controls that are potentially bad. Therefore, if a researcher is not sure about whether to control for a variable or not, it is better to leave it out.	Work on econometrics shows that inclusion of irrelevant variables does little harm. Because omitted variable bias is a severe concern, researchers should rather control for variables that they are not sure about. However, this applies only to controls that are exogenous.	Identify potentially relevant controls systematically and determine their relationships with other variables. Pay attention to what is the sources of variance in the potential controls. Leave out bad controls and include the rest.
Specific Recommendations	<p>Omit impotent control variables (<math> r_{cy}  &lt; .1</math>).</p> <p>Avoid proxies.</p> <p>Beware of bad controls.</p> <p>Run models with and without controls.</p> <p>Control variable justification based on theory.</p>	<p>Include many controls to prevent omitted variable bias and because inclusion of irrelevant controls has little negative consequences.</p> <p>Proxies should be used because an unreliable proxy is better than none.</p> <p>Overcontrolling by including endogenous controls should be avoided.</p> <p>Control variable justification based on theory.</p>	<p>Consider many potential controls based on prior theory and research.</p> <p>Consider what is the source of variation in each potential control.</p> <p>Rule out bad controls preferably with a causal graph.</p> <p>Document both included and excluded controls as an online supplement.</p> <p>Empirical rules for control variable selection should not be used.</p>

We hope that our manuscript encourages more rigorous control variable selection in three different ways. First, in our review we found that while the articles advocating the frugal perspective strongly argue that control variables should be chosen based on theory, it is the empirical rules from the frugal perspective that were applied in research practice. In our study, we used simulations to demonstrate that while these rules (“beware of impotent controls”,

“avoid proxies”) sound reasonable, they are at best useless and can often lead to incorrect results. While our analysis focused on regression, which is perhaps the most common analysis tool in organizational research, these same principles have been also derived in the context of structural causal models (e.g., Morgan & Winship, 2007; Pearl, 2012). For example, the result that controlling for a proxy generally reduces bias has been proven in the context of structural causal models as well (Ogburn & Vanderweele, 2013). Because these models make no assumptions about functional forms, these principles apply to also non-linear models (e.g., Poisson regression). More generally, they apply to any conditioning strategy, including for instance various matching techniques.

On a more general level, one can wonder why the empirical rules have been introduced in the first place. Both perspectives agree that control variables should be chosen based on theory, which is perhaps best exemplified by Breugh (2008), who states that “If theory suggests a variable should be controlled, it should be controlled” (p. 219). There is also a general agreement that the key limitation of the statistical control strategy is that it is impossible to control for every possible variable, but researchers should focus on the theoretically relevant ones (Antonakis et al., 2010, p. 1099; Cohen et al., 2003, sec. 12.1.4; Morgan & Winship, 2007, p. 5.4.2). But if controls should be determined based on theory, then empirical rules such as “avoid impotent controls” should play no role in control variable selection. Unfortunately, while the recent literature emphasizes the role of theory, it has failed to explain *how exactly researchers can use theory to guide control variable selection* beyond providing general recommendation on looking at variables that are related to both the dependent variable and the independent variable(s) so that all relevant alternative explanations can be ruled out (Spector, 2019). This might be one of the reasons why many management articles contain control variables that are just weakly correlated with the focal variables, creating an “illusion of statistical control” (Carlson & Wu, 2012). Our article clearly shows that the empirical rules should be

abandoned, and we hope that this would foster more thoughtful control variable selection.

Second, general recommendations such as “when in doubt, leave them out” or “err on the side of caution by including more than fewer control variables” cast the control variable decisions as a many or few choice, which is not ideal. The problem with the frugal perspective recommendation is that it might lead researchers to pick a couple of obvious controls and then declare that as sufficient instead of going through more rigorous control variable selection procedures. Similarly, following the prolific perspective one might just conclude that “The greater the number of variables that are controlled [...], the greater the likelihood that the relationship is not spurious.” (Singleton & Straits, 2018, p. 102) and mindlessly include a large number of controls, some of which are inevitably bad, leading to severe bias as the results from our Simulation 3 show. As Hünermund and coauthors (2022) put it, “the debate on whether to include fewer or more variables is not a productive one.” (p. 5).

Third, there is a better way of choosing control variables. The literature on econometrics (e.g., Greene, 2012; Wooldridge, 2010, 2013) and structural causal models (e.g., Cinelli et al., 2022; Huntington-Klein, 2022) tells us 1) it is important to include all relevant controls, 2) that bad or endogenous controls should not be included, and 3) including other controls is generally safe, but can increase or decrease the precision of estimates. The key challenge is how specifically to identify the relevant controls to be included and the bad controls to be excluded and doing this solely based on theory. The recent literature on control variable selection using causal graphs in sociology and psychology (Cinelli et al., 2022; Wysocki et al., 2022) provides an answer for how theory-based selection of controls might work. Hünermund et al. (2022) explain one possible workflow and demonstrate it in the context of leadership studies. Control variables selection should start by identifying a long list of potential controls based on prior theory, prior empirical results, and authors’ intuition. The list of controls should then be narrowed down by classifying them as relevant controls, bad controls, and unnecessary or

irrelevant controls by using causal graphs (Cinelli et al., 2022; Hünermund et al., 2022; Wysocki et al., 2022). It is also recommended that the control selection process should be documented (Hünermund et al., 2022). This is consistent with the frugal perspective’s call for more transparency on how and why control variables were selected as well as reporting control variable statistics and effects (e.g., Becker et al., 2016; Bernerth et al., 2018). While the full explanation of the causal graph is technical and beyond the scope of this work, we offer a simplified workflow that should be considered in Table 5 below.

**Table 5**

*Simplified Workflow for Control Variable Selection.*

Step 1	Start with a long list of control variable candidates based on prior theory, controls used in prior similar studies, and your own intuition.
Step 2	Consider the potential endogeneity of each control variable candidate by asking what is the source of variance in that variable (Guide & Ketokivi, 2015). Two common cases of endogenous controls are controlling for a mediator and controlling for a variable that depends on the dependent variable (as in our Simulation 3) (Cinelli et al., 2022).
Step 3	Leave out the endogenous controls and include the rest. Eliminating irrelevant controls or otherwise optimizing the control variable set (e.g., by identifying minimal adjustment sets Knüppel & Stang, 2010) can be done, but is beyond the scope of this guideline. However, variables with no relationship with the study variables can be excluded for parsimony (Scenario 1: Uncorrelated control in Table 1).
Step 4	Document the list of variables from Step 1 and how they were classified (included, omitted as bad, omitted as uncorrelated) and include it as an appendix or online supplement to an article.

The number of controls that comes out of this kind of rigorous process is the right number for a study — regardless of whether many or few. If a researcher is still not sure about which controls should be included there is a large literature on model uncertainty that can be consulted (Huntington-Klein, 2022, sec. 22.2).



## References

- Adams, M., & Hardwick, P. (1998). An Analysis of Corporate Donations: United Kingdom Evidence. *Journal of Management Studies*, 35(5), 641–654. <https://doi.org/10.1111/1467-6486.00113>
- Aerts, W., & Cormier, D. (2009). Media legitimacy and corporate environmental communication. *Accounting, Organizations and Society*, 34(1), 1–27. <https://doi.org/10.1016/j.aos.2008.02.005>
- Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1), 569–595. <https://doi.org/10.1146/annurev-orgpsych-031413-091231>
- Aigner, D. J. (1974). MSE dominance of least squares with errors-of-observation. *Journal of Econometrics*, 2(4), 365–372. [https://doi.org/10.1016/0304-4076\(74\)90020-7](https://doi.org/10.1016/0304-4076(74)90020-7)
- Al-Khazali, O. M., & Zoubi, T. A. (2005). Empirical testing of different alternative proxy measures for firm size. *Journal of Applied Business Research (JABR)*, 21(3), Article 3. <https://doi.org/10.19030/jabr.v21i3.1471>
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6), 1086–1120. <https://doi.org/10.1016/j.leaqua.2010.10.010>
- Atinc, G., Simmering, M. J., & Kroll, M. J. (2012). Control variable use and reporting in macro and micro management research. *Organizational Research Methods*, 15(1), 57–74. <https://doi.org/10.1177/1094428110397773>
- Basu, D. (2020). Bias of OLS estimators due to exclusion of relevant variables and inclusion of irrelevant variables. *Oxford Bulletin of Economics and Statistics*, 82(1), 209–234. <https://doi.org/10.1111/obes.12322>

- Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, 8(3), 274–289. <https://doi.org/10.1177/1094428105278021>
- Becker, T. E., Atinc, G., Breugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *Journal of Organizational Behavior*, 37(2), 157–167. <https://doi.org/10.1002/job.2053>
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6), 1315–1344. <https://doi.org/10.1093/rof/rfac033>
- Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1), 229–283. <https://doi.org/10.1111/peps.12103>
- Bernerth, J. B., Cole, M. S., Taylor, E. C., & Walker, H. J. (2018). Control variables in leadership research: A qualitative and quantitative review. *Journal of Management*, 44, 131–160. <https://doi.org/10.1177/0149206317690586>
- Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. Sage Publications.
- Bono, J. E., & McNamara, G. (2011). Publishing in AMJ — part 2: Research design. *Academy of Management Journal*, 54(4), 657–660. <https://doi.org/10.5465/amj.2011.64869103>
- Bor, A. (2020). Evolutionary leadership theory and economic voting: Warmth and competence impressions mediate the effect of economic perceptions on vote. *The Leadership Quarterly*, 31(2), 101295. <https://doi.org/10.1016/j.leaqua.2019.05.002>
- Breugh, J. A. (2008). Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. *Human Resource Management Review*, 18(4), 282–293. <https://doi.org/10.1016/j.hrmr.2008.03.001>
- Calderwood, C., & Mitropoulos, T. (2021). Commuting spillover: A systematic review and

- agenda for research. *Journal of Organizational Behavior*, 42(2), 162–187.  
<https://doi.org/10.1002/job.2462>
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511811241>
- Carlson, K. D., & Wu, J. (2012). The Illusion of Statistical Control: Control Variable Practice in Management Research. *Organizational Research Methods*, 15(3), 413–435.  
<https://doi.org/10.1177/1094428111428817>
- Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*. <https://doi.org/10.1177/00491241221099552>
- Clark, O. L., & Walsh, B. M. (2016). Civility climate mitigates deviant reactions to organizational constraints. *Journal of Managerial Psychology*, 31(1), 186–201.  
<https://doi.org/10.1108/JMP-01-2014-0021>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
- Dai, L., Parwada, J. T., & Zhang, B. (2015). The Governance Effect of the Media’s News Dissemination Role: Evidence from Insider Trading: GOVERNANCE EFFECT OF THE MEDIA’S NEWS DISSEMINATION ROLE. *Journal of Accounting Research*, 53(2), 331–366. <https://doi.org/10.1111/1475-679X.12073>
- de Vries, T. A., Walter, F., Van der Vegt, G. S., & Essens, P. J. M. D. (2014). Antecedents of individuals’ interteam coordination: Broad functional experiences as a mixed blessing. *Academy of Management Journal*, 57(5), 1334–1359.  
<https://doi.org/10.5465/amj.2012.0360>
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.  
<https://doi.org/10.1016/j.socscimed.2017.12.005>

- Den Hartog, D. N., De Hoogh, A. H. B., & Belschak, F. D. (2020). Toot your own horn? Leader narcissism and the effectiveness of employee self-promotion. *Journal of Management*, 46(2), 261–286. <https://doi.org/10.1177/0149206318785240>
- DesJardine, M. R., Marti, E., & Durand, R. (2021). Why activist hedge funds target socially responsible firms: The reaction costs of signaling corporate social responsibility. *Academy of Management Journal*, 64(3), 851–872. <https://doi.org/10.5465/amj.2019.0238>
- Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2), 127–136. <https://doi.org/10.1198/000313005X41337>
- Gnyawali, D. R., & Song, Y. (2016). Pursuit of rigor in research: Illustration from cooperation literature. *Industrial Marketing Management*, 57, 12–22. <https://doi.org/10.1016/j.indmarman.2016.05.004>
- Green, J. P., Tonidandel, S., & Cortina, J. M. (2016). Getting through the gate: Statistical and methodological issues raised in the reviewing process. *Organizational Research Methods*, 19(3), 402–432. <https://doi.org/10.1177/1094428116631417>
- Greene, W. H. (2012). *Econometric analysis* (7<sup>th</sup> ed). Prentice Hall.
- Guide, D., & Ketokivi, M. (2015). Notes from the editors: Redefining some methodological criteria for the journal. *Journal of Operations Management*, 37, v–viii. [https://doi.org/10.1016/S0272-6963\(15\)00056-X](https://doi.org/10.1016/S0272-6963(15)00056-X)
- Heckman, J. J. (2008). Econometric causality. *International Statistical Review*, 76(1), 1–27. <https://doi.org/10.1111/j.1751-5823.2007.00024.x>
- Hernández, A. V., Steyerberg, E. W., & Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, 57(5), 454–460. <https://doi.org/10.1016/j.jclinepi.2003.09.014>

- Hitchcock, C. (2010). Causation. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (1. publ. in paperback, pp. 317–326). Routledge.
- Hünemann, P., Louw, B., & Rönkkö, M. (2022). The choice of control variables: How causal graphs can inform the decision. *Academy of Management Proceedings*, 2022(1), 15534. <https://doi.org/10.5465/AMBPP.2022.294>
- Huntington-Klein, N. (2022). *The effect: An introduction to research design and causality*. <https://theeffectbook.net/index.html>
- Jaccard, J., & Jacoby, J. (2020). *Theory construction and model-building skills: A practical guide for social scientists* (Second edition). The Guilford Press.
- Jensen, P. H., & Webster, E. (2009). Another look at the relationship between innovation proxies. *Australian Economic Papers*, 48(3), 252–269. <https://doi.org/10.1111/j.1467-8454.2009.00374.x>
- Johns, G. (2018). Advances in the treatment of context in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, 5(1), 21–46. <https://doi.org/10.1146/annurev-orgpsych-032117-104406>
- Knüppel, S., & Stang, A. (2010). DAG program: Identifying minimal sufficient adjustment sets. *Epidemiology*, 21(1), 159. <https://doi.org/10.1097/EDE.0b013e3181c307ce>
- Lewis, J. W., & Escobar, L. A. (1986). Suppression and enhancement in bivariate regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35(1), 17–26. <https://doi.org/10.2307/2988294>
- Li, M. (2021). Uses and abuses of statistical control variables: Ruling out or creating alternative explanations? *Journal of Business Research*, 126, 472–488. <https://doi.org/10.1016/j.jbusres.2020.12.037>
- Liu, D., Gong, Y., Zhou, J., & Huang, J.-C. (2017). Human resource systems, employee creativity, and firm innovation: The moderating role of firm ownership. *Academy of*

- Management Journal, 60(3), 1164–1188. <https://doi.org/10.5465/amj.2015.0230>
- Lund, T. (1981). Meehl and the ex post facto design. *Scandinavian Journal of Psychology*, 22(1), 93–96. <https://doi.org/10.1111/j.1467-9450.1981.tb00382.x>
- Mändli, F. B., Amer, E., & Bonardi, J.-P. (2023). Private politics, dynamics of negative media exposure, and firm strategy. [Working Paper].
- Matta, F. K., Scott, B. A., Koopman, J., & Conlon, D. E. (2015). Does seeing “eye to eye” affect work engagement and organizational citizenship behavior? A role theory perspective on LMX agreement. *Academy of Management Journal*, 58(6), 1686–1708. <https://doi.org/10.5465/amj.2014.0106>
- McClellan, E. J., Burris, E. R., & Detert, J. R. (2013). When does voice lead to exit? It depends on leadership. *Academy of Management Journal*, 56(2), 525–548. <https://doi.org/10.5465/amj.2011.0041>
- Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Analyses of theories and methods of physics and psychology* (p. 373–402). University of Minnesota Press.
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.
- Nielsen, B. B., & Raswant, A. (2018). The selection, use, and reporting of control variables in international business research: A review and recommendations. *Journal of World Business*, 53(6), 958–968. <https://doi.org/10.1016/j.jwb.2018.05.003>
- Ogburn, E. L., & Vanderweele, T. J. (2013). Bias attenuation results for nondifferentially mis-measured ordinal and coarsened confounders. *Biometrika*, 100(1), 241–248. <https://doi.org/10.1093/biomet/ass054>
- O’Neill, T. A., McLarnon, M. J. W., Schneider, T. J., & Gardner, R. C. (2014). Current misuses of multiple regression for investigating bivariate hypotheses: An example from the

- organizational domain. *Behavior Research Methods*, 46(3), 798–807.  
<https://doi.org/10.3758/s13428-013-0407-1>
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). The Guilford Press.
- Russo, M. V., & Fouts, P. A. (1997). A resource-based perspective on corporate environmental performance and profitability. *Academy of Management Journal*, 40(3), 534–559.  
<https://doi.org/10.2307/257052>
- Sahai, R., & Frese, M. (2019). If you have a hammer, you only look for nails: The relationship between the einstellung effect and business opportunity identification. *Journal of Small Business Management*, 57(3), 927–942. <https://doi.org/10.1111/jsbm.12346>
- Schjoedt, L., & Bird, B. (2014). Control variables: Use, misuse and recommended use. In A. Carsrud & M. Brännback (Eds.), *Handbook of research methods and applications in entrepreneurship and small business* (pp. 136–155). Edward Elgar Publishing.  
<https://doi.org/10.4337/9780857935052.00013>
- Schoot, R. van de, & Miočević, M. (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge. <https://doi.org/10.4324/9780429273872>
- Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (2017). *Understanding regression analysis: An introductory guide*. SAGE Publications, Inc.  
<https://doi.org/10.4135/9781506361628>
- Semenova, N., & Hassel, L. G. (2015). On the Validity of Environmental Performance Metrics. *Journal of Business Ethics*, 132(2), 249–258. <https://doi.org/10.1007/s10551-014-2323-4>
- Singleton, R., & Straits, B. C. (2018). *Approaches to social research* (Sixth edition). Oxford University Press.
- Smith, R. L., Ager, J. W., & Williams, D. L. (1992). Suppressor variables in multiple

- regression/correlation. *Educational and Psychological Measurement*, 52(1), 17–29.  
<https://doi.org/10.1177/001316449205200102>
- Spector, P. E. (2019). Do not cross me: Optimizing the use of cross-sectional designs. *Journal of Business and Psychology*. <https://doi.org/10.1007/s10869-018-09613-8>
- Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2), 287–305.  
<https://doi.org/10.1177/1094428110369842>
- Spoelma, T. M., Chawla, N., & Ellis, A. P. J. (2020). If you can't join 'em, report 'em: A model of ostracism and whistleblowing in teams. *Journal of Business Ethics*.  
<https://doi.org/10.1007/s10551-020-04563-9>
- Sturman, M. C., Sturman, A. J., & Sturman, C. J. (2022). Uncontrolled control variables: The extent that a researcher's degrees of freedom with control variables increases various types of statistical errors. *Journal of Applied Psychology*, 107(1), 9–22.  
<https://doi.org/10.1037/apl0000849>
- Sudzina, F. (2018). Impact of UTAUT/UTAUT2 motives on intention to use deal sites. In M. H. Bilgin, H. Danis, E. Demir, & U. Can (Eds.), *Consumer behavior, organizational strategy and financial economics* (pp. 63–71). Springer International Publishing.  
[https://doi.org/10.1007/978-3-319-76288-3\\_5](https://doi.org/10.1007/978-3-319-76288-3_5)
- Venus, M., Stam, D., & van Knippenberg, D. (2019). Visions of change as visions of continuity. *Academy of Management Journal*, 62(3), 667–690.  
<https://doi.org/10.5465/amj.2015.1196>
- Waddock, S. A., & Graves, S. B. (1997). The corporate social performance-financial performance link. *Strategic Management Journal*, 18(4), 303–319.  
[https://doi.org/10.1002/\(SICI\)1097-0266\(199704\)18:4<303::AID-SMJ869>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-0266(199704)18:4<303::AID-SMJ869>3.0.CO;2-G)



Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2<sup>nd</sup> ed.). MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5<sup>th</sup> ed). South-Western Cengage Learning.

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, 5(2), 1–19.

<https://doi.org/10.1177/25152459221095823>

Zax, J. (2011). *Introductory econometrics: Intuition, proof, and practice*. Stanford University Press. <https://doi.org/10.11126/stanford/9780804772624.001.0001>

## **Automated Sentiment Analysis vs. Manual Coding:**

### **Examining the Tone of News on Companies' Environmental Issues.**

Fabian Mändli, HEC Lausanne

#### **Abstract**

Automated approaches for sentiment analysis are gaining increasing popularity in management research, as they allow to save time and resources. While some have praised these techniques as reaching human level accuracy, they have also been criticized for their poor performance and lack of agreement between each other. In this paper, I examine the performance of dictionary and supervised machine learning approaches for sentiment analysis used most frequently in management research by applying them to a large hand-coded quality dataset of news articles on companies' environmental issues. I find that dictionary approaches yield very different results than the hand-coding baseline, mostly because they are unable to account for the complexity of news articles. Although machine learning approaches performed better than dictionary approaches at the task, their results still differ substantially from hand-coding. I also show that novel approaches using large language models are able to outperform both types of approaches at the task. I discuss that although there are research settings where commonly used automated sentiment analysis might provide satisfactory results, these results might in many cases not be of sufficient quality. I recommend researchers to include state of the art machine learning models in the stage of model decision instead of turning to commonly used approaches. I conclude that manual coding of a subset of the data remains vital for training and validation of all types of automated approaches.

**Keywords:** *sentiment analysis, dictionary, machine learning*

## Introduction

Qualitative and quantitative research in management is increasingly employing automated procedures to analyze text. This is driven by two forces: First, there has been growing availability of digitized information on companies from various sources such as websites and databases, driven by global digitalization of society and business (Ritter & Pedersen, 2020). Analyzing these large amounts of information is laborious if done manually. At the same time, new information on investigated topics emerges, often at an increasing rate (Zwitter, 2014), potentially resulting in a Sisyphus-type of task. Second, the development of automated tools for computer-aided text analysis permits to treat large amounts of information efficiently and at low cost, and allows to potentially replicate studies (Duriau et al., 2007). Consequently, automated text analysis approaches have found their way from information science to disciplines such as political science, communication studies and management science (Piryani et al., 2017).

Sentiment analysis is one typical task executed with automated approaches<sup>1</sup>. Sentiment itself is defined as “[...] emotional dispositions formed toward an object [...] that] are enduring.” (Munezero et al., 2014, p. 7). For example, sentiment can be the tone of coverage that “refers to the level of support for an organization expressed in a news article” (Carroll & Deephouse, 2014, p. 84). Building on this, sentiment analysis is described as “[...] a task of natural language processing that aims to extract sentiments and opinions from texts” (Birjali et al., 2021, p. 1). Sentiment analysis can be applied to written texts of various nature, for example social media posts (Drus & Khalid, 2019), product reviews (Micu et al., 2017), or news articles (Young & Soroka, 2012), to determine the sentiment in a given piece of text. In the context of management and information systems research, sentiment analysis can be broadly applied in two ways: First, to understand how different groups of stakeholders evaluate industries or companies and/or

---

<sup>1</sup> Sentiment analysis is sometimes referred to as opinion mining. For simplicity, in the remainder of the paper I use the term sentiment analysis.

their management figures, strategies or products and services. For example, sentiment analysis can be used to proxy organizational legitimacy by assessing social media content (Etter et al., 2018), to grasp consumer sentiment by examining online forums, product reviews and social media content (Rambocas & Pacheco, 2018), to understand short-term stock market movements by analysing Twitter<sup>2</sup> discussions (Rao & Srivastava, 2012), or to assess CEO communication styles by examining interviews (Choudhury et al., 2019). Sentiment analysis has also been used to determine media sentiment on companies related to specific topics such as for example environmental issues (Bettinazzi et al., 2023). Second, sentiment analysis can also be employed within companies, such as to examine processes, projects, employees or teams (Gelbard et al., 2018). It has been used for example to understand the relationship between supply chain performance and supply chain members' opinions (Swain & Cao, 2019), to grasp challenges in project management (Guzman & Bruegge, 2013), to enhance the capability of employee appraisal systems (Aqel & Vadera, 2010), or to determine employee satisfaction (Moniz & De Jong, 2014).

In management studies, a particular application of sentiment analysis is the operationalization of social evaluation constructs, often times using news content. For example, to construct a measure of organizational celebrity, Hubbard et al. (2018) calculated positive and negative affect of news reporting using sentiment analysis. Similarly, stigma intensity has been measured using the amount of negative news articles relative to the total number of articles (Piazza & Perretti, 2015). Also, sentiment analysis of social media content (Etter et al., 2018) and/or media reports on organizations can be operationalized to proxy (media) reputation and/or legitimacy (Deephouse & Carter, 2005; Pfarrer et al., 2010; Wei et al., 2017). Even though these constructs are built using the same underlying data, news articles and/or social media

---

<sup>2</sup> Twitter has been recently renamed to "X". As the social media platform is still commonly known as Twitter at the time of writing and the research I am referring to is referring to the platform as Twitter, I kept the old name for consistency.

posts, they differ in terms of selection strategy for the news articles examined as well as how the proxy measures are calculated (Bitektine et al., 2020; Pollock et al., 2019). The commonality is that they require determining whether a piece of information is endorsing or praising, thus conveying positive sentiment, or critical and disapproving, thus conveying negative sentiment. This is where tools for automated sentiment analysis have been applied in recent years.

Despite the increasing popularity, automated procedures to analyze sentiment and opinion received criticism to produce results that are limited in their reliability or sometimes even unsuitable (Barberá et al., 2021; Boukes et al., 2020). This mainly has to do with the complexity of language and the different meanings of words depending on the context, factors that are difficult to take into account using automated approaches. Indeed, “[...] sentiment categorizations [are] more difficult than topic classification.” (Pang et al., 2002, p. 83). These approaches are usually employed instead of manual coding, thus manual coding commonly represents the benchmark for comparison of performance (e.g., Boukes et al., 2020; van Atteveldt et al., 2021). While it appears reasonable to assume that trained human coders are superior at determining tonality of texts, as they are able to take into account more complexity than algorithms, there is evidence that human coding is subject to other limitations, such as fatigue, text features and individual coder characteristics (Weber et al., 2018). Given the rising popularity of automated approaches for sentiment analysis in management science, there is a necessity to take stock of these approaches and compare their performance. This paper examines the following research questions: How do automated approaches commonly used in management science perform compared to manual coding in a context of news articles where there is an object (company) as well as a subject (environmental issues)? How does finetuning of the approaches and the textual data affect performance of the approaches examined? Do potential differences in coding between the approaches manifest themselves in different

conclusions drawn from an analysis based on the coding data? Finally, can manual coding (still) be considered as the baseline given the recent advances of large language models?

To answer these questions, I conducted a literature search in 12 leading management journals, to map the types of approaches that are used for automated sentiment analysis. In a second step, I analysed the popularity of the individual approaches for sentiment across research disciplines. Based on this analysis, I chose the most popular approaches and briefly introduce them. In a next step, I examined them on the manually pre-coded dataset of environmental company news from Mändli et al. (2023).

The reason why I chose manual coding as a benchmark is twofold: 1) Even if most of the automated approaches have been validated against manual coding for other kinds of data, predominantly shorter texts such as social media posts, it is imperative to evaluate them also in the context I examined. For example, an approach that performs well with Twitter posts could perform very differently with news articles. Being able to examine those approaches against a manually coded dataset of news articles thus presents a unique opportunity. 2) The manual coding of the tone of news articles in the dataset used was conducted meticulously, based on an extensive codebook and intensive training of the coders. Consequently, its quality should be particularly adequate to be used as a benchmark.

Based on the initial results I obtain by the approaches I selected, I explore how their performance can be enhanced by finetuning as well as combining them. I also investigate how two large language models perform at the same task of sentiment classification. Using the machine-coded datasets, I then replicate a simplified analysis from the original paper (Mändli et al., 2023) and compare this to the results based on manually coded data. This allows to understand the potential impact of using automated tools on conclusions drawn from data. To my knowledge, this paper is the first in management science to provide a comprehensive

overview of the most frequently used techniques in the field, testing them on a large news dataset and comparing their performance against manual coding.

My analysis yields that machine learning approaches outperform dictionary approaches applied to the examined dataset. Nevertheless, the performance of machine learning approaches remains below quality levels usually required for manual coding. The application of several refinements recommended in the literature only marginally improve results. However, results of two state-of-the-art large language models support that these models present a third novel category of algorithms for sentiment classification that might be approaching human coding accuracy. Finally, the abridged data analysis that compares results obtained based on the hand-coded data with those from automated coding, shows that conclusions could vary by coding approach chosen.

The paper contributes to understand the implications of using automated approaches for sentiment analysis in management science. It shines a critical light on the approaches commonly employed in the field: Even with careful examination and validation using manual coding, many of the approaches used are likely to yield low performance, particularly the dictionary approaches appear to struggle with news articles. It further shows that conclusions drawn based on automated approaches may vary substantially by approach chosen. This shows the importance of contrasting different approaches and comparing results, also in applied studies. My results mirror recent discussions on the accuracy of these approaches in political science and communication studies (Barberá et al., 2021; Boukes et al., 2020), where particularly the dictionary-based approaches are criticized. Yet, in some cases commonly used automated approaches might yield sufficient quality for certain kinds of data analysis, for example when aggregated coarse measures of media tenor are required (e.g., Hubbard et al., 2018). My analysis also provides support for the use of large language models that do not require large amounts of training data to perform well for tasks such as sentiment analysis. Researchers

exploring these new tools might be able to reach new levels of accuracy and be able to save time and resources on training and validation.

### **Theory**

Procedures of automated sentiment analysis can be broadly separated into two categories: Dictionary approaches and supervised machine learning<sup>3</sup> approaches (Barberá et al., 2021). The prior rely on a pre-constructed dictionary that matches expressions with valence (e.g., negative or positive), thus there is no need for manual coding or training. The latter are machine learning algorithms that can be used for classification tasks. In machine learning, sentiment analysis represents a classification task, where a text is classified according to a specific set of features or attributes using an algorithm. They usually demand a pre-coded training set to “learn” the classification rules. Thus, they are usually also more resource-intensive than dictionary approaches.

Sentiment analysis of written texts from various origins has been applied in numerous fields within management science. For example to assess sentiment in news media in the context of corporate reputation (e.g., Fombrun & Shanley, 1990) and social media sentiment analysis to measure organizational legitimacy (Etter et al., 2018), to predict stock returns employing sentiment analysis of news (e.g., Tetlock, 2007; Uhl, 2014), or to determine CEO communication styles by using sentiment analysis of interview transcripts (Choudhury et al., 2019). Also, in marketing research, sentiment analysis has gained popularity to analyse consumer opinions online (Rambocas & Pacheco, 2018). The increasing popularity of sentiment analysis is reflected in the number of publications making use of this tool: A Web of Science search using the keywords “sentiment analysis” in the research domains of business and management confirms this trend (see Figure 1 below).

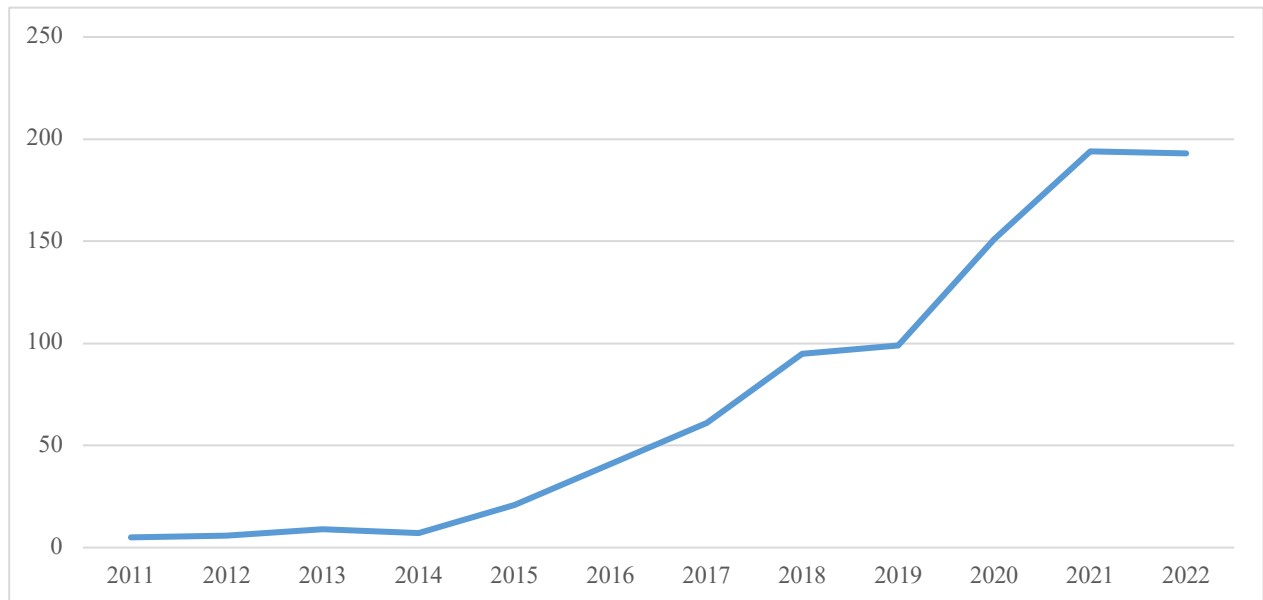
---

<sup>3</sup> For simplicity, I refer to supervised machine learning approaches only as machine learning approaches in the remainder of the paper.



**Figure 1**

*Articles in Web of Science in Research Field “Business and Management” that Contain the Term “sentiment analysis”(2011 – 2021).*



Albeit its increasing popularity, automated sentiment analysis has been criticized due to its low performance and poor overlap of results across different techniques (Barberá et al., 2021; Boukes et al., 2020). Dictionary approaches specifically have further been scrutinized because they have seldomly been developed for the research questions they are being used for (Loughran & McDonald, 2015). This critique is in sharp contrast to the works of researchers that have proposed these techniques, which often times claim that their approaches are not only superior to other existing automated approaches, but even able to reach human accuracy (e.g., Hutto & Gilbert, 2014; Pang et al., 2002).

Next, I conduct a literature analysis to identify the approaches and specific tools used for sentiment analysis in recent years in management research. Thereafter, I determine the overall popularity of each tool for the task of sentiment analysis in research more broadly and select the approaches I examine in the remainder of the paper.

## Literature Analysis

To determine the popularity of automated sentiment analysis in management science, and to take stock of the kinds of tools that are being used, I conducted a literature analysis in a set of highly ranked management journals. The 12 journals examined are: *Academy of Management Annual Meeting Proceedings*<sup>4</sup>, *Academy of Management Journal*, *Academy of Management Review*, *Accounting, Organizations and Society*, *Entrepreneurship Theory and Practice*, *Journal of Business & Economic Statistics*, *Journal of International Business Studies*, *Journal of Management Studies*, *Organization Science*, *Organizational Research Methods*, *Strategic Management Journal*, and *The Leadership Quarterly*.

In each of these journals, I conducted a full-text search of the 2011 – 2021 period with the keywords “sentiment analysis”. This yielded a list of 43 articles, of which I was able to obtain the full text of 30 articles. I coded the full articles in more detail, recorded what class of approach was used (dictionary and/or machine learning) as well as what specific kind (algorithm or dictionary type). Of the articles examined, 16 used an automated approach and provided sufficient information on the procedure, the remainder were editorials or reviews, only referring to the technique. An overview of the number of articles using machine learning and/or dictionary approaches by journal can be found in Table 1 below. Although the numbers are limited, they clearly show that dictionary approaches are employed more frequently.

Table 2 below presents an overview of the respective dictionaries and/or machine learning algorithms the 16 identified papers applied. It shows that there is no clear dominance in popularity of a single or even a group of approaches. Further, it shows that some also used a combination of different algorithms or dictionaries.

---

<sup>4</sup> I included Academy of Management Annual Meeting Proceedings in the list because this allows to include more recent developments in the field of management that have not necessarily reached publication yet. However, it is not a peer reviewed journal.

**Table 1**

*Number of Articles in Highly Ranked Management Journals Employing Different Types of Automated Sentiment Analysis (2011 – 2021).*

Journal Name	Approach Type	Number of Articles
Academy of Management Annual Meeting Proceedings	Dictionary	4
	Machine Learning	2
Entrepreneurship Theory and Practice	Dictionary	2
Journal of Management Studies	Dictionary	1
	Dictionary & Machine Learning	1
Organization Science	Dictionary	2
Organizational Research Methods	Dictionary	2
Strategic Management Journal	Dictionary	2
Total		16

*Note.* Academy of Management Journal, Academy of Management Review, Accounting, Organization and Society, Journal of Business & Economic Statistics, Journal of International Business Studies, and The Leadership Quarterly did not return any results in the search query.

Using the dictionary techniques identified in the literature analysis above as well as by extending the list with commonly used dictionaries in communication studies and political science (Ribeiro et al., 2016; Zhang et al., 2014), I conducted a second examination to determine the broader popularity of a given approach. I carried out a full-text search for each approach in Web of Science, where I searched for the word “sentiment” in combination with the name of each approach (i.e., “sentiment” and “VADER”) without any other restrictions. An overview of the results from this analysis can be found in Table 3 further below. VADER, SentiStrength,

LIWC, TextBlob and AFINN are the five most frequently applied approaches in the works covered by the Web of Science database. I chose these five approaches for further examination. They have also been found to be among the best-performing in examinations based on different datasets (Ribeiro et al., 2016).

**Table 2**

*Types of Machine Learning Algorithms or Dictionaries Used for Sentiment Analysis in the Articles of the Highly Ranked Management Journals in Table 1 (2011 – 2021).*

Approach Type	Name	Number of Times Used
Machine Learning	IBM Watson Tone Analyzer	2
	Multinomial Naïve Bayes	1
	Bernoulli Naïve Bayes	1
	Stochastic Gradient Decent	1
	Support Vector Machine	1
	Linear Support Vector Machine	1
	Logistic Regression	1
Dictionary	Own list of Keywords	2
	AFINN	2
	HuLiu	2
	LIWC	2
	SentiStrenght	1
	DICTION	1
	VADER	1
	AnalyzeSentiment	1
	Rinker	1
	Bing	1
	Syuzhet R	1
	MPQA	1
	General Inquirer	1
	Nvivo 11	1
Total		26

*Note.* Some Articles Employed Multiple Approaches.

**Table 3**

*Number of Articles in Web of Science in All Research Areas Covered by the Database Containing the Term “Sentiment” and the Below Mentioned Dictionary Approaches.*

Dictionary Name	Number of Articles
VADER	92
SentiStrength	66
LIWC	60
TextBlob	43
AFINN	31
MPQA	22
General Inquirer (GI)	12
NRC Emotion Lexicon	11
DICTION	10
Lexicoder	6
Hu Liu	6
Syuzhet	6
IBM Watson Tone Analyzer	3
NVIVO	2
Total	370

*Note.* Retrieval date 18.03.2022.

With regards to machine learning approaches, a similar analysis was not possible, since the algorithms employed in these types of approaches are also used in other types of statistical analyses (i.e., logistic regression is used in sentiment analysis as well in data analysis more generally) and results of such a literature search would not be trustworthy. Hence, I searched for the terms “machine learning” and “sentiment analysis” in Web of Science and coded the machine learning algorithms employed in the 20 highest cited papers that this search returned. The results of this analysis are presented in Table 4 below. The three most frequently employed algorithms are Support Vector Machines (SVM), Naïve Bayes and Maximum Entropy. This largely corresponds with the algorithms identified in the management literature search (cf. Table 2). In what follows, I further examine these three algorithm types.

**Table 4**

*Machine Learning Algorithms Employed in the 20 Highest Cited Articles of A Web of Science Search With the Terms “machine learning” and “sentiment analysis”.*

<u>Algorithm Name</u>	<u>Count per Type</u>
SVM	16
Naive Bayes	10
Maximum Entropy	6
Logistic Regression	3
KNN	2
Random Forest	2
Information not provided	1
DAN2	1
Gradient Boosting Trees	1
J48	1
Negative Binomial	1
ETC	1
OneR Algorithm	1
Radial Basis Function Neural Network	1
Decision Tree	1
SGD	1
BFTree	1
XGBoost	1
Artificial Neural Network	1
Bagging	1
Total	53

*Note.* Retrieval date 18.03.2022.

Next, I discuss the characteristics of dictionary and machine learning approaches and describe each of the specific approaches chosen for examination.

### **Dictionary Approaches**

Dictionary approaches, also referred to as “lexicon approaches”, rely on a matching mechanism between words in a given text and a dictionary that assigns valence or other types

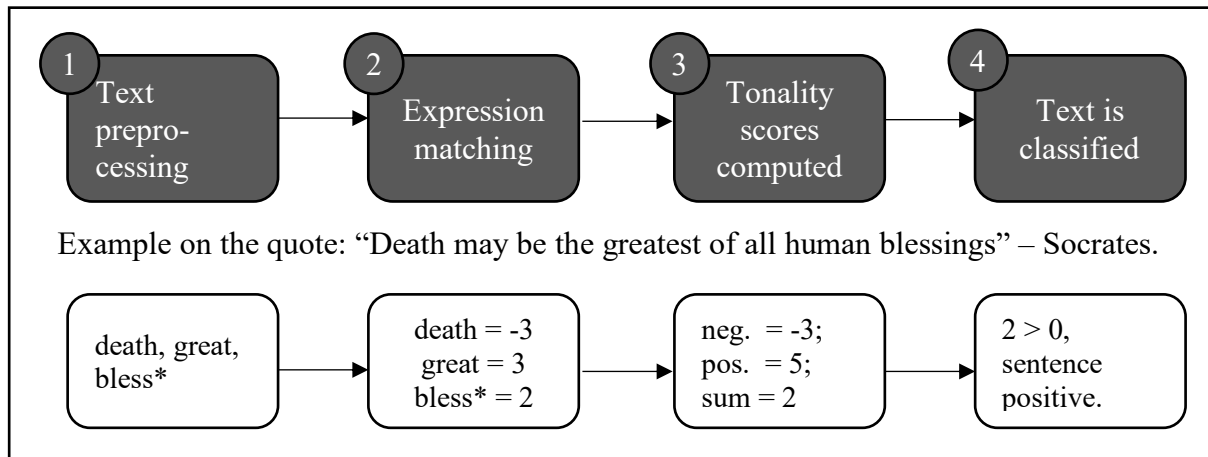
of evaluations to a given word. Dictionary approaches date back to the 1960's, when the General Inquirer (GI) was proposed as the first computer assisted content analysis tool (Stone & Hunt, 1963). Over the years, various dictionaries and associated algorithms have emerged, particularly driven by communication studies (e.g., Boukes et al., 2020; Young & Soroka, 2012). They differ on how the dictionaries are developed and what they contain: While some are expert-coded (Bradley & Lang, 1999), others make use of crowdsourcing to code individual expressions in the dictionary (Haselmayer & Jenny, 2017). Further, some dictionaries contain only adjectives (e.g., TextBlob), while others contain more diverse expressions including emoticons, abbreviations, or slang (e.g., VADER). The calculation of tonality usually differs by approach, making comparisons between them challenging.

Due to the growing number of dictionaries as well as the expansion in languages covered, it is increasingly difficult for researchers to identify the dictionary that is most suitable to use for a given task at hand. Moreover, some of the dictionaries are not freely accessible, are only available in certain languages, or are released only in a limited set of data analysis tools.

A standard procedure of sentiment analysis using dictionaries is represented in Figure 2 below and progresses through the following steps: (1) The algorithm is being fed with the words from a piece of text, the so called “bag of words” (or sometimes expressions of two or more words), the sequence and context is usually not taken into account (Scharkow, 2017). (2) By matching with the content of the dictionary, each word or expression is being assigned a score to characterize its amount of negativity or positivity (in some cases also neutrality). (3) The score for each dimension is being computed, and this differs by algorithm. (4) The final output is converted to classes of ‘negative’, ‘neutral’, and ‘positive’.

**Figure 2**

*Sentiment Analysis Workflow for Dictionary Approaches With an Example of Applying SentiStrenght.*



While dictionary approaches are straightforward to apply as they do not require training data, there is mixed evidence regarding their performance: There have been claims of human level accuracy (Hutto & Gilbert, 2014), yet others have found little overlap between the different tools, stating that they are inadequate for research purposes (Boukes et al., 2020). Performance is strongly driven by choice and potential adaptation of the dictionary (Chan et al., 2021) and validation of the results is usually encouraged. As dictionaries are “off-the shelf” (Boukes et al., 2020) and can be directly used, they are a convenient solution compared to manual coding and do not demand a pre-coded training set. This might be one reason for their popularity. Instead of using existing pre-defined dictionaries, researchers may also construct a custom dictionary for a specific task (Muddiman et al., 2019). Apart from time and resource investments, the drawback of constructing a custom dictionary would be a potential reluctance to accept results of the emerging dictionary in the peer-review process, as the novel dictionary is not considered to be established. Next, I briefly introduce the five dictionaries I chose for further examination.



The most popular dictionary I found in my search is the *VADER* dictionary (Valence Aware Dictionary for sEntiment Reasoning) (Hutto & Gilbert, 2014). It has been developed in 2014 as an extension of the LIWC dictionary discussed below, specifically for the context of social media sentiment analysis. As such, it evaluates words according to valence (non-binary categories of how positive or negative a given expression is). It also covers slang and abbreviated expressions often used in online conversations (i.e., “LOL” or “nah”) as well as emoticons (i.e., “:-)”). Moreover, it contains heuristics that cover sentiment intensity as well as double negation, among others (i.e., “The company has not been doing great.”), cases other algorithms often ignore or misinterpret. It has been shown to perform well on different types of datasets (Ribeiro et al., 2016). *VADER* has been used for example to code the sentiment of comments made on Kickstarter projects (Butticè et al., 2017).

*SentiStrength*, the second most popular dictionary approach according to my analysis, has been developed around 2010 and was subsequently refined several times (Thelwall, 2017; Thelwall et al., 2010, 2012). It’s current version contains around 2,310 expressions, and it is particularly adapted to classify sentiment of short informal texts such as social media content. It covers several linguistic heuristics (i.e., double negation) and abbreviations as well as emoticons and punctuation. It has for example been applied to classify comments from backers of Kickstarter projects to calculate the overall backer sentiment for a given project (Courtney et al., 2017).

Ranking third in my analysis is *LIWC* (Linguistic Inquiry and Word Count). *LIWC* was developed around the 1990’s to study structures of language, namely to detect negative and positive emotions in therapeutic writing (Pennebaker, 1993). It is commercially licensed and the dictionary of the most recent version contains around 6,400 words (Pennebaker et al., 2015). Apart from tone, it is also able to calculate other variables of interest, such as authenticity,

number of female/male references, or time orientation variables. An example from management research is the use of LIWC to determine media tenor towards companies (Shipilov et al., 2019).

*TextBlob* ranks fourth in popularity according to my analysis. It is part of a larger library of tools for Python that contains various text processing tools, which was started in 2013 (Loria, 2020). The sentiment dictionary approach is implemented from the *pattern* Python package that contains more than 2,900 hand-coded adjectives (Desmedt & Daelemans, 2012). It's sentiment command returns a score in the range of [-1.0, 1.0], which is the average of the respective polarities of adjectives in a text input. The documentation does not discuss any specific applications, yet it has commonly been applied for social media sentiment analysis (Micu et al., 2017).

The final approach I examined is *AFINN*, which was originally developed to analyze tweets related to the COP15-summit in 2009 as an extension of the ANEW dictionary (Bradley & Lang, 1999). The current version of *AFINN* contains almost 2,500 expressions and yields a sentiment score in the range of [-5, 5]. The algorithm has been developed and tested by the author specifically to deal with microblogging services such as Twitter (Nielsen, 2011). An example of *AFINN*'s use is the coding of polarity of tweets to determine sentiment by different stakeholders related to a company (Castelló et al., 2016).

## **Machine Learning Approaches**

Whereas machine learning algorithms for sentiment analysis pursue the same goal as dictionary approaches, determining the tonality of text, machine learning approaches differ from dictionary approaches in their application as well as how they work.

Machine learning approaches to sentiment analysis are usually supervised classification algorithms, they classify a given piece of text into pre-defined categories (e.g., 'negative', 'neutral' and 'positive') based on a subset of the data on which the algorithm has been trained

before (Kobayashi et al., 2018). Usually, this requires pre-labelling a fraction of the dataset for the training phase, thus researchers need to manually code a segment of the data they wish to classify. Supposing the algorithm performs well on the training set, there is no additional validation needed (Scharkow, 2017). Notably, one could also use an algorithm that has been pre-trained on a different dataset with similar properties and only validate the algorithm on the data of interest<sup>5</sup>.

A simplified blueprint procedure for machine learning based sentiment analysis can be found in Figure 3 below and is as follows: (1) First, the texts are pre-processed to remove unnecessary information such as stop words and words are stemmed. (2) Features from the texts are extracted. This provides a statistical model of indicative expressions such as individual words or groups of words. (3) The data is split into training and testing set. (4) The training set is manually coded into categories of ‘negative’, ‘neutral’, and ‘positive’. (5) The classifier is trained and tuned on the training set, to learn the classification rules. (6) The trained classifier is applied to the testing set, where it classifies the remaining texts, based on what it has learnt on the training set before (Grimmer & Stewart, 2013).

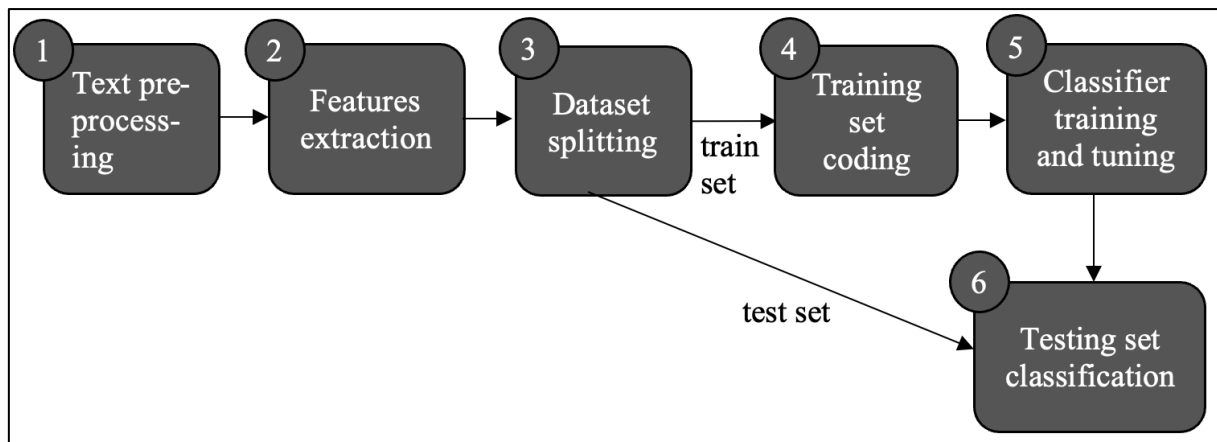
While the performance of these approaches varies with the type of classification algorithm, the finetuning of its parameters and the quality of the training data, they have been shown to outperform dictionary approaches in various applications (e.g., Nauhaus et al., 2021). Yet still, there is a risk that the algorithms might pick up certain aspects of the training set that are not necessarily accuracy enhancing or even might be misleading. For example, an unpopular politician’s name might be taken as a negative feature and hence even positive or neutral texts with this politicians name could be coded as negative (Thelwall, 2017).

---

<sup>5</sup> Large language models are a category of pre-trained algorithms, usually pre-trained on very extensive datasets. I discuss and apply two large language models later in this paper.

**Figure 3**

*Sentiment Analysis Workflow for Machine Learning Approaches.*



There are three categories of algorithms that can be used for sentiment analysis: geometric, probabilistic, and logical (Flach, 2012). Geometric algorithms map the decision objects (in this case the news articles) on a multidimensional plane that represents the decision criteria, where closeness of objects represents similarity. Classifiers then separate the different groups of objects according to certain criteria (e.g., maximize distance between groups) (Kobayashi et al., 2018). A typical example of such a classifier is *Support Vector Machines*. Probabilistic algorithms on the other hand are based on probability of attributes of an object (e.g., words in news articles) belonging to a certain classification category (Flach, 2012). An object is thus classified into the category it has the highest probability of belonging. *Naïve Bayes* is a popular algorithm belonging to this category. Finally, logical classifiers rely on decision rules that yield the classification of a given object (Kotsiantis et al., 2006). For example, if a news article contains the word “spill” but not the word “prevent”, it is classified into the negative category. A typical classifier in this category is the *Decision Tree* algorithm. Below, I introduce the three machine learning algorithms I identified in my prior analysis, two of these algorithms are probabilistic, one is geometric. My analysis does not include a logical algorithm as I did not identify one belonging to this category in my literature analysis.

As a geometric algorithm, *Support Vector Machines (SVM)* classifies data points by separating them using hyperplanes. More specifically, it attempts to construct the hyperplane(s) that have the maximum distance to the nearest training data points in each classification category (Cristianini & Shawe-Taylor, 2000). SVM requires larger training datasets (Kotsiantis et al., 2006) and does not perform well if the data is noisier, i.e., datapoints from different categories overlap, as it becomes more difficult to find a separating hyperplane (Sarker, 2021). SVM is a binary classifier, but there are various strategies to obtain a multi-category classifier and also different kernels are employed in the optimization of the hyperplane, therefore several subvariants of this algorithm exist (Liu et al., 2017). For example, two subvariants of SVM have been employed to classify sentiment of news articles from trade press that discuss capital allocations of pharmaceutical companies (Nauhaus et al., 2021).

Belonging to a probabilistic category, *Naïve Bayes* algorithms are based on Bayes' theorem which assumes the independence of attributes predictive of the classification (John & Langley, 1995), allowing for a relatively small training data set (Kotsiantis et al., 2006). The algorithm works with binary and multi-class classification categories and performs well with noisier data (Sarker, 2021). Conversely, its rather simplistic assumptions and potential data scarcity might limit performance and thus present a weakness (Samuel et al., 2020). Particularly the independence assumption is hard to defend in practice, one solution to overcome this is to use a classifier such as Maximum Entropy (Bird et al., 2009). As with SVM above, the same article also uses two subvariants of the Naïve Bayes family to classify trade press news (Nauhaus et al., 2021).

*Maximum Entropy* also belongs to the probabilistic algorithm category. It is a generalized form of the Naïve Bayes model, that needs input on what labels and features it should use (Bird et al., 2009). It has been found to work well with text classification of websites and online discussion groups (Nigam et al., 1999). Initially, the algorithm assumes a uniform

distribution for all categories, thus maximizing entropy. Based on the training data, the algorithm learns when to be minimally non-uniform, and accordingly estimates the conditional distribution of a classification category (Bergert, 1996). In my review, I encountered no example of an application of this algorithm in the management literature. Outside management, an example is the classification of Twitter data for sentiment analysis (Gautam & Yadav, 2014).

What is absent from this overview are more recent approaches using large pre-trained language models to perform sentiment analysis, these models appear to have not yet found their way into the management journals I examined. I examine two different large pre-trained models in a post-hoc analysis further below.

## **Data and Methodology**

### **Data**

#### *Data Retrieval*

The data analysis is based on the dataset from Mändli et al. (2023). This data consists of news articles on environmental issues in the 1995 – 2014 period of the 30 largest U.S. companies in terms of market capitalization, which were listed in the Standard & Poor's 500 index on December 31st, 2014. The articles had been retrieved in 2015 from Lexis-Nexis and hand-coded regarding their tonality in three categories, 'negative', 'neutral', and 'positive'. The coding was done by three independent and trained coders based on an extensive codebook which was specifically developed for the task. Krippendorff's alpha between those coders' is reported at .82 ( $p = .025$ ), values of this metric of .8 or above are considered reliable (Krippendorff, 2004, p. 241). An article was coded as 'negative' if there was any criticism towards the respective company or its' industry (and the company is mentioned as being part of the industry) in relation to environmental issues. If there was no criticism and there was remark of an improvement of environmental practices or praise of the environmental record of

either a company or its respective industry, an article was coded as positive. Finally, if an article could not be coded as neither positive nor negative, it was considered neutral.

During the manual coding procedure, only the paragraphs or sentences that were considered important to reproduce the classification were kept in a database. To obtain the full articles in a database format for the present analysis again, I coded a script that searched for the article titles and matched the date on Lexis-Nexis and retrieved the full text of the articles. Since not all articles could be matched to their originals due to ambiguity of the title (e.g., “Business News”) and because Lexis-Nexis had adapted parts of the database structure, the dataset was reduced from 16,059 to 12,929 articles. Of these, 6,581 (50.9%) were manually labeled as ‘negative’, 5,243 (40.6%) were ‘positive’ and 1,105 (8.5%) were ‘neutral’. While the reduction would be problematic if I wanted to replicate the results of the original study, this does not present an issue for the present analysis.

### *Preprocessing*

The preprocessing of the news articles was done in Python using the NLTK (natural language toolkit) library (Wang & Hu, 2021)<sup>6</sup>. Following the recommended steps on preprocessing (Bird et al., 2009), numbers, whitespace, and special characters were removed. Further, the words were transformed to their stems (e.g., “companies”, “company”, and “company’s” is transformed to ‘compan’). Additionally, the data was randomly split into two subsets, the training and finetuning set which consists of 20% of the articles (2,586) and the testing set, which consists of 80% of the articles (10,343)<sup>7</sup>. In a case of unequally distributed categories, it is recommended to use stratified sampling (Kotsiantis et al., 2006), to allow

---

<sup>6</sup> <https://pypi.org/project/nltk/>

<sup>7</sup> In machine learning, it is common to use around 2/3 of the data for training (e.g., Kotsiantis et al., 2006). While using larger proportions of data for training, in this context, 20% of training data seems a more reasonable amount, considering that time savings and resources only apply if there is less manual coding to be done. I examine the effect of size of the training set on performance later in this paper.

machine learning algorithms to learn each category of tone sufficiently well. Nevertheless, I chose random sampling under the assumption that a researcher needs to select testing and training sets before having hand-coded the training data, thus she/he cannot determine the distribution of tonalities ex-ante. I repeated the analysis, each time with a novel random 20% / 80% split of training and testing data over 100 replications and averaged the resulting metrics for all machine learning approaches in the paper to rule out that results are driven by chance.

### *Dictionary Approaches*

Each of the dictionary approaches – AFINN, LIWC, SentiStrength, TextBlob, and VADER – were run on the preprocessed training set first. AFINN<sup>8</sup> and TextBlob<sup>9</sup> are provided in ready-made packages for Python by the authors, VADER is included in the NLTK package<sup>10</sup>. For LIWC<sup>11</sup> and SentiStrength<sup>12</sup>, the applications were not originally developed for Python, thus the dictionaries had to be obtained and loaded manually with respective user-developed packages.

Depending on the dictionary, the calculations of tonality that are returned differ, which makes their comparison more challenging. Table 5 further below presents the output score of each dictionary approach examined as well as the proposed standard cutoffs from the literature. Since the articles needed to be classified into the three categories ‘negative’, ‘neutral’, and ‘positive’ and none of the approaches provided these measures directly, the cutoffs had to be set manually. The choice of these cutoffs obviously influences the performance parameters of the respective approaches. This is more challenging, if there is a neutral category: One needs to determine what is an adequate window to choose for the neutral category, or differently, how

---

<sup>8</sup> <https://pypi.org/project/afinn/>

<sup>9</sup> <https://pypi.org/project/textblob/>

<sup>10</sup> [https://www.nltk.org/\\_modules/nltk/sentiment/vader.html](https://www.nltk.org/_modules/nltk/sentiment/vader.html)

<sup>11</sup> <https://pypi.org/project/liwc/>

<sup>12</sup> <https://pypi.org/project/sentistrength/>



negative or how positive can an article be to be considered as ‘neutral’? An additional reason for this step was the aim to closely mirror the manual coding, where articles that contained criticism were categorized as ‘negative’ and only absent criticism and presence of praise was coded as ‘positive’. To examine whether the proposed cutoffs should be adapted, I used an algorithm based on Python’s `scipy.optimize` package<sup>13</sup>, which maximized the respective Macro-F1<sup>14</sup> score between the manual coding of the training set and the one obtained by each of the dictionary approaches. There is a risk that the algorithm allocates as many news as possible to the categories negative and positive while ignoring the neutral category, due to the following reasons: First, because the neutral category is in the middle of the other two in terms of cutoffs, the smaller this window is chosen, the higher the likelihood that the other two categories score high in accuracy. Second, the dataset contains fewer news that have been manually categorized as neutral, hence accuracy would be higher if the neutral category is completely ignored. As Macro-F1 weights each category equally, this reduces the risk of such a scenario. A comparison of the performance of both, standard and optimized cutoffs, can be found in Table 6 below.

The optimized cutoffs performed better with regards to the Macro-F1 score for all dictionary approaches, consequently those were chosen for the analysis of the testing set.

---

<sup>13</sup> <https://docs.scipy.org/doc/scipy/reference/optimize.html>

<sup>14</sup> A more detailed description of the Macro-F1 score is provided in the section Results below.

**Table 5**

*Output Scores and Respective Standard Cutoffs for Trinary Sentiment Classification.*

Name	Output Score	Standard Cutoffs		Source
		Lower	Upper	
AFINN	Single polarity score (-5 to 5).	-0.25	0.25	(Nielsen, 2011)
LIWC	Number of positive emotions - number of negative emotions.	-1	1	(Pennebaker et al., 2015)
SentiStrenght	Sum of negative (-5 to -1) and positive (1 to 5) score.	-1	1	(Thelwall et al., 2017)
TextBlob	Single polarity score (-1 to 1).	-0.05	0.05	(Loria, 2020)
VADER	Adjusted score of negative, positive and neutral words (-1 to 1).	-0.05	0.05	(Hutto & Gilbert, 2014)

*Note.* Classification as ‘negative’ if score  $\leq$  lower cutoff, as ‘positive’ if score  $\geq$  upper cutoff, else as ‘neutral’.

**Table 6**

*Performance Metrics of Dictionary Approaches With Standard and Adapted Cutoffs for Trinary Sentiment Classification.*

Name	Standard Cutoffs		Standard Cutoff Performance			Optimized Cutoffs		Optimized Cutoff Performance		
	Lower	Upper	Overall accuracy	Krippendorff's alpha	Macro-F1	Lower	Upper	Overall accuracy	Krippendorff's alpha	Macro-F1
AFINN	-0.25	0.25	0.62	0.30	0.44	5.16	10.02	0.62	0.34	0.49
LIWC	-1	1	0.48	0.10	0.37	6.40	8.15	0.60	0.28	0.45
SentiStrenght	-1	1	0.19	-0.29	0.18	-1.60	-0.62	0.38	0.02	0.32
TextBlob	-0.05	0.05	0.35	-0.10	0.26	0.08	0.09	0.51	0.14	0.41
VADER	-0.05	0.05	0.59	0.22	0.41	0.81	0.97	0.56	0.28	0.47

*Note.* As Krippendorff's alpha is adjusted by agreement that could be obtained by chance, negative values occur if classification performance is worse than chance.

## *Machine Learning Approaches*

The machine learning approaches were implemented in Python using the scikit-learn package<sup>15</sup>. In a first step, the texts needed to be transformed to a readable format for the machine learning algorithms. I used the Term Frequency Inverse Document Frequency (TFIDF) command<sup>16</sup> from the scikit-learn package, which calculates the term frequency (how many times a word appears in a text divided by the total number of words in the text) and multiplies this with the inverse document frequency (the log of the total number of texts, divided by the number of texts that contain the a given word). The advantage of this approach is that instead of simply counting the number of times a given word appears in each text, words that appear frequently across all documents (such as “this”, “are”, etc.) are deemphasized (Li-Ping Jing et al., 2002).

This stage of data preparation is important as it potentially affects the results (Kotsiantis et al., 2006). Apart from determining the number of terms used as well as the upper and lower limits of their frequency, one can also vary the nature of the terms, whether they are individual characters, words or groups of words (called n-grams). Tuning the parameters for feature selection on the training data should increase the performance on the data to test (Yu & Liu, 2004). I conducted an analysis using scikit-learn’s Pipeline<sup>17</sup> and GridsearchCV<sup>18</sup> modules, where I varied the feature selection parameters and examined their respective performance on the training data, to find the optimal combination of the feature selection parameters. Table 7 below presents the optimal parameters found for each of the machine learning approaches examined. In this step, I simultaneously removed all English stop words.

Next, all three classifiers (Support Vector Machines, Maximum Entropy, and Naïve Bayes) were trained on the training dataset. This means that given the text and the respective

---

<sup>15</sup> <https://scikit-learn.org/>

<sup>16</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfVectorizer.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

<sup>17</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>

<sup>18</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)

labelling as ‘negative’, ‘positive’, or ‘neutral’, the classifiers determined the features that were most indicative of each label based on their respective algorithms. For the Support Vector Machine class of algorithms, of the available subtypes, the C-Support Vector Classification (SVC) algorithm performed best on the training set. Similarly, from the subtypes of Naïve Bayes algorithms, Bernoulli Naïve Bayes provided the best results on the training data. For Maximum Entropy, there are no subcategories, the actual implementation is based on Logistic Regression.

Finally, I optimized all three above mentioned classifiers’ hyperparameters (parameters such as decision function shape, regularization parameters and bounds for support vectors) on the training data set using Python’s scikit-learn GridsearchCV module<sup>19</sup>. This divides the testing set into random parts used as testing and training sets for the optimal hyperparameter combination, to prevent biased overfitting of the parameters on the entire training data set (Kotsiantis et al., 2006), where the algorithm would not perform well on the testing data (Vabalas et al., 2019).

For example, the data is divided into 10 different validation parts, each of them not containing data from the other parts. In a first iteration, the model is validated on part 1 and trained on parts 2, 3, ... , 10. In a second iteration, the algorithm uses part 2 for validation and the remaining parts 1, 3, 4, ... , 10 for training. After 10 iterations, the results across each of the 10 parts are averaged for every combination of hyperparameters. The combination that yields the highest average Marco-F1 score is returned. Table 8 below reports the initial metrics for each of the three machine learning approaches, as well as the metrics after hyperparameter tuning. Only non-default hyperparameters are reported.

---

<sup>19</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)

**Table 7**

*Optimal Feature Selection Parameters for Machine Learning Approaches.*

Name	Maximum Number of Features	Maximum Document Frequency*	Minimum Number of Documents†	N-Grams‡
Bernoulli Naïve Bayes	1500	0.60	50	Unigrams, bigrams, trigrams
Maximum Entropy	2800	0.45	5	Unigrams
Support Vector Machines	2900	0.70	15	Unigrams

*Note.* \* Maximum of documents (e.g., news articles) that contain a specific feature (e.g., word). † Minimum number of documents (e.g., news articles) that contain a specific feature (e.g., word). ‡ Features can be individual unigrams, such as words (e.g., "environment") or multigrams, such as tuples of words (e.g., "environmental pollution" or "increased environmental pollution").

**Table 8**

*Performance Metrics of Machine Learning Classifiers Applied to Manually Coded Training Set, Before and After Hyperparameter Tuning.*

Name	With Default Hyperparameters			After Hyperparameter Optimization			Hyperparameters Set Differently from Default
	Overall accuracy	Krippendorff's alpha†	Macro-F1	Overall accuracy	Krippendorff f's alpha	Macro-F1	
Bernoulli Naïve Bayes	0.71	0.51	0.65	0.75	0.59	0.69	alpha = 0.01, binarize = None, fit_prior = False C = 10, penalty = 'l2', solver = 'newton-cg' C = 3, gamma = .4
Maximum Entropy	0.85	0.73	0.71	0.96	0.93	0.95	
Support Vector Machines	0.93	0.88	0.88	0.95	0.92	0.92	

## Results

Table 9 below reports the results of the chosen dictionary and machine learning approaches, run on the testing set and compared to manual coding, providing three different measures of performance: 1) Accuracy is the percentage of articles that are correctly classified in their respective category divided by the total number of articles (Zhang et al., 2014). This is the same measure as the agreement measure for intercoder reliability (Krippendorff, 2004). Since accuracy ignores how the categories are distributed in the data, in cases where the underlying data is not equally distributed among categories, this metric might be misleading. 2) Macro-F1 considers the effectiveness of classification in each category (Ribeiro et al., 2016). F1 is calculated as the harmonic mean of precision, the ratio of the number of news correctly classified to the number of news classified within a given category, and recall, which is the ratio of number of news correctly classified to the number of news that belong in this category (Lewis, 1995). Macro-F1 is then computed by averaging over the classes (Lewis et al., 2004). Both, accuracy and Macro-F1 are measures that do not consider that correct classification could also happen by chance. 3) A metric that takes into account the agreement between raters that could have been obtained by chance is Krippendorff's alpha (Krippendorff, 2004), in simple terms it is calculated as accuracy adjusted by chance probability. As each of these measures captures different aspects of classification quality, I calculate all three, accuracy, Krippendorff's alpha and Macro-F1 for each of the approaches I examined and report them. I also report the accuracy per category of tone, i.e., the percentage of 'negative' news correctly classified<sup>20</sup>.

A common approach to improve results of classification is using votes among different classifiers or combining machine learning and dictionary approaches (Dhaoui et al., 2017).

---

<sup>20</sup> As detailed earlier, all Machine Learning approaches are run 100 times with a 20/80 random split of the training/testing data, thus the reported metrics represent averages over these 100 replications.

Analogous to political voting, there are different procedures that could be used (Kuncheva & Rodríguez, 2014). I implemented three different majority votes: The first votes among the dictionary approaches (Dictionary Vote), the second votes among the machine learning approaches (Machine Learning Vote) and the third among all implemented algorithms (Global Vote). The scores of these voting classifiers are also shown in Table 9 below.

The results indicate that for classification of news on environmental issues of companies, machine learning approaches outperform dictionary approaches against the benchmark of manual coding, yet also those results are far from accurate. Krippendorff's alpha is recommended to be at least above .6, ideally above .8 (Krippendorff, 2004), yet no approach scores higher than .61. A second baseline is to consider the accuracy when the most dominant sentiment classification is chosen. In the present case, 50.9% of the articles are 'negative', hence if an algorithm classifies all articles as negative, the accuracy would be .51. The results show that two of the dictionary approaches are unable to surpass this default accuracy. Support Vector Machines performs highest in terms of accuracy and Krippendorff's alpha, yet it suffers from relatively low accuracy when classifying 'neutral' news. Bernoulli Naïve Bayes yields the best results in classification of 'neutral'. Comparable relative performance among the three machine learning algorithms I examined have been found when classifying movie reviews (Samal et al., 2017). Similar Macro-F1 scores are reported from trinary classification of reviews using dictionary approaches (e.g., Ribeiro et al., 2016), and machine learning approaches (e.g., Ahmad et al., 2018).

Finally, the voting algorithms provide marginal improvements in terms of performance. In particular, the Global Vote, that combines dictionary and machine learning approaches, yields a marginal improvement recorded in all three performance measures. It is notable that voting increases the accuracy of the 'positive' classification. In what follows, I explore several parameters that could additionally be examined to further improve performance.

**Table 9**

*Performance Metrics of Dictionary and Machine Learning Approaches Applied to Testing Set, Evaluated Against a Benchmark of Manual Coding.*

Approach Type	Name	Overall Accuracy	Krippendorff's alpha	Macro-F1	Accuracy Negative	Accuracy Neutral	Accuracy Positive
Machine Learning	Bernoulli Naïve Bayes	0.72	0.54	0.64	0.68	<b>0.52</b>	0.82
	Maximum Entropy	0.77	0.59	0.63	0.84	0.22	0.81
	Support Vector Machines	<b>0.79</b>	0.61	0.64	<b>0.85</b>	0.18	0.83
	<i>Machine Learning Vote</i>	<b>0.79</b>	<b>0.62</b>	<b>0.65</b>	0.84	0.23	0.84
Dictionary	AFINN	0.6	0.31	0.47	0.56	0.1	0.76
	LIWC	0.57	0.25	0.44	0.56	0.09	0.68
	SentiStrenght	0.39	0.03	0.32	0.15	0.28	0.71
	TextBlob	0.51	0.16	0.41	0.56	0.13	0.54
	VADER	0.56	0.27	0.46	0.48	0.23	0.72
	<i>Dictionary Vote</i>	0.6	0.3	0.46	0.55	0.09	0.78
<i>Global Vote</i>		0.78	0.6	0.64	0.8	0.21	<b>0.87</b>

*Note.* The highest value in each column is highlighted in bold.



## **Post Hoc Analysis I – Improving Performance**

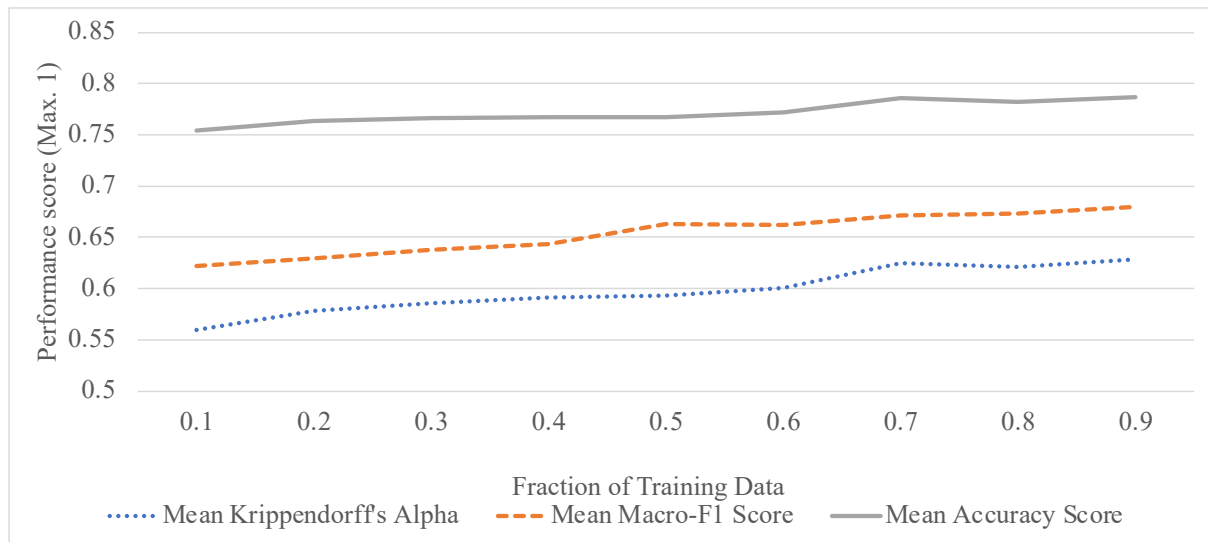
### *Training Set Size*

The performance of machine learning approaches depends on the training set chosen, both in terms of size and representativeness of the data (Vabalas et al., 2019). While more training data generally improves results, the goal would be to determine a reasonable amount of training data that yields adequate results (Cho et al., 2016), as classifying news articles manually is resource intensive. Not only the amount of training data matters, but also the quality and its representativeness of the total data (Batista et al., 2004). If it is heavily skewed or poorly coded, then an algorithm will learn such biases or encounter more noise when learning. In both cases, this usually results in lower performance. I varied the amount of training data and computed the average performance score for each of the three metrics over the three machine learning algorithms, Bernoulli Naïve Bayes, Maximum Entropy, and Support Vector Machines. The results from this analysis are shown in Figure 4 below.

With increasing size of the training dataset, all performance metrics improve. With increasing size of the training data set, performance improves. Beyond 70% training data, this trend appears to stall. Thus, the ideal choice in this case would be 70%, yet higher amounts of training data also would lead to substantially more manual coding labor. In machine learning, a training set size of around 70% is common practice (Samal et al., 2017). A cutoff of 20% as chosen in the current setup nevertheless is more realistic, given that when applied, coding 70% of the data manually would diminish the benefits of applying such an approach in terms of cost and time. In practice, this type of information is usually not available ex-ante, as the testing dataset is not pre-coded.

**Figure 4**

*Mean Performance Measures of all Three Machine Learning Algorithms With Increasing Training Data Fraction.*



*Note.* Algorithms Examined: Bernoulli Naïve Bayes, Maximum Entropy and Support Vector Machines

*Sentences, Excerpts, and Full News Articles*

An approach that could potentially increase the performance of both categories of classifiers would be to only select sentences that contain the respective company's name. As news articles might contain a lot of additional unnecessary information, also in terms of tonality, sentences which contain the company's name might reduce the amount of noisy and/or irrelevant information (Balahur et al., 2013). Also, journalists might actually position their criticism or praise for the company around the company name, which is why selecting sentences containing the company name could be representative of the tonality. I tested this in an additional analysis: In a first step, I identified the variants of company names. This is challenging in some cases. For example, if the full name 'Cisco Systems' is used to identify the sentences containing the company, cases where the company is called 'Cisco System', 'Cisco-

systems' or 'Cisco' are missed. On the other hand, I would identify too many false-positive sentences using only one of the terms. 'Cisco' is problematic, because it is contained in the city name of San Francisco and the term 'systems' might relate to many other issues that could be discussed in the news articles. In this case, I had to ensure that San Francisco is not identified and then use 'cisco' as the identifying term. After running a script that isolates sentences with the company's name in them, I dropped those articles where no company name was identified through this process. The dataset subsequently was reduced to 9,652 articles that only contained sentences with company names.

Further, from the original hand-coding of the data, excerpts of the articles were available. These extracts had been archived during the coding process in order to keep the most important passages of each article and should allow to reconstruct the hand coded classification decision. Those reduced news articles are more extensive than only the sentences containing the company names and might capture the tonality towards the company and the specific environmental issue(s) mentioned in the article more accurately. I thus also examined whether these extracts could yield superior performance. In sum, I analyzed and compared three different versions of the news articles: The full news articles as in the main analysis, the sentences containing only company names, and the archived excerpts from the hand coding process. Table 10 further below reports the performance results for each examined approach and each machine learning classifier as well as AFINN and VADER, which had the highest Macro-F1 score among the dictionary approaches on the entire news articles.

As these results show, using only the sentences with the company's name provides worse results than using the full news articles. Exploiting the excerpts improves the performance of all approaches examined. However, manually pre-scanning each article for the relevant parts that contain sentiment towards the company and the company's environmental

issues would represent a substantial investment and mitigate the benefit of using an automated approach in the first place.

### *Binary vs. Trinary Classification*

In the results I obtain, all approaches score particularly low in the ‘neutral’ category. With increasing number of categories, performance of text classification is generally decreasing (Bouazizi & Ohtsuki, 2016). To examine whether performance improves when only two categories are used, I collapsed the news which were hand-labeled as ‘positive’ and ‘neutral’ into the category ‘non-negative’. Table 11 below compares the results of using two and three categories for all approaches examined. Since the optimal parameters for trinary classification are not necessarily ideal for binary classification, I recalculated the optimal cutoffs for the dictionary approaches, as well as the optimal parameters for both, the feature selection and the algorithms for the machine learning approaches.

These results show, that within each approach, using a binary classification yields better performance. This is not surprising, as for the dictionary approaches, it is easier to determine the optimal cutoffs when only two categories are used. Machine learning approaches should also perform better as they have more data by category to train on. It is notable that in terms of performance, AFINN remains the best performing dictionary approach and Support Vector Machines the best performing machine learning approach when switching from trinary to binary classification. Depending on the research question at hand, if one is solely interested in the negative articles and uses the non-negative as a control variable or is not using this information, binary classification might be sufficient and should thus be preferred.

**Table 10**

*Performance Metrics of the Best Performing Dictionary and Machine Learning Approaches Applied to Variants of the News Articles Testing Set, Evaluated Against a Benchmark of Manual Coding.*

Name	Data Version	Overall Accuracy	Krippendorff's alpha	Macro-F1	Accuracy Negative	Accuracy Neutral	Accuracy Positive
Bernoulli Naïve Bayes	Full News Articles	0.72	0.54	0.64	0.68	<b>0.52</b>	0.81
	Sentences with Company Name*	0.72	0.52	0.63	0.71	0.42	0.83
	Excerpts from Hand Coding <sup>†</sup>	0.76	0.59	<b>0.67</b>	0.72	0.48	<b>0.87</b>
Maximum Entropy	Full News Articles	0.77	0.58	0.63	0.85	0.19	0.8
	Sentences with Company Name*	0.75	0.52	0.62	0.88	0.23	0.7
	Excerpts from Hand Coding <sup>†</sup>	0.79	0.62	0.65	0.86	0.21	0.83
Support Vector Machines	Full News Articles	0.78	0.6	0.63	0.85	0.16	0.82
	Sentences with Company Name*	0.76	0.54	0.62	<b>0.89</b>	0.18	0.72
	Excerpts from Hand Coding <sup>†</sup>	<b>0.8</b>	<b>0.64</b>	0.65	0.86	0.19	0.86
AFINN	Full News Articles	0.6	0.31	0.47	0.56	0.1	0.77
	Sentences with Company Name*	0.5	0.13	0.39	0.53	0.1	0.57
	Excerpts from Hand Coding <sup>†</sup>	0.63	0.35	0.49	0.6	0.12	0.76
VADER	Full News Articles	0.56	0.27	0.46	0.48	0.22	0.74
	Sentences with Company Name*	0.49	0.12	0.39	0.63	0.23	0.34
	Excerpts from Hand Coding <sup>†</sup>	0.63	0.34	0.49	0.69	0.13	0.65

*Note.* The highest value in each column is highlighted in bold.

\* Only sentences that contained the company name were used. Dataset reduced to N = 9,652 news articles.

<sup>†</sup> Excerpts only contain the relevant parts of each article that lead to the manual coding assessment.

**Table 11**

*Performance Metrics of Trinary and Binary Classification using Dictionary and Machine Learning Approaches Applied to Testing Set, Evaluated Against a Benchmark of Manual Coding.*

Approach Type	Name	Classification	Overall Accuracy	Krippendorff's alpha	Macro-F1	Accuracy Negative	Accuracy Neutral	Accuracy Positive	Accuracy Non-neg.
Machine Learning	Bernoulli	Binary	0.8	0.61	0.8	0.78			0.83
	Naïve Bayes	Trinary	0.72	0.54	0.64	0.68	<b>0.52</b>	0.82	
	Maximum Entropy	Binary	0.82	0.65	0.82	0.81			0.83
	Support Vector Machines	Trinary	0.77	0.59	0.63	0.84	0.22	0.81	
	<i>Machine Learning Vote</i>	Binary	<b>0.83</b>	<b>0.65</b>	<b>0.83</b>	0.82			0.84
		Trinary	0.79	0.61	0.64	<b>0.85</b>	0.18	0.83	
		<i>Binary</i>	0.82	<b>0.65</b>	0.82	0.81			0.84
		<i>Trinary</i>	0.79	0.62	0.65	0.84	0.23	0.84	
Dictionary	AFINN	Binary	0.69	0.38	0.69	0.56			0.83
		Trinary	0.6	0.31	0.47	0.56	0.1	0.76	
	LIWC	Binary	0.65	0.3	0.65	0.56			0.74
		Trinary	0.57	0.25	0.44	0.56	0.09	0.68	
	SentiStrenght	Binary	0.62	0.25	0.62	0.55			0.7
		Trinary	0.39	0.03	0.32	0.15	0.28	0.71	
	TextBlob	Binary	0.59	0.17	0.59	0.6			0.57
		Trinary	0.51	0.16	0.41	0.56	0.13	0.54	
	VADER	Binary	0.64	0.23	0.61	0.36			<b>0.94</b>
		Trinary	0.56	0.27	0.46	0.48	0.23	0.72	
<i>Dictionary Vote</i>	<i>Binary</i>	<i>0.68</i>	<i>0.35</i>	<i>0.67</i>	<i>0.52</i>			<i>0.85</i>	
	<i>Trinary</i>	<i>0.6</i>	<i>0.3</i>	<i>0.46</i>	<i>0.55</i>	<i>0.09</i>	<i>0.78</i>		
<i>Global Vote</i>	<i>Binary</i>	<i>0.82</i>	<i>0.64</i>	<i>0.82</i>	<i>0.78</i>			<i>0.87</i>	
	<i>Trinary</i>	<i>0.78</i>	<i>0.6</i>	<i>0.64</i>	<i>0.8</i>	<i>0.21</i>	<b><i>0.87</i></b>		

*Note.* The highest value in each column is highlighted in bold.

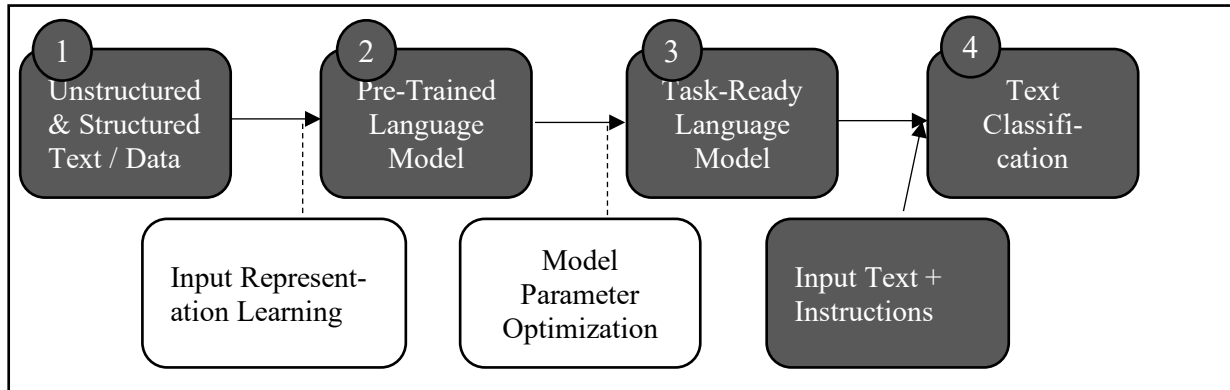
## **Post Hoc Analysis II – Large Language Models**

Pre-trained language models have emerged as a promising approach to sentiment analysis in recent years (Kheiri & Karimi, 2023). These models are mostly based on a transformer architecture that was initially developed for translation tasks (to transform text from one language to another) (Vaswani et al., 2023). They are pretrained on large collections of texts such as Wikipedia and have millions or even billions of parameters. A popularly known language model group are the GPT models, version 3 of this model has around 175 billion parameters (Brown et al., 2020). These models take into account the context of text, making them suitable for classifications, translations and/or summarizing text among other tasks (Li et al., 2022). For example, ChatGPT and variants of the BERT models have been shown to perform well at sentiment analysis tasks (e.g., Zhong et al., 2023). It is probable that such models will be regularly used for sentiment analysis in management research in the future. Apart from cost and time considerations, higher accuracy would permit their use also from a research perspective (Xu et al., 2021).

While large language models usually also need to be fine-tuned (optimization of model parameters) to the specific task intended, one particular benefit of ChatGPT is that it can excel at diverse tasks without specific adaptations (Li et al., 2022). A simplified view of a process such as sentiment analysis (which represents a classification task) using a large language model is shown in Figure 5 below. The original transformer architecture is an encoder-decoder type of model, as translation tasks require original language text to be encoded, which would be to generate a numeric representation of the text, and then decoded again, which would be to generate target language text out of the numeric representation. Yet some models are encoder or decoder only. Both models I examine subsequently are encoder-only models, thus they convert a text input into a meaningful representation to have an understanding of what it is, in this case what kind of sentiment.

**Figure 5**

*Simplified Sentiment Analysis Workflow for Large Pre-Trained Language Models.*



The BERT-family (Bidirectional Encoder Representations from Transformers) of models is a group of pretrained language models that are specifically adapted to understand the context of a given text (Devlin et al., 2019). It is highly popular in Natural Language Processing research and has been improved over time, making it a standard baseline model in many experiments (Rogers et al., 2020). A straightforward application of BERT models is sentiment classification, as this requires an understanding of the input provided. I examined DeBERTaV3<sup>21</sup> (Decoding-enhanced BERT with disentangled attention version 3), which is an improved version of the original BERT, that has been shown to perform excellent on various tasks of natural language understanding (He et al., 2023). This model has been pre-trained on Wikipedia content, a conversational archive built on Reddit, and a book corpus (He et al., 2021). I fine-tuned the model on a training set of 2,400 randomly selected news articles (around 20% of the dataset) and ran the model on a subset of 20% of the news, randomly selected among the

<sup>21</sup> <https://github.com/microsoft/DeBERTa>



shorter 50% of news articles<sup>22</sup> using python's Transformer package<sup>23</sup>. Performance results of this analysis are reported in Table 12 below.

GPT (Generative Pre-trained Transformer) is another group of models, that are generalists in their capabilities. ChatGPT has been proposed as a suitable tool for sentiment analysis, one approach being to pass instructions to the model, similar to a codebook for human coders (Kheiri & Karimi, 2023). Using the dataiku<sup>24</sup> platform, I passed on brief coding instructions similar to the ones used in (Mändli et al., 2023) along with each news article to examine by the algorithm. I used the GPT 3.5 Chat (a.k.a. Turbo). The model then returned the classification it determined the most likely. Interestingly, this approach allows to return a textual description of the reasoning behind the chosen classification of sentiment. It would have also allowed to pass on examples of classifications for the model, a step which I skipped after initial examination due to two reasons: First, this would have resulted in additional amounts of text being transferred along with the actual news articles to classify, stressing even further the input amount limits. Second, as the classification of the articles is highly complex, passing only few examples might bias the results in unintended ways. I examined 20% of randomly selected news articles as for DeBERTaV3, results are reported in Table 12 below.

The results show that DeBERTaV3 outperforms all approaches examined thus far on the subsample of news articles on all overall metrics by a notable margin. It also shows that ChatGPT reaches similar performance levels as the best performing machine learning approaches. This is remarkable, as apart from brief coding instructions, ChatGPT was not trained on the specific task of classification of the news articles, thus there is no need to

---

<sup>22</sup> In both cases where I test the large language models, I was unable to test the entire dataset because these models have limits in terms of input length. Many of the news articles exceed this length, with the ChatGPT model there are also instructions sent with each article, thus there is even less available space for news articles. Nevertheless, using 20% of the data at random for comparable results regarding their performance vis-à-vis the other approaches.

<sup>23</sup> <https://pypi.org/project/transformers/>

<sup>24</sup> <https://www.dataiku.com>

manually code a subset of the articles prior to running the algorithm (called zero-shot). While DeBERTaV3 was trained on a subset of the pre-coded data, similar to the other machine learning approaches, the amount of work involved for the ChatGPT approach is comparable using dictionary approaches, yet it is substantially more accurate.

**Table 12**

*Results of Large Language Models DeBERTaV3 and ChatGPT Run on a Subset of 20% of the News Articles Reported Among the Best Performing Machine Learning and Dictionary Approaches.*

Approach Type	Name	Overall Accuracy	Krippendorff's alpha	Macro-F1	Accuracy Negative	Accuracy Neutral	Accuracy Positive
Large Language Models	ChatGPT	0.78	0.61	0.63	0.8	0.26	0.84
	DeBERTa V3	<b>0.85</b>	<b>0.72</b>	<b>0.72</b>	<b>0.87</b>	0.38	<b>0.9</b>
Machine Learning	Support Vector Machines	0.79	0.61	0.64	0.85	0.18	0.83
	<i>Machine Learning Vote</i>	0.79	0.62	0.65	0.84	0.23	0.84
Dictionary	AFINN	0.6	0.31	0.47	0.56	0.1	0.76
	<i>Dictionary Vote</i>	0.6	0.3	0.46	0.55	0.09	0.78
<i>Global Vote</i>		0.78	0.6	0.64	0.8	0.21	<b>0.87</b>

*Note.* The highest value in each column is highlighted in bold.

### Post Hoc Analysis II – Impacts of Using Automated Coding on Research Results

Apart from a consideration of how closely manual sentiment coding can be matched using either dictionary or machine learning approaches, it is important to consider what impact such automated coding might have on the conclusions drawn when answering specific research questions. For this purpose, I compare results of a simplified panel-data model inspired by the

original paper (Mändli et al., 2023) to examine potential differences using the best performing approaches from the analysis above. In particular, I ran panel data fixed-effects regression models, where the dependent variable is the number of negative news in a given month  $t$ . The independent variables are constructed as the sum of news in each category of negative, neutral, and positive, over the annual period preceding the focal month  $t$ . I also employ the same control variables as in the original analysis (MSCI ESG environmental score, company assets, profitability, leverage, net sales, and cash flow).

Column 1 in Table 13 below shows the results of the company fixed-effects model run on hand-coded data. Columns 2, 3, and 4 report the results of company fixed-effects models on the machine-coded data from the best dictionary approach (AFINN), the best machine learning approach (SVM), and the Global Vote among all examined approaches, respectively<sup>25</sup>. Standard errors are clustered at the company level.

The Global Vote approach provides results that match the results from the original hand coding the closest, in terms of which estimates are significant as well as their direction. The approaches scoring higher in the prior examination of all approaches (c.f. Table 9 above), appear to perform better in this actual analysis. However, estimates of coefficients differ in size. For example, the variable “Sum of Neutral Articles”, has an estimated coefficient size of .14 in the hand-coded model, while the SVM model estimates this coefficient at .35 and the global vote at .05.

This shows that given this particular dataset and the respective manual coding procedure as well as the examined automated approaches, in an analysis such as the one I conducted above, automated approaches are unable to match manual coding. However, this does not allow to rule out that other automated sentiment approaches, other datasets or a different analysis would provide a similar verdict.

---

<sup>25</sup> I could not compare results from the large language models (ChatGPT and DeBERTaV3), as they did not include the entire sample examined in the other approaches.

**Table 13**

*Results From Panel Fixed-Effect Regressions Based on Data From the Best Performing Coding Approaches and the Manually Coded Data.*

Model	1	2	3	4
News Article Coding Type	Hand-Coding	AFINN Coding	SVM Coding	Global Vote Coding
Variables	Sum of Negative Articles in Month t			
Sum of Negative Articles	0.040*** (0.006)	0.030*** (0.009)	0.038*** (0.008)	0.030** (0.012)
Sum of Positive Articles	-0.001 (0.001)	0.011 (0.012)	0.001 (0.003)	-0.003 (0.003)
Sum of Neutral Articles	0.140** (0.068)	0.003 (0.042)	0.346* (0.198)	0.052** (0.019)
MSCI ESG Environmental Score	0.012 (0.025)	-0.009 (0.018)	-0.000 (0.033)	-0.015 (0.020)
Assets	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Profitability	0.027 (0.032)	0.007 (0.013)	0.021 (0.027)	0.012 (0.025)
Leverage	-0.002 (0.003)	-0.003 (0.003)	-0.002 (0.002)	-0.004 (0.003)
Net Sales	-0.000 (0.001)	-0.000 (0.001)	0.001 (0.002)	-0.003 (0.002)
Cash Flow	0.026 (0.020)	0.035 (0.022)	0.008 (0.012)	0.043 (0.028)
Constant	0.191 (0.205)	0.294 (0.191)	0.366 (0.219)	0.561** (0.208)
Observations	4,393	4,393	4,393	4,393
R-squared	0.104	0.051	0.102	0.063
Number of Companies	30	30	30	30

*Note.* Robust standard errors in parentheses.

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

All independent and control variables are annual values corresponding to months t-1 to t-12.

## Discussion

My analysis yields mixed results regarding the appropriateness of automated approaches commonly used in management science to code tonality of news articles. While careful examination and setup might warrant the use of some approaches, it is unclear whether the results would be of sufficient quality. In particular, depending on how the coded data is used afterwards, conclusions based on the coded data might be conditional on the choice of approach.

A first outcome from my analysis is the superiority of machine learning approaches over dictionary approaches on the examined dataset. This is in contrast to other studies that have found similar performance levels for both, dictionary-based and machine learning based sentiment analysis (e.g., Dhaoui et al., 2017). A potential explanation for this could be the structure of the data: News articles are usually longer than for example social media posts or product reviews and they may contain a lot of additional information unrelated to the subject examined. While dictionary approaches would use the entire text, machine learning approaches “learn” to ignore the unrelated parts of text given a quality training set. Further, the tonality of shorter texts might be more homogeneous than the tonality of longer texts, for example news articles try to capture the reader’s attention by using narration and different aspects of or views on an issue (Curran et al., 2017). A dictionary approach that reports the overall tonality in an article most likely fails with increasing heterogeneity of tonality in the text.

Second, common practice of using approaches that have been employed in prior literature might lead to missing out more promising approaches proposed in other fields. As I show in my supplemental analysis, a model that does not require any training like ChatGPT performs with similar precision as the supervised machine learning models after several steps of pre-training and adjusting. A pre-trained large language model such as DeBERTaV3 might even be able to deliver results that approach manual coding accuracy. I thus recommend

researchers to include such novel models in their choice set of algorithms when starting a new project and examine their performance in pre-tests.

Third, as my analysis shows, close examination, cross-validation, and tuning of parameters at all stages of the process is pivotal. These steps also rely on manually pre-coded subset of the data used for testing and calibration, which ideally is of high quality and representative of the data. If done correctly, improvements are possible at every step of the sentiment analysis process. It is consequently important for researchers to understand what can be improved at every stage and how. Moreover, the results also show that there are not only differences between classes of approaches (dictionary and machine learning) but also between individual approaches. Careful pretesting and comparison of different algorithms and dictionaries are thus important.

Fourth, as I have found in my analysis, there are considerable differences in terms of performance by tonality, particularly the ‘neutral’ category suffers from low performance metrics. Similar problems have been reported elsewhere (e.g., Boukes et al., 2020). In the case of the news articles examined, a reason for why this might be more pronounced could be attributed to the coding regime. While several approaches examined would classify an article that contains equal amounts of negative and positive sentiment as ‘neutral’, the aim in Mändli et al. (2023) was to capture any criticism that was voiced regarding environmental issues of companies. Consequently, an article containing equal amounts of negative and positive sentiment was classified as ‘negative’ in the hand-coding process. Results show that even when passing on this coding regime to ChatGPT, the ‘neutral’ category is plagued by lower accuracy. Depending on the research question at hand and the importance of each sentiment category, some approaches might be better suited than others.

Fifth, my results support that manual coding should still be considered an essential part of the research process of sentiment analysis (Boukes et al., 2020) in two ways: For the training

of machine learning and fine-tuning of large language models, a meticulously coded training and validation subset of the data ensures higher classification accuracy. For the use of dictionary approaches and models that do not rely on finetuning such as ChatGPT, validation of performance relies on accurately pre-coded data as well. While it might be tempting to use ChatGPT or similar models that do not require training as this allows to save time, the results should nevertheless be examined and validated. Further, manually coding at least part of the data allows researchers to develop a deeper understanding of the subject of interest as well as to validate whether the data they are examining is really what they are intending to investigate. For example, requesting a sample of news articles on environmental issues and the company 'Apple' from a database might actually contain articles about environmental issues related to pesticide use on apple farms.

Importantly, if Krippendorff's alpha is considered as a benchmark criterion, there is no approach in my examination that reaches a recommended level of alpha above .80 (Krippendorff, 2004). Yet, as I show in my supplemental analysis, where I used the outcomes from the automated coding to conduct an actual analysis, the output of machine learning classification and even dictionary classification might be sufficient to draw certain conclusions, depending on the research question. For example, if the goal is to measure average media tonality over a longer period of time, the performance of the approaches I examined might be sufficient, as individual misclassifications might not matter as much on an aggregate level.

These results contribute to management research and in particular applications of sentiment analysis in the following ways: First, they show that for longer texts and more complex subjects, such as environmental issues, manual coding is most likely superior to automated approaches commonly employed in management. This is an important insight for literatures that rely on media sentiment, for example the legitimacy (e.g., Haack et al., 2012; Pollock & Rindova, 2003), celebrity (e.g., Pfarrer et al., 2010; Rindova et al., 2006), or

reputation literatures (e.g., Bundy et al., 2021; Wartick, 1992). Second, when automated approaches are used, a consideration of several approaches, including the most recent ones proposed in information science, as well as manual validation are important. Additionally, the reporting of these validation measures should be considered standard in research papers. Third, the results indicate that performance levels of more novel approaches such as ChatGPT or BERT might approach desirable performance levels, yet require less effort than many approaches thus far. Nevertheless, they still require manually coded data for fine-tuning and validation. It is therefore essential that researchers are able to conduct manual coding reliably, as this approach is plagued by different shortcomings as well (Riffe et al., 2006). Vice-versa, this also implies that papers solely based on manual coding are likely to benefit from quality underlying data.

### **Limitations**

I limited my analysis to examine the most popular dictionary and machine learning approaches currently used in management literature and tested them on a dataset of news articles. This presents several restrictions that demand consideration: First, there might be approaches that perform substantially better that I have not identified through my literature analysis. I have included two state of the art large language models, yet still, there is for example an emerging literature on applications of deep learning and neural networks for sentiment analysis (e.g., Tul et al., 2017; Yadav & Vishwakarma, 2020). Further work could integrate more of such recently proposed methods for sentiment analysis from other disciplines and examine their performance in the context of typical tasks for management science.

Second, another limitation of my analysis is that I examined the large language models only on a subset of the news articles, randomly drawn from the half of the sample with the smaller article length. This was due to the so-called sequence length limitations (the maximum length of a news article passed on to the algorithm). Drawing a subsample limits the ability to



compare results with other approaches in my analysis. On the other hand, performance might also be influenced by the length of the news articles, smaller articles might for example contain less noisy information. Possible solutions to overcome the issue of sequence length limitations, are to summarize input data using large language models, only using portions of the text, or passing on different parts to multiple instances of the algorithm simultaneously (e.g., Grail et al., 2021; Pascual et al., 2021; Zaheer et al., 2020). Whether either of these techniques work well for larger news articles' sentiment analysis remains an additional topic of investigation.

Third, the approaches I examined might have performed differently on other datasets, for example dictionary approaches such as AFINN tend to perform well on shorter Twitter messages (Ribeiro et al., 2016). Adequate performance of sentiment classification of short texts such as social media posts and product reviews is documented well (e.g., Dos Santos & Gatti, 2014; Kiritchenko et al., 2014), some authors found that both types of approaches – dictionary and machine learning – perform equally well on such short texts (e.g., Dhaoui et al., 2017). Yet my results show that for longer and potentially less homogeneous texts, automated sentiment classification becomes more challenging and might demand more sophisticated techniques. In particular, the performance superiority of machine learning over dictionary approaches I observe points towards higher diversity of narration, topics and tones used in those texts. Another aspect that might contribute to this is that dictionary-based approaches are using words from the specific domains they were developed for and might not be adequate for environmental issues in a business context. Future research might thus compare automated approaches' performance on different types of longer texts, such as for example sections of annual reports, transcripts of CEO speeches, or analyst reports, to understand what text specific elements make sentiment classification more challenging, in order to propose solutions to overcome them.

Fourth, the manual coding scheme and its specificities regarding tonality categories in the dataset I used additionally raised the bar for automated approaches, in particular the

dictionary ones. As detailed earlier, the aim of the manual coding process was to capture criticism, which is why an article was coded as ‘negative’ by the coders once there was at least one element of criticism against the company and its environmental issues. Machine learning algorithms are theoretically able to learn any type of classification scheme, what matters is the quality and amount of the training data. This is most likely one of the reasons why machine learning and large language models outperformed dictionary approaches in my analysis. To further investigate this, an approach would be to capture how performance of automated approaches varies with the level of complexity of the manual coding scheme. It would also be interesting to understand whether more sophisticated coding schemes for manual coders require more training data for algorithms.

A final potential limitation to the performance levels of the approaches I examined might be that the manual coders did not agree perfectly. As mentioned earlier, Krippendorff’s alpha between the three coders is reported at .82. It is thus questionable, also from a statistical point of view, whether any type of algorithm might be able to reach levels of this metric above the intercoder reliability of the human coders. First, using the hand-coded dataset for training, will introduce these individual disagreements of coders into the model. Second, as all approaches are evaluated against the hand-coded data that most likely contains a minor proportion of articles where coders disagree. Randomization of training and testing data should be able to deal with the prior issue, as this ensures that news articles coded by each of the coders should be equally likely to be drawn into the training dataset. The latter concern is more complex to deal with. It is theoretically possible, yet unlikely, that currently an algorithm is able to learn the preferences of each individual coder, in this case a precision above the one reported between human coders should be possible. On the other hand, it might be that some algorithms are better at capturing the true tonality of the articles than human coders. In this case, reported metrics could be higher or lower, depending on what kinds of issues are affecting

the automated and hand-coding approaches. To investigate this, a dataset that contains a ‘ground of truth’ different from manual coding (for example customer reviews that contain star ratings) could be examined. It would also be useful to investigate whether different levels of agreement between coders limit the reliability of automated approaches.

### **Conclusion**

This paper underscores that careful manual coding of data remains an important step when conducting sentiment analysis (Boukes et al., 2020), for training and validation of even the most sophisticated models. Given the results of my analysis, several automated approaches might yield results that are sufficiently close to manual coding, using them would result in substantial savings in terms of time and cost. On the other hand, my analysis also shows that the data used in original article (Mändli et al., 2023) which is based entirely on hand coding, cannot be closely replicated with any of the approaches examined.

Advances in machine learning and artificial intelligence make it more likely that different and better solutions to such problems such as sentiment analysis might become popular in the near future (Cambria et al., 2013). For example, developments of large language models (e.g., ChatGPT or BERT) have been shown to perform well at sentiment analysis tasks (e.g., Zhong et al., 2023), a finding that my results also support. Most likely, such tools will be regularly used in management research in the future as well, apart from cost and time considerations, higher accuracy would permit their use also from a research perspective (Xu et al., 2021).

Even if these approaches might go beyond current limitations in terms of performance, sentiment of a given piece of text ultimately is not objectively determined by a machine, but a subjective evaluation of each reader. Similar as to microlevel legitimacy (Bitektine & Haack, 2015), it represents a judgment made at the individual level of the reader. Crowd coding, where a larger group of individuals determine sentiment of given piece of text, might thus present an

additional promising avenue (van Atteveldt et al., 2021). A potential consequence of this could be that instead of binary or trinary evaluations, researchers would employ a distribution of sentiment on a continuum. Coding of text sentiment would probably become costlier (more coders assessing the same piece of text), but not necessarily more time consuming, as platforms would allow to recruit large numbers of coders and to allocate smaller proportions of the data to each coder. On the other hand, quality of coding might suffer if coders are untrained or try to earn as much as possible by coding the texts quicker. Finally, ChatGPT and SentiStrenght both classify this last paragraph as ‘neutral’, whether the reader agrees depends on his or her individual assessment.

## References

- Ahmad, M., Aftab, S., Salman, M., & Hameed, N. (2018). Sentiment analysis using SVM: A systematic literature review. *International Journal of Advanced Computer Science and Applications*, 9(2). <https://doi.org/10.14569/IJACSA.2018.090226>
- Aqel, D., & Vadera, S. (2010). A framework for employee appraisals based on sentiment analysis. *Proceedings of the 1<sup>st</sup> International Conference on Intelligent Semantic Web-Services and Applications*, 1–6. <https://doi.org/10.1145/1874590.1874598>
- Balahur, A., Steinberger, R., Kabadjov, M., & Zavarella, V. (2013). Sentiment analysis in the news. *ArXiv Preprint ArXiv:1309.6202*. <https://doi.org/10.48550/arXiv.1309.6202>
- Barberá, P., Boydston, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19–42. <https://doi.org/10.1017/pan.2020.8>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Bergert, A. L. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), 39–71. <https://dl.acm.org/doi/abs/10.5555/234285.234289>
- Bettinazzi, E. L. M., Jacqueminet, A., Neumann, K., & Snoeren, P. (2023). Media coverage of firms in the presence of multiple signals: A configurational approach. *Academy of Management Journal*, amj.2020.1791. <https://doi.org/10.5465/amj.2020.1791>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1<sup>st</sup> ed). O'Reilly.

- Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226, 107134. <https://doi.org/10.1016/j.knosys.2021.107134>
- Bitektine, A., & Haack, P. (2015). The “Macro” and the “Micro” of Legitimacy: Toward a Multilevel Theory of the Legitimacy Process. *Academy of Management Review*, 40(1), 49–75. <https://doi.org/10.5465/amr.2013.0318>
- Bitektine, A., Hill, K., Song, F., & Vandenberghe, C. (2020). Organizational legitimacy, reputation, and status: Insights from micro-level measurement. *Academy of Management Discoveries*, 6(1), 107–136. <https://doi.org/10.5465/amd.2017.0007>
- Bouazizi, M., & Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. 2016 IEEE International Conference on Communications (ICC), 1–6. <https://doi.org/10.1109/ICC.2016.7511392>
- Boukes, M., van de Velde, B., Araujo, T., & Vliegthart, R. (2020). What’s the tone? Easy doesn’t do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2), 83–104. <https://doi.org/10.1080/19312458.2019.1671966>
- Bradley, M. M., & Lang, P. J. (1999). Affective norms for english words (ANEW): Instruction manual and affective ratings. 49.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://dl.acm.org/doi/abs/10.5555/3495724.3495883>

- Bundy, J., Iqbal, F., & Pfarrer, M. D. (2021). Reputations in flux: How a firm defends its multiple reputations in response to different violations. *Strategic Management Journal*, 42(6), 1109–1138. <https://doi.org/10.1002/smj.3276>
- Butticè, V., Colombo, M. G., & Wright, M. (2017). Serial Crowdfunding, Social Capital, and Project Success. *Entrepreneurship Theory and Practice*, 41(2), 183–207. <https://doi.org/10.1111/etap.12271>
- Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013). Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2), 12–14. <https://doi.org/10.1109/MIS.2013.45>
- Carroll, C. E., & Deephouse, D. L. (2014). The foundations of a theory explaining organizational news: The VT4 Framework of Organizational News Content and five levels of influence on its production. In *Organizations and the Media* (pp. 81–95). Routledge. <https://doi.org/10.4324/9780203068052>
- Castelló, I., Etter, M., & Årup Nielsen, F. (2016). Strategies of Legitimacy Through Social Media: The Networked Strategy: Strategies of Legitimacy Through Social Media. *Journal of Management Studies*, 53(3), 402–432. <https://doi.org/10.1111/joms.12145>
- Chan, C., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using ‘off-the-shelf’ dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, 3(1), 1–27. <https://doi.org/10.5117/CCR2021.1.001.CHAN>
- Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2016). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? (arXiv:1511.06348). arXiv. <http://arxiv.org/abs/1511.06348>

- Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, 40(11), 1705–1732. <https://doi.org/10.1002/smj.3067>
- Courtney, C., Dutta, S., & Li, Y. (2017). Resolving Information Asymmetry: Signaling, Endorsement, and Crowdfunding Success. *Entrepreneurship Theory and Practice*, 41(2), 265–290. <https://doi.org/10.1111/etap.12267>
- Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press. <https://dl.acm.org/doi/abs/10.5555/345662>
- Curran, J., Esser, F., Hallin, D. C., Hayashi, K., & Lee, C.-C. (2017). International news and global integration: A five-nation reappraisal. *Journalism Studies*, 18(2), 118–134. <https://doi.org/10.1080/1461670X.2015.1050056>
- Deepphouse, D. L., & Carter, S. M. (2005). An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, 42(2), 329–360. <https://doi.org/10.1111/j.1467-6486.2005.00499.x>
- Desmedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13, 2063–2067. <https://www.jmlr.org/papers/v13/desmedt12a.html>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). arXiv. <http://arxiv.org/abs/1810.04805>
- Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6), 480–488. <https://doi.org/10.1108/JCM-03-2017-2141>
- Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. 69–78. <https://aclanthology.org/C14-1008>



- Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161, 707–714. <https://doi.org/10.1016/j.procs.2019.11.174>
- Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods*, 10(1), 5–34. <https://doi.org/10.1177/1094428106289252>
- Etter, M., Colleoni, E., Illia, L., Meggiorin, K., & D'Eugenio, A. (2018). Measuring Organizational Legitimacy in Social Media: Assessing Citizens' Judgments With Sentiment Analysis. *Business & Society*, 57(1), 60–97. <https://doi.org/10.1177/0007650316683926>
- Flach, P. A. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511973000>
- Fombrun, C., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of Management Journal*, 33, 233–258. <https://doi.org/10.5465/256324>
- Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3), 437–442. <https://doi.org/10.1109/IC3.2014.6897213>
- Gelbard, R., Ramon-Gonen, R., Carmeli, A., Bittmann, R. M., & Talyansky, R. (2018). Sentiment analysis in organizational work: Towards an ontology of people analytics. *Expert Systems*, 35(5), e12289. <https://doi.org/10.1111/exsy.12289>
- Grail, Q., Perez, J., & Gaussier, E. (2021). Globalizing BERT-based Transformer Architectures for Long Document Summarization. *Proceedings of the 16<sup>th</sup> Conference of the European*

- Chapter of the Association for Computational Linguistics: Main Volume, 1792–1810.  
<https://doi.org/10.18653/v1/2021.eacl-main.154>
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297.  
<https://doi.org/10.1093/pan/mps028>
- Guzman, E., & Bruegge, B. (2013). Towards emotional awareness in software development teams. *Proceedings of the 2013 9<sup>th</sup> Joint Meeting on Foundations of Software Engineering*, 671–674. <https://doi.org/10.1145/2491411.2494578>
- Haack, P., Schoeneborn, D., & Wickert, C. (2012). Talking the talk, moral entrapment, creeping commitment? Exploring narrative dynamics in corporate responsibility standardization. *Organization Studies*, 33(5–6), 815–845. <https://doi.org/10.1177/0170840612443630>
- Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity*, 51(6), 2623–2646. <https://doi.org/10.1007/s11135-016-0412-4>
- He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing (arXiv:2111.09543). arXiv. <http://arxiv.org/abs/2111.09543>
- He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention (arXiv:2006.03654). arXiv. <http://arxiv.org/abs/2006.03654>
- Hubbard, T. D., Pollock, T. G., Pfarrer, M. D., & Rindova, V. P. (2018). Safe bets or hot hands? How status and celebrity influence strategic alliance formations by newly public firms. *Academy of Management Journal*, 61(5), 1976–1999.  
<https://doi.org/10.5465/amj.2016.0438>

- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. *Proceedings from the 11<sup>th</sup> Conference on Uncertainty in Artificial Intelligence*. <https://doi.org/10.48550/arXiv.1302.4964>
- Kheiri, K., & Karimi, H. (2023). SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning (arXiv:2307.10234). arXiv. <http://arxiv.org/abs/2307.10234>
- Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50, 723–762. <https://doi.org/10.1613/jair.4272>
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Classification for Organizational Researchers: A Tutorial. *Organizational Research Methods*, 21(3), 766–799. <https://doi.org/10.1177/1094428117719322>
- Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190. <https://doi.org/10.1007/s10462-007-9052-3>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2<sup>nd</sup> ed). Sage. <https://lccn.loc.gov/2003014200>
- Kuncheva, L. I., & Rodríguez, J. J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, 38(2), 259–275. <https://doi.org/10.1007/s10115-012-0586-6>
- Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. *Proceedings of the 18<sup>th</sup> Annual International ACM SIGIR Conference on Research and*

<https://doi.org/10.1145/215206.215366>

- Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5. <https://dl.acm.org/doi/abs/10.5555/1005332.1005345>
- Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2022). Pretrained language models for text generation: A survey (arXiv:2201.05273). arXiv. <http://arxiv.org/abs/2201.05273>
- Li-Ping Jing, Hou-Kuan Huang, & Hong-Bo Shi. (2002). Improved feature selection approach TFIDF in text mining. *Proceedings. International Conference on Machine Learning and Cybernetics*, 2, 944–946. <https://doi.org/10.1109/ICMLC.2002.1174522>
- Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications*, 80, 323–339. <https://doi.org/10.1016/j.eswa.2017.03.042>
- Loria, S. (2020). TextBlob documentation release 0.16.0. <https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf>
- Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. *Journal of Behavioral Finance*, 16(1), 1–11. <https://doi.org/10.1080/15427560.2015.1000335>
- Mändli, F. B., Amer, E., & Bonardi, J.-P. (2023). Private politics, dynamics of negative media exposure, and firm strategy. [Working Paper].
- Micu, A., Micu, A. E., Geru, M., & Lixandriou, R. C. (2017). Analyzing user sentiment in social media: Implications for online marketing strategy. *Psychology & Marketing*, 34(12), 1094–1100. <https://doi.org/10.1002/mar.21049>
- Moniz, A., & De Jong, F. (2014). Sentiment analysis and the impact of employee satisfaction on firm earnings. In M. De Rijke, T. Kenter, A. P. De Vries, C. Zhai, F. De Jong, K. Radinsky, & K. Hofmann (Eds.), *Advances in Information Retrieval* (Vol. 8416, pp.

- 519–527). Springer International Publishing. [https://doi.org/10.1007/978-3-319-06028-6\\_51](https://doi.org/10.1007/978-3-319-06028-6_51)
- Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, 36(2), 214–226. <https://doi.org/10.1080/10584609.2018.1517843>
- Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2), 101–111. <https://doi.org/10.1109/TAFFC.2014.2317187>
- Nauhaus, S., Luger, J., & Raisch, S. (2021). Strategic Decision Making in the Digital Age: Expert Sentiment and Corporate Capital Allocation. *Journal of Management Studies*, 58(7), 1933–1961. <https://doi.org/10.1111/joms.12742>
- Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs (arXiv:1103.2903). arXiv. <http://arxiv.org/abs/1103.2903>
- Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1, 61–67. <http://www.kamalnigam.com/papers/maxent-ijcaiws99.pdf>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques (arXiv:cs/0205070). arXiv. <http://arxiv.org/abs/cs/0205070>
- Pascual, D., Luck, S., & Wattenhofer, R. (2021). Towards BERT-based automatic ICD coding: Limitations and opportunities. ArXiv Preprint ArXiv:2104.06709. <https://doi.org/10.48550/arXiv.2104.06709>
- Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, 31(6), 539–548. [https://doi.org/10.1016/0005-7967\(93\)90105-4](https://doi.org/10.1016/0005-7967(93)90105-4)

- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2015). Linguistic inquiry and word count: LIWC2015. Pennebaker Conglomerates.
- Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, 53(5), 1131–1152. <https://doi.org/10.5465/amj.2010.54533222>
- Piazza, A., & Perretti, F. (2015). Categorical stigma and firm disengagement: Nuclear power generation in the United States, 1970–2000. *Organization Science*, 26(3), 724–742. <https://doi.org/10.1287/orsc.2014.0964>
- Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, 53(1), 122–150. <https://doi.org/10.1016/j.ipm.2016.07.001>
- Pollock, T. G., Lashley, K., Rindova, V. P., & Han, J.-H. (2019). Which of these things are not like the others? Comparing the rational, emotional, and moral aspects of reputation, status, celebrity, and stigma. *Academy of Management Annals*, 13(2), 444–478. <https://doi.org/10.5465/annals.2017.0086>
- Pollock, T. G., & Rindova, V. P. (2003). Media legitimation effects in the market for initial public offerings. *Academy of Management Journal*, 46(5), 631–642. <https://doi.org/10.5465/30040654>
- Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing*, 12(2), 146–163. <https://doi.org/10.1108/JRIM-05-2017-0030>
- Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. <http://dx.doi.org/10.1109/ASONAM.2012.30>

- Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1), 23. <https://doi.org/10.1140/epjds/s13688-016-0085-1>
- Riffe, D., Lacy, S., Fico, F., Riffe, D., Lacy, S., & Fico, F. G. (2006). *Analyzing media messages* (0 ed.). Routledge. <https://doi.org/10.4324/9781410613424>
- Rindova, V. P., Pollock, T. G., & Hayward, M. L. A. (2006). Celebrity firms: The social construction of market popularity. *Academy of Management Review*, 31(1), 50–71. <https://doi.org/10.5465/amr.2006.19379624>
- Ritter, T., & Pedersen, C. L. (2020). Digitization capability and the digitalization of business models in business-to-business firms: Past, present, and future. *Industrial Marketing Management*, 86, 180–190. <https://doi.org/10.1016/j.indmarman.2019.11.019>
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. [https://doi.org/10.1162/tacl\\_a\\_00349](https://doi.org/10.1162/tacl_a_00349)
- Samal, B., Behera, A. K., & Panda, M. (2017). Performance analysis of supervised machine learning techniques for sentiment analysis. *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 128–133. <https://doi.org/10.1109/SSPS.2017.8071579>
- Samuel, J., Ali, G. G. Md. N., Rahman, Md. M., Esawi, E., & Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information*, 11(6), 314. <https://doi.org/10.3390/info11060314>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>

- Scharkow, M. (2017). Content analysis, automatic. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods* (1<sup>st</sup> ed., pp. 1–14). Wiley. <https://doi.org/10.1002/9781118901731.iecrm0043>
- Shipilov, A. V., Greve, H. R., & Rowley, T. J. (2019). Is all publicity good publicity? The impact of direct and indirect media pressure on the adoption of governance practices. *Strategic Management Journal*, 40(9), 1368–1393. <https://doi.org/10.1002/smj.3030>
- Stone, P. J., & Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, 241–256. <https://doi.org/10.1145/1461551.1461583>
- Swain, A. K., & Cao, R. Q. (2019). Using sentiment analysis to improve supply chain intelligence. *Information Systems Frontiers*, 21(2), 469–484. <https://doi.org/10.1007/s10796-017-9762-2>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- Thelwall, M. (2017). The heart and soul of the web? Sentiment strength detection in the social web with sentistrength. In J. A. Holyst (Ed.), *Cyberemotions: Collective Emotions in Cyberspace* (pp. 119–134). Springer International Publishing. [https://doi.org/10.1007/978-3-319-43639-5\\_7](https://doi.org/10.1007/978-3-319-43639-5_7)
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163–173. <https://doi.org/10.1002/asi.21662>
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558. <https://doi.org/10.1002/asi.21416>



- Tul, Q., Ali, M., Riaz, A., Noureen, A., Kamranz, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: A review. *International Journal of Advanced Computer Science and Applications*, 8(6).  
<https://doi.org/10.14569/IJACSA.2017.080657>
- Uhl, M. W. (2014). Reuters sentiment and stock returns. *Journal of Behavioral Finance*, 15(4), 287–298. <https://doi.org/10.1080/15427560.2014.967852>
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, 14(11).  
<https://doi.org/10.1371/journal.pone.0224365>
- van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2), 121–140.  
<https://doi.org/10.1080/19312458.2020.1869198>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need (arXiv:1706.03762). arXiv.  
<http://arxiv.org/abs/1706.03762>
- Wang, M., & Hu, F. (2021). The application of NLTK library for Python natural language processing in corpus research. *Theory and Practice in Language Studies*, 11(9), 1041–1049. <https://doi.org/10.17507/tpls.1109.09>
- Wartick, S. L. (1992). The relationship between intense media exposure and change in corporate reputation. *Business & Society*, 31(1), 33–49.  
<https://doi.org/10.1177/000765039203100104>
- Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting latent moral information

- from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, 12(2–3), 119–139. <https://doi.org/10.1080/19312458.2018.1447656>
- Wei, J., Ouyang, Z., & Chen, H. A. (2017). Well known or well liked? The effects of corporate reputation on firm value at the onset of a corporate crisis: the effects of corporate reputation on firm value. *Strategic Management Journal*, 38(10), 2103–2120. <https://doi.org/10.1002/smj.2639>
- Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., ... Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4), 100179. <https://doi.org/10.1016/j.xinn.2021.100179>
- Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6), 4335–4385. <https://doi.org/10.1007/s10462-019-09794-5>
- Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2), 205–231. <https://doi.org/10.1080/10584609.2012.671234>
- Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5. <https://dl.acm.org/doi/abs/10.5555/1005332.1044700>
- Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., & Yang, L. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33, 17283–17297.
- Zhang, H., Gan, W., & Jiang, B. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. 2014 11<sup>th</sup> Web Information System and Application Conference, 262–265. <https://doi.org/10.1109/WISA.2014.55>

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT (arXiv:2302.10198). arXiv. <http://arxiv.org/abs/2302.10198>

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2), 1–6. <https://doi.org/10.1177/2053951714559253>

## Conclusion

In this last section, I briefly outline the main contributions and limitations of the chapter in the thesis. Finally, I tie together the individual results of the chapters with regards to their interrelatedness.

Chapter 1 contributes to the understanding of the relationship of firms and the media. We provide evidence that negative media exposure on environmental issues has a dynamic effect, therefore it increases the subsequent likelihood of additional exposure of this type for firms. As negative media exposure dynamically accumulates over time, firms might be in danger of ending up in a spiral of negative reporting. However, we also show that these spirals can be broken by positive media reports and press releases, as both have a protective effect. This indicates that firms can not only use impression management tactics in a reactive manner, but strategically communicate – as in our study for example by emitting press releases – to diminish negative bias in news reports.

Dynamic effects in media exposure appear to be a complex phenomenon. A valuable approach towards gaining additional understanding could be to investigate potential dynamic effects across varying media types, including print, TV, and social media. Another approach would be to explore the implications of these dynamic effects for the consequences of media exposure. For example, do negative dynamics additionally pressure firms to improve on environmental dimensions? Finally, it would also be essential to understand the repercussions of dynamic media effects for firm reputation, legitimacy, and other social evaluations' constructs.

The contributions of chapter 2 are not tied to a specific literature, but more broadly to quantitative research in management that employs control variables. Our discussion of both perspectives on the matter of using few or many control variables and the simulations show that empirical rules related to control variable selection are misleading and that the many or few control variables debate is futile. Instead, we propose an integrated approach that combines

the merits of both perspectives. We also recommend that researchers focus on existing theory to identify relevant controls, and use causal graphs to detect endogenous controls. Quantitative research in management science should aim at delivering causal claims, correct model specifications and inclusion of relevant controls are essential steps on this path.

While simulations are a fitting tool to answer our research questions, we are unable to quantify actual biases that applications of either perspective might have caused. Further studies would be required to understand whether studies that did or did not adhere to a specific approach were more or less likely to use fewer or more control variables and whether these studies are plagued by omitted variable bias and/or bias caused by inclusion of endogenous controls. More work is also needed to ensure that techniques such as causal graphs become more widely known and more frequently applied in research practice.

The findings from chapter 3 contribute to management research using sentiment analysis. My investigation puts forth that frequently used automated approaches for sentiment analysis in the field are unable to replicate manual coding with necessary precision when applied to news articles. However, I also show that two more recent state of the art large language models deliver more promising results than commonly used approaches in the field. I consequently propose that researchers test different coding procedures, both manual and automated, and include in this test more novel approaches such as the large language models I examined, before settling on a methods choice.

Given that I considered the most frequently used approaches in management science in this study, further research is needed to investigate whether other approaches, that were outside the scope of my analysis, might be more accurate. Moreover, while news articles are one type of data investigated in management research, an understanding of the performance of automated approaches across various kinds of data commonly examined in management would be important. Finally, as in my investigation, the “ground truth” in sentiment analysis is usually

manual coding, often times coded by researchers involved in the project. Investigations into the limitations of using manual coding as a baseline and potential alternative measures that could be used as a reference should present additional valuable steps forward.

The results of chapters 2 and 3, beyond their individual contributions, are in support of the empirical choices made in chapter 1, lending credibility to its results as well as the theoretical and practical implications. A main outcome of the critiques and simulations in chapter 2 is that control variable selection should be driven by existing theory and to abandon empirical rules. This supports our approach in chapter 1 to identify drivers of media exposure in existing works and to control for them, irrespective of how many or few. For example, if we would have chosen the frugal approach, we would have dropped the control variables *Assets* and *Profitability*, as they are not significantly correlated with the dependent variables. As shown in the simulations of chapter 2, this would have possibly caused an omitted variable problem and biased our estimates of interest, as both, *Assets* and *Profitability* are significantly correlated with other variables in the model. Chapter 2 also shows that it would have probably been a bad idea to abstain from proxies from our analysis in chapter 1 (most control variables in this chapter are proxies), as this would have most likely resulted in additional bias.

Second, our choice of using manual coding for the environmental news articles is supported by the results of the analysis of automated sentiment approaches in chapter 3. It shows that if we had used automated approaches to determine the tonality of the news articles, we would have obtained different tonality patterns conditional on the type of approach used. Most likely, this would have also changed the conclusions drawn from the data. One potential reason why automated approaches struggle in this case, is because the news articles contain additional information that is unrelated to the firm's environmental issues, but is taken into account by any algorithm. Moreover, in the manual coding procedure, we were aiming to detect even a minimal amount of criticism, something that is difficult to replicate using basic automated

approaches such as a dictionary, but might be attained using more capable algorithms such as ChatGPT. Finally, the manual coding approach additionally provided the authors with a detailed and systematic understanding of the underlying news articles (each article had been read by at least one person, two of the three coders are authors), knowledge that would not have been obtained if we had used an automated approach.