Bots on Social Media

The Past, Present and Future

Rossi, Sippo

Document Version
Final published version

DOI:
10.22439/phd.19.2024

Publication date:
2024

COPENHAGEN BUSINESS SCHOOL

Solbjerg Plads 3
DK-2000 Frederiksberg
Danmark

www.cbs.dk

ISSN 0906-6934

Print ISBN: 978-87-7568-269-0
Online ISBN: 978-87-7568-270-6

BOTS ON SOCIAL MEDIA

PhD Series 19-2024

CBS

CBS PhD School
Department of Digitalization

PhD Series 19.2024

SIPPO ROSSI

BOTS ON SOCIAL MEDIA

The Past, Present and Future

CBS

Bots on Social Media

The Past, Present and Future

Sippo Rossi

Department of Digitalization

Supervisor(s): Raghava Rao Mukkamala, Jason Bennett Thatcher

CBS PhD School
Copenhagen Business School

Sippo Rossi
Bots on Social Media
The Past, Present and Future

First edition 2024
Ph.D. Series 19.2024

ISSN 0906-6934

Print ISBN: 978-87-7568-269-0
Online ISBN: 978-87-7568-270-6

DOI: https://doi.org/10.22439/phd.19.2024

Foreword

Three and a half years have passed quickly. This time has shaped my academic identity and laid the groundwork for the beginning of my career as a researcher. Being supervised by a computer scientist and an information systems scholar has resulted in what I hope will be seen as an interesting interdisciplinary dissertation that strikes a balance between researching technology and studying the effects of technology on society. This dual identity is reflected in the dissertation's papers, which fall somewhere between information systems and computer science. Despite the challenges this introduced, ultimately, I am extremely satisfied with how the PhD project turned out and would like to thank all those who contributed to its success.

First and foremost, I would like to thank my supervisors, Raghava Rao Mukkamala and Jason Thatcher. My primary supervisor, Raghava Rao Mukkamala, entrusted me with an incredible amount of freedom to pursue the type of research that I found interesting and was always available to help and guide me through the challenges of completing a PhD. I aspire to become a researcher who is as calm, supportive, and optimistic as Raghava. I also want to praise my secondary supervisor, Jason Thatcher, whose guidance and support have been invaluable throughout my PhD. Jason pushed me just the right amount to aim higher, and I believe his efforts will help me stay competitive in the academic job market.

The research conducted for this PhD was truly a collaborative effort, and I would like to thank my research assistant, Odd Harald Auglend, and co-authors Matti Rossi, Bikesh Raj Upreti, Yong Liu, Peter Ractham, and Yogesh K Dwivedi, who contributed to the articles presented in this dissertation. Furthermore, I want to express my gratitude for the general academic mentoring that I have received during my PhD studies from Matti Rossi and Nenad Jukić.

I also want to thank the community at the CBS Department of Digitalization—in particular, Stig Nyman, Timothy Charlton-Czaplicki, Lucas Göbeler, Maylis Saigot, and Irfan Kanat. Besides the DIGI community, I am also thankful for Matti Nelimarkka and Silvia Masiero, who hosted my research stays at the University of Helsinki and the University of Oslo, respectively.

Finally, I thank my wife, Katya, for all her love and support throughout these three and a half years, and our cat, Naomi, who would routinely sit in my lap or next to my screen purring or meowing during the long days of writing. Moreover, I thank my parents, Tuuli and Matti, and siblings, Saana and Samu, for their support.

Abstract

Bots, and their more sophisticated variant, social bots, have become a ubiquitous part of social media. From a technological standpoint, the modern social bot is a remarkable achievement, having survived countless attempts at detection and removal by social networking sites through the evolution and persistence of those who operate bots. Social bots have been accused of being capable of influencing opinions and even manipulating election results, although recent work has also explored benign and benevolent use cases for social bots. This dissertation uses a variety of methods, from machine learning and network analysis to experiments, to study these different types of social bots.

The dissertation is based on five publications that contribute to our overall understanding of social bots and how to study them. The first two publications represent the early and naive era of social bot research, where the goal was to use machine learning to detect and study bots as manipulators of elections or spreaders of conspiracy theories. The third and fourth publications jump to the modern era of generative AI-powered social bots and focus on the bot detection capabilities of humans rather than machine learning models, an understudied area of social bot research. The fifth and final publication builds on the methods developed in the third paper and proposes a more generalized approach to using foundation models and generative AI in experiments to study phenomena such as social bots. Overall, this dissertation describes the history and evolution of both social bots and the field of study itself and concludes with an epilogue that speculates on the future of bot research in an era where Twitter is no longer a viable data source.

Abstrakt

Bots, og deres mere sofistikerede variant, sociale bots, er blevet en allestedsnærværende del af sociale medier. Fra et teknologisk perspektiv er den moderne sociale bot en bemærkelsesværdig bedrift, som har overlevet utallige forsøg på opdagelse og fjernelse fra sociale netværkssider gennem evolutionen og vedholdenheden hos dem, der opererer bots. Sociale bots er blevet beskyldt for at være i stand til at påvirke meninger og endda manipulere valgresultater, selvom nyere forskning dog også har udforsket godartede og velgørende anvendelsesmuligheder for sociale bots. Denne afhandling bruger en række metoder, fra maskinlæring og netværksanalyse til eksperimenter, til at studere disse forskellige typer af sociale bots.

Denne afhandling bygger på fem artikler, som bidrager til vores overordnede forståelse af sociale bots og hvordan man studerer dem. De første to artikler repræsenterer den tidlige og naive æra af forskning i sociale bots, hvor målet var at bruge maskinlæring til at detektere og studere bots, som forsøgte at manipulere valgresultater eller sprede konspirationsteorier. Tredje og fjerde artikel omhandler den moderne æra af generativ AI-drevne sociale bots og fokuserer på menneskers, ikke maskinlæringsmodellers, evne til at opdage bots, hvilket er et underbelyst område inden for forskning i sociale bots. Den femte og sidste artikel bygger på metoderne udviklet i den tredje artikel og foreslår en mere generaliseret tilgang til brug af fundamentale modeller og generativ AI i eksperimenter til at studere fænomener såsom sociale bots. Overordnet beskriver denne

afhandling historien og evolutionen af både sociale bots og studiefeltet selv og afslutter med en epilog, der spekulerer om fremtiden for bot-forskning i en æra, hvor Twitter ikke længere er en datakilde.

Table of Contents

1 INTRODUCTION 9

2 A BRIEF HISTORY OF SOCIAL BOT RESEARCH 15

3 SUMMARY OF ARTICLES AND CONTRIBUTIONS 25

4 CONCLUSION 41

5 EPILOGUE 45

6 REFERENCE LIST 49

7 APPENDIX 57

PAPER I 58

PAPER II 72

PAPER III 87

PAPER IV 103

PAPER V 126

1 Introduction

This chapter first presents the background of social bots and social bot research, providing a context for the topic of the dissertation. This is followed by an explanation of the definition of social bots that is used in this dissertation and the delimitation of which areas of bot research are considered relevant. Then, the research objectives of the PhD project are introduced. Lastly, the structure of the dissertation is outlined.

1.1 Background

Social media is ubiquitous and has become increasingly important over the last 20 years for individuals, organizations, and even state actors as a means of sharing information and communicating with relevant stakeholders. At its peak, it would have been theoretically possible to interact with hundreds of millions of active users on Twitter alone. Unsurprisingly, various entities, from regular users to corporate and government-affiliated accounts, have found this ability to reach the masses valuable and are actively using social media to interact, ultimately influencing what happens in the physical world.

This dissertation focuses on one particular entity that has been designed to control discourse and influence users—the entity of the *social bot*. Social bots have been described in various ways and at the height of public and academic interest in the subject, the tone was increasingly menacing. Figure 1 below presents some examples.

"Today's social bots are sophisticated and sometimes menacing. Indeed, their presence can endanger online ecosystems as well as our society." *-Ferrara et al. (2016)*

"They are considered to be capable of massively distributing propaganda in social and online media and their application is even suspected to be partly responsible for recent election results" *-Grimme et al. (2017)*

"Bots increasingly tamper with political elections and economic discussions. Tracing trends in detection strategies and key suggestions on how to win the fight." *-Cresci (2020)*

Figure 1: Examples of descriptions and images of social bots as shown in journals

Social bots, broadly defined as accounts on social networking sites that mimic humans and are controlled by a program (Ferrara et al., 2016), have existed for over a decade (Cresci, 2020a). Recent advances in technology such as machine learning (ML) and natural language processing (NLP) have helped bots evolve from mere crude spammers to sophisticated autonomous agents capable of shaping online discourse (Ferrara et al., 2016). From a technological point of view, the modern social bot is a remarkable achievement, as it is increasingly capable of disseminating information (Salge et al., 2022), influencing public opinion (Ross et al., 2019), and avoiding human detection by blending into crowds of humans on social networking sites (Cresci et al., 2017).

To detect these evolving social bots, researchers have employed increasingly complex detection methods, such as machine learning models that use hundreds of features to determine whether individual accounts (or even groups of accounts) are bots (Cresci, 2020a). Rather than building a bespoke model for each paper, many researchers have also resorted to using services such as the Botometer, a bot detection tool maintained by researchers and accessible through an API (Davis et al., 2016; Sayyadiharikandeh et al., 2020). This cat-and-mouse game between bot developers and researchers trying to detect newer bots has also been described as an "arms race" against social bots (Cresci et al., 2017).

Despite the recent emphasis on advanced and autonomous social bots, traditional or simple bots operated by more elementary scripts to curate content or spam or boost the visibility of content on social media are still very much present on social media (Rossi, 2022). Scholars have been interested in such bots from the early days of social networking sites, with sporadic publications appearing in the early 2010s describing them as "spambots" or "astroturfing bots" that, for example, share links to phishing websites or pretend to support a politician by following them and liking and retweeting their posts (Boshmaf et al., 2011; Chu et al., 2010). The study of bots saw a general explosion of interest after the public became aware of the use of social bots in the 2016 US Presidential Election (Cresci, 2020a). Due to the popularity of studying specifically political bots, most bot research has focused on either proposing new bot detection methods (Cresci, 2020a) or presenting case studies of bots in various settings such as elections (Bessi & Ferrara, 2016; Brachten et al., 2017; Fernquist et al., 2018; Ferrara, 2017). The commonality of these two categories of papers is that they rely predominantly on Twitter data and posit social bots as inherently malicious entities, using descriptions like those shown in the text of Figure 1.

More recent works have started exploring social bots through a broader set of lenses, proposing benevolent use cases for them (Blasiak et al., 2021) and studying their use in more neutral settings such as in the context of sales on LinkedIn (Goldstein & DiResta, 2022). Furthermore, studies have shown that bots can both intentionally and unintentionally have a positive or negative influence when distributing content such as information coming from the site of a natural disaster (Hofeditz et al., 2019). For example, if a social bot distributes content that is truthful, it can help spread valuable information to an audience that may otherwise not have been reached. Conversely, if the same social bot shares a post containing misinformation, it can mislead those who come across it. Thus, social bots can take many roles and are not clearly definable as positive or negative entities in the social media ecosystem.

The recent advancements in the capabilities of foundation models and generative AI (Bommasani et al., 2022) will most likely also have an impact on bots and bot research, as they make producing high-quality and unique content at scale easier. We may soon see an increase in generative AI-infused bots that can leverage the power of large language models such as LLaMA 2 and GPT-4 in writing posts and responses while hiding behind seemingly genuine and unique profile pictures generated with NVIDIA's StyleGAN2. While we know that bot developers are already using deep learning-generated images in bot profiles and ChatGPT to produce tweets for bots, academic research on the matter is still scarce due to the recency of this development and challenges related to the detection of generated text (Bond, 2022; Goldstein & DiResta, 2022; Yang & Menczer, 2023). Furthermore, early experiments have shown that profiles created with these methods can indeed be indistinguishable from genuine profiles and posts in social media feeds (Rossi et al., 2023).

For bot researchers, generative AI influencing bot designs and making them more challenging to detect is not the only pressing concern. Researching social bots has recently become significantly more difficult as a result of the social media landscape changing and becoming less supportive of academic research. With the majority of social bot studies relying on data from Twitter (now X), the future is uncertain, as the company has changed its pricing policies, making access to its API prohibitively expensive for academic users (Kupferschmidt, 2023). X is not alone in its decision to cut off academic access; for example, competing companies such as Meta have even attempted to actively prevent researchers from collecting data through other means such as scraping (Hatmaker, 2021). Overall, social media companies have shifted towards being more closed and hostile to researchers, and many studies, including two of the ones included in this dissertation, are quickly becoming relics of a time when data was easily available.

1.2 Definition of social bots in this dissertation

One of the early challenges in social bot research was defining what exactly counts as a social bot.[1] While there is an ongoing discussion on the exact definition of social bots, the terms bot and social bot are sometimes used interchangeably to describe automated accounts on social networking sites (Gorwa & Guilbeault, 2020; Grimme et al., 2017; Stieglitz et al., 2017). One of the significant ambiguities is whether social bots refer to accounts meant to mimic and deceive humans (Stieglitz et al., 2017) and whether accounts that have disclosed that they are automated or computer managed also count as social bots, despite lacking the element of deception.

To align with what seems to be the consensus in two widely cited review articles (Gorwa & Guilbeault, 2020; Stieglitz et al., 2017), in this dissertation, bots are defined as autonomous accounts or agents operated by a computer program rather than a human. Furthermore, social bots are defined as bots that interact with humans and mimic their behavior. Also, this dissertation is limited to studying bots that operate in publicly accessible social networking sites and have a profile similar to that of humans. In other words, the bot profile must be visually and functionally indistinguishable from other accounts, which is not the case, for example, with customer service bots that operate only within chat functionalities and lack profiles that are similar to human users.

[1] It should be noted that the term “social robot” (bot being an abbreviation of robot) has been used with a much broader definition in the field of robotics even before the existence of social media. Originally, the term social robot meant robots designed to interact with humans and behave in a humanlike manner (Duffy, 2003). However, in contemporary research, “social bot” is used predominantly in the context of bots on social media.

Therefore, chatbots and social bots in enterprise social networks are not considered in this dissertation for two reasons. First, whereas work on bots for commercial purposes primarily considers implications for organizations (Meske & Amojo, 2018; Stieglitz et al., 2018), this dissertation is interested in furthering the discourse on bots that create societal impact (Grimme et al., 2017) by participating in public social networks (Salge & Karahanna, 2018). Second, the closed environment in which chatbots and enterprise social bots operate makes their behavior predictable; in contrast, bots operating in more open environments like X (formerly Twitter) or Facebook often exhibit rather unpredictable behavior (Salge et al., 2021). From this point onward, the above definition will apply to the word social bot unless explicitly defined otherwise.

1.3 Research objectives, gap, and questions

As bot research matured and new areas of interest arose within the field between 2020 and 2023, the research objectives of this PhD project reactively evolved. Ultimately, this dissertation poses three separate research objectives and research questions, each responding to a distinct gap in bot research.

1.3.1 The initial research objective

Initially, the research objective was to develop methodologies for detecting social bots and then applying these methods and documenting the existence of bots and their behavior in new contexts such as Finnish politics and the COVID-19 pandemic. This objective was set during the "arms race" period of social bot research, when both bots and proposed detection methods were becoming increasingly intricate. This resulted in the identification of a research gap—namely that light and relatively simple detection methods were not being studied in academic papers as much as complex and computationally intensive bot detection methods, which had become the norm. Thus, the goal then became to test whether sufficient performance in detection could be achieved with machine learning models that rely on a smaller number of features and the reduced use of the Twitter API. Therefore, for papers I and II of this dissertation, the research question was:

RQ1: How can bots be detected and studied in specific contexts with computationally light approaches?

1.3.2 The second research objective

During the process of writing papers I and II, a new research gap was identified—namely the dearth of studies on the human ability to distinguish social bots from genuine human accounts. Moreover, due to the productization and increased accessibility of generative AI models from 2022 onwards, producing high-quality content and profile images for bot profiles became easier, and evidence began to appear that generative AI was being used by those operating bots (Bond, 2022; Goldstein & DiResta, 2022; Yang & Menczer, 2023). As a result of these developments, the limited previous findings regarding the human ability to qualitatively identify accounts might no longer hold if producing more humanlike bots has become not only possible but easy to do. Thus, pursuant to the proliferation of generative AI, determining whether humans can still identify bots on social media via qualitative assessment has become a timely topic to investigate.

Further, contributing to the need to change the research objective, the previously described developments coincided with Twitter becoming increasingly unreliable as a data source due to the change of management at the company. Shortly after the social networking site Twitter (now X), on which most bot studies had relied for data on bots, became practically infeasible as a data source due to the introduction of much higher costs to use the platform's API. Thus, conducting research with the existing methodologies and contributing to bot detection research that would build upon previous findings became difficult.

To study this newly identified research gap and to address the changes in social bots and data collection, the dissertation took a new approach to studying the social bot phenomenon. Moving the research objective away from developing detection methods, the new focus shifted towards using the latest AI tools for developing bots to produce fake accounts, followed by experiments on the human ability to detect these AI-infused bot profiles. Thus, papers III and IV sought to answer the following research question:

RQ2: Can humans detect AI-generated bot profiles on social media?

1.3.3 The third research objective

To answer RQ2, papers III and IV applied generative AI to produce bot profiles to show to experiment participants. The application of generative AI in these papers was quite innovative, as there were few examples of existing research where generative AI was used to produce content for experiments. As this approach to producing content could be beneficial not only for studies on social bots but for other social media phenomena as well, the fifth paper's research objective was to write a generalized guide for using generative AI as part of existing research methods. This research objective sought to address a gap in the literature—namely that there were few practical guides related to using generative AI, as most similar papers and editorials were focused on discussing the ethics of using generative AI. Thus, the research question for paper V was:

RQ3: How can generative AI be used to augment existing research methods?

The overarching research objective of this PhD project was to increase our understanding of social bots. This objective resulted in the three research questions described above, which the five papers of this PhD project answered by providing evidence of bots operating in different contexts, developing methods for studying social bots, and demonstrating through experiments that humans can no longer reliably detect AI-driven social bots. The final paper of this project used the approach to using generative AI in an experiment developed in papers III and IV to create a more general guide to using generative AI in research, which could be used in future bot studies as well as in other research areas.

1.4 Outline of the dissertation

The remainder of the dissertation is organized as follows. Chapter 2 presents a literature review via an essay entitled "A Brief History of Social Bot Research." This essay summarizes relevant bot literature from the past 15 years to explain how the study of bots has evolved throughout the history of social media, emphasizing seminal papers mainly from computer science as well as publications that have appeared in information systems (IS) journals and conferences. The essay also discusses studies that are similar to papers I–IV of this dissertation, providing a historical context for the four papers and aligning them with their respective periods of social bot research.

Chapter 3 starts by presenting extended abstracts of the five papers of this dissertation to summarize the research questions, methods, key findings, and limitations of each study. This is followed by a more general discussion of the overall contribution of these studies and this dissertation to the study of social bots. The chapter concludes by outlining future research that could build upon the research in the individual papers by addressing some of their limitations and by providing evidence that further generalizes their findings.

Chapter 4 is the final formal chapter and contains a summary of the findings and contributions, highlighting the most significant findings of the dissertation. Following this chapter is an epilogue that contains a more informal and speculative essay on the future of social bots. More specifically, this essay discusses both the challenges related to collecting data on social bots in a post-Twitter world as well as the opportunities that this development offers. The epilogue ends with an optimistic prediction for the future of social bot research. Following the epilogue, a reference list for the dissertation is provided as well as an appendix, which contains the five publications in their respective publication or review formats.

2 A Brief History of Social Bot Research

This essay is both a literature review as well as a commentary on the history of social bot research. The goal is to provide a succinct but comprehensive look into how social bot research has evolved from the early years of social media (beginning in 2010) to the present. Initially, the literature review focuses on papers from computer science journals and conferences, as most of the early and seminal works on social bots are from this field. Later, the emphasis shifts to reflect bot research that has started to appear in information systems conferences and journals. However, seminal and highly cited papers from all fields and interdisciplinary outlets are also included.

The history of social bots can be divided into four approximate periods that have somewhat fluid starting and ending dates. The first period started around 2010, when social media companies such as Twitter and Facebook were gaining popularity and the first articles describing bots in social media began to appear. This period lasted until approximately 2014 and thus precedes the widespread use of the term "social bot," as many articles were still simply referring to "bots" during this time. While studies in the early years of this period were few and far between, they formed the foundation for social bot research and definitively influenced future work on the topic. At this point, there was already evidence of social bots being used to influence operations—for example, regarding elections—dating from as early as 2010 (Ratkiewicz et al., 2010, 2011). However, the research stream focused on studying political bots had not yet become dominant.

The second period started around 2015, when the existence and threat posed by social bots were already clear to researchers and governmental organizations—exemplified, for instance, by the DARPA (the Defense Advanced Research Projects Agency of the United States) "Twitter Bot Challenge," where six teams from companies and universities competed to detect bots (Subrahmanian et al., 2016). During the next year, two seminal works were published: a paper on BotOrNot (later known as the Botometer) (Davis et al., 2016), which presented what was arguably the most popular bot detection tool, followed by the review article "The Rise of Social Bots" (Ferrara et al., 2016), which would become the most cited paper on social bots. In 2016 and continuing into 2017, public awareness and general interest in social bots increased dramatically as a result of the widely publicized Russian interference in the 2016 US presidential election through the use of fake accounts and social bots (Guilbeault & Woolley, 2016; O'Connor & Schneider, 2017). Consequently, the number of publications directly or indirectly discussing bots also started to proliferate.

The third period was characterized by the "arms race" and research saturation. The number of proposed approaches to bot detection continuously increased, while previous methods became obsolete or disappeared into obscurity, as they could no longer detect the latest versions of social bots or were overshadowed by more efficient methods. Two overlapping streams of research were prominent in social bot studies from this period. The first stream consisted of papers that presented case studies documenting the existence of bots and typically focused on specific elections, topics, or regions. The second stream consisted of papers contributing to the development of bot detection methods, with supervised machine learning methods being especially common (Cresci, 2020a).

The majority of the social bot research that has been published in information systems outlets appeared during this period; thus, this section of the essay will contain more field-specific evaluation than the other sections.

The fourth and final period, began in early 2023 after Twitter made it more difficult to use the social media's API. The impacts of this change on social bot research remain to be seen because much of the extant research relied on Twitter data and many of the existing bot detection tools were built on the premise that the Twitter API would continue to be accessible at a reasonable price. However, as many studies were conducted before the changes and the resulting manuscripts are still under review, there will undoubtedly still be a dwindling stream of new research appearing in the near future relying on methods and tooling that are no longer available. Due to this uncertainty, the discussion of the final period will be limited; instead, the epilogue will focus on further speculation and research predictions.

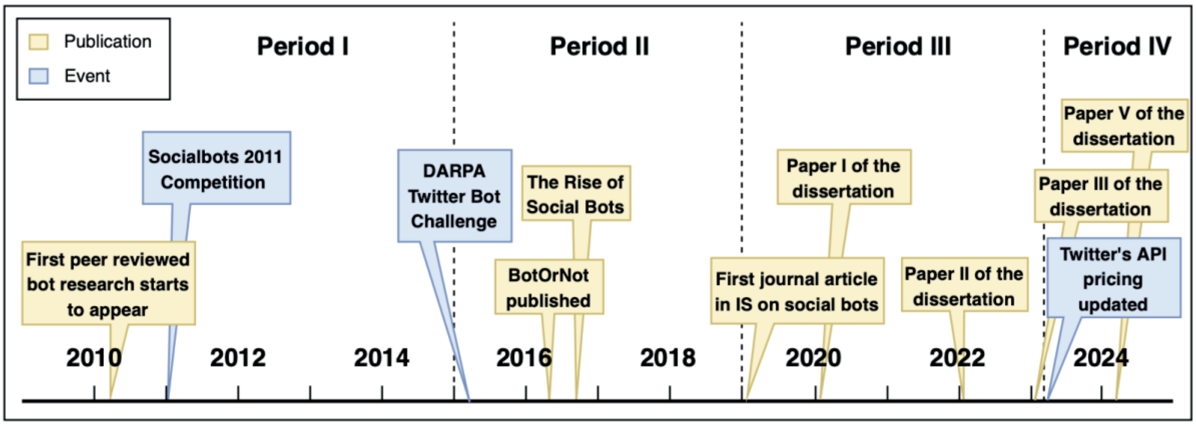

Publication
Event
Period I
Period II
Period III
Period IV
Socialbots 2011 Competition
First peer reviewed bot research starts to appear
DARPA Twitter Bot Challenge
The Rise of Social Bots
BotOrNot published
Paper I of the dissertation
First journal article in IS on social bots
Paper II of the dissertation
Paper V of the dissertation
Paper III of the dissertation
Twitter's API pricing updated
2010
2012
2014
2016
2018
2020
2022
2024

Figure 2: A timeline of social bot research

The following sections of this essay examine each of these four periods more closely and highlight important studies and trending research streams at each point in time. Figure 2 above summarizes key information from the introduction, presents landmark papers and events in social bot research, and positions the dissertation papers within the time periods.

2.1 Early studies on bots on social media (2010–2014)

By the end of 2010, social networking sites were growing rapidly; while Twitter had almost 200 million registered users (Rao, 2010), by July 2010, Facebook already had more than half a billion registered users (Pepitone, 2010). During the same year, the first peer-reviewed publications on bots started appearing, with many using Twitter data (Chu et al., 2010; A. H. Wang, 2010) and some others also other sources, such as Facebook (Stringhini et al., 2010) or MySpace (Lee et al., 2010). In the beginning, bots were often associated with spam, and the term social bot had not yet been coined. At this point, bots were not always presented as a separate topic but were instead discussed in research related to astroturfing and spamming on social media (Chu et al., 2010; Lee et al., 2010; Ratkiewicz et al., 2010; Stringhini et al., 2010). This resulted in some publications referring to the bots as spambots, and researchers also began to use the term "cyborg" to refer to hybrid accounts that were partially automated and partially operated by a human and operated in similar fashion to fully automated bots or spambots (Chu et al., 2010).

As the descriptions imply, spambots or astroturfing bots were often repeatedly sharing unoriginal content or links (spamming), liking or sharing content, or following specific profiles (Chu et al., 2010). This predictable and repetitive behavior was distinct from that of humans, making them easy to detect both manually and algorithmically. Nevertheless, likely due to the vast number of accounts that needed to be labeled, the algorithmic approach was more popular. To identify the bot accounts, many early studies were already employing machine learning-based classifiers (Stringhini et al., 2010; Wang, 2010), which would eventually develop into the most common approach to detect bots (Cresci, 2020a). Compared to the models used in contemporary research, those featured in early studies were simple and relied on a handful of features, such as the posting rate, number of followers, friends, and the ratios of these (Lee et al., 2010, 2011; Stringhini et al., 2010; Wang, 2010).

Many of these early studies relied on honeypot accounts to collect the account dataset they studied (Lee et al., 2010, 2011; Stringhini et al., 2010). In practice, this meant setting up accounts that would attract followers, some of which would also turn out to be bots. This contrasts with the later commonly used practice of choosing to collect, for example, all accounts that tweeted using specific keywords during a specific time frame. At this point, some researchers were also experimenting with bots they developed to collect data. For example, in Boshmaf et al. (2011), the authors created bots themselves and tested their ability to infiltrate Facebook, showing that using bots made it easy to harvest information.

Boshmaf et al.'s (2011) paper, was also interesting for another reason. Their article, titled "The Socialbot Network: When Bots Socialize for Fame And Money," was one of the first academic sources instances to use the term "social bot." Notably, the term "socialbot" was stylized as one word here, and other works in the 2011–2014 time frame also adopted this spelling (Boshmaf et al., 2012; Elyashar et al., 2013; Hwang et al., 2012; Mitter et al., 2014). However, the two-word "social bot" term also appeared in publications around this time (Wagner et al., 2012), eventually becoming the standard, though it is unclear why.

Although the aforementioned 2015 DARPA Twitter bot challenge is possibly the most well-known bot themed competition, it was not the first of its kind. In fact, beginning already in 2011, the Web Ecology Project hosted the SOCIALBOTS 2011 event, where participating teams raced to develop bots that would influence a group of 500 unsuspecting Twitter users, gaining points for friending them and inducing them to perform certain actions.[2] This event ultimately resulted in multiple papers, but the publications focused on humans and how suspectable they are to bots, rather than on the bots themselves (Wagner et al., 2012; Wald et al., 2013a, 2013b). Interestingly, this type of experimentation with bot designs did not become a popular stream of research, despite

[2] http://www.webecologyproject.org/2011/01/help-robots-take-over-the-internet-the-socialbots-2011-competition/

its potential to offer direct insights into human-bot interaction at a level difficult to achieve as an external observer—for example, in studies relying on detecting bots in the wild.

By the end of 2014, the number of articles on the "characterization, detection and impact estimation of social bots" was rapidly increasing (Cresci, 2020a). While the numbers in the early 2010s were modest, in both 2013 and 2014, Scopus was registering approximately 50 new publications per year (Cresci, 2020a). In these early studies, bots and social bots were portrayed almost exclusively as malicious entities, which is not surprising given that early bots were associated with activities such as spamming and astroturfing. The first period ends in 2014, representing the last year in which bot research was arguably still in the slow-growth phase and gaining interest but not yet of interest to wider audiences.

2.2 The rise of social bot research (2015–2018)

The rapid rise in social bot literature starting in 2015 signaled the start of a new period, characterized by papers presenting evidence of bots interfering with elections and revealing an interest in developing increasingly sophisticated detection models. In terms of numbers, in 2016, there were more than 100 new publications (indexed in Scopus), doubling to over 200 per year in 2018 (Cresci, 2020a). Bots were also now being mentioned in articles belonging to another trending and closely related and sometimes overlapping field—the study of fake news and online misinformation—further increasing public awareness and interest in bots (Lazer et al., 2018; Vosoughi et al., 2018). At this point, the term "social bot" had become clearly established.

The beginning of the second period of social bot research featured the publication of two of the most influential publications. The first is the review article "The Rise of Social Bots" (Ferrara et al., 2016), which provided the inspiration for the name of the second period. This paper has been cited more than any other bot paper, with over 2500 citations at the time of writing. It provided a clear description of the state of the art of social bots and proposed a taxonomy of social bot detection systems from crowdsourcing to supervised machine learning-based systems. The second influential publication is the conference proceeding "BotOrNot: A System to Evaluate Social Bots" (Davis et al., 2016), which presented what would become the most well-known (Grimme et al., 2018) and most widely used bot detection tool (Rauchfleisch & Kaiser, 2020). While the tool BotOrNot had been available since 2014, in 2017, it was renamed as the Botometer.[3] The Botometer was the tool of choice for many landmark social bot papers (Shao et al., 2018; Vosoughi et al., 2018), although its reliability and accuracy were criticized in later years (Gallwitz & Kreil, 2021; Grimme et al., 2018; Rauchfleisch & Kaiser, 2020). This will be discussed further in the next section of the essay focusing on the third period of social bot research.

In addition to these two papers, many other widely cited (several hundred citations or more) papers appeared in the 2015–2018 period and had the typical characteristics common to many social bot studies of this period—namely a focus on detection, characterization, and the presentation of

[3] https://botometer.osome.iu.edu/faq#name-change

evidence of bots in political contexts. These include studies showing that social bots were present in the US 2016 presidential election (Bessi & Ferrara, 2016; Howard et al., 2018; Shao et al., 2017, 2018), the UK-EU referendum that resulted in Brexit (Howard & Kollanyi, 2017), the 2017 French presidential election (Ferrara, 2017), and the 2017 Catalan referendum (Stella et al., 2018). More general papers also appeared during this period, and the term political bot became commonly used to describe a specific subset of social bots (Woolley, 2016). Ultimately, these papers likely provided the catalyst for the popularity of the two dominant streams of bot research, which focus on detecting and/or presenting cases of bots interfering with elections.

Table 1: Notable groups, institutions, researchers, and their primary contributions

Groups and institutions	Notable affiliated researchers	Contributions
The Observatory on Social Media (OSoMe) at Indiana University	Alessandro Flammini, Filippo Menczer, Onur Varol, Clayton Allen Davis, Kai-Cheng Yang	Bot detection by developing and maintaining the Botometer (originally BotOrNot). Co-authors of "The Rise of Social Bots."
The University of Southern California	Emilio Ferrara	Co-author of "The Rise of Social Bots" and multiple well-known studies on social bots in elections.
The Computational Propaganda Project at the Oxford Internet Institute	Phil Howard, Lisa-Maria Neudert	Produced multiple reports and papers written with audiences such as governments and policy makers in mind in addition to academic papers.

During this period, as bot research grew, so did the dominance of several groups of researchers. Examining the authors of the publications of this and the following periods described in this essay, a distinct pattern emerges. Certain names and institutions were very heavily represented during this period, coalescing around several influential research groups, whose publications were both widely cited and whose opinions on the matter attracted ample attention from by media outlets (Gallwitz & Kreil, 2021). These groups and the affiliated researchers are listed in Table 1, along with their major contributions. Beyond the three groups listed in Table 1, the researcher Stefano Cresci, a co-author in several highly cited studies (Cresci et al., 2015, 2017; Cresci, 2020a), should also be mentioned. These groups and individuals contributed significantly to establishing social bot research, although in the later years, some also voiced criticism about how their works presented social bots as more powerful than evidence would suggest (Assenmacher et al., 2020; Gallwitz & Kreil, 2021).

Nearing the end of this period, social bot research had started to appear in information systems conference proceedings, and journal publications would soon follow. Some of the earliest examples include Stieglitz et. al.'s (2017) categorization of social media bots, Brachten et al.'s (2017) study on social bots in the 2017 German state election, and Wang et al.'s (2018) *n*-gram-based approach for detecting bots. As can be seen, these early studies closely followed the general trends of bot research by focusing on bot detection and case studies. However, adding diversity to the study of social bots, one IS paper from this early period investigated benign examples of social bots (Brachten et al., 2018). While prior works had already raised critical comments about the field's focus on malicious bots over benign or benevolent alternatives (Oentaryo et al., 2016), they focused on simple bots rather than social bots.

Of all the boundaries between the four periods discussed in this essay, that demarcating the second and the third is perhaps the blurriest, as it is difficult to identify a specific distinguishing landmark other than a reduction in the growth of the pace of publications appearing after 2018 and a shift towards more incremental additions to the new knowledge introduced by papers. Further, at the end of this period, criticism of the dystopian portrayal of social media being controlled by seemingly intelligent social bots was already starting to appear, signaling a shift towards more critical attitudes, in contrast to the excitement and retrospectively even alarmist tone of some of the earlier papers (Assenmacher et al., 2020; Gallwitz & Kreil, 2021). Thus, the second period concluded at the end of 2018.

2.3 The arms race (2019–2023)

The third period began in 2019 and lasted until early 2023. The period is characterized by the "arms race" nature of bot research at this time, as bots were evolving in response to social networking sites taking more aggressive measures to remove them. Consequently, bot detection models had to be regularly updated and expanded to maintain their ability to detect of the next iteration of social bots (Cresci, 2020b). Developing bot detection models and showcasing bots in different contexts remained the primary contribution of many studies during this period. In 2020, academic publications documented the use of bots in the elections of 39 countries across the world (Cresci, 2020a). One such study is paper I of this dissertation, which presents a metadata-based approach to bot detection and demonstrates it through the detection of bots following Finnish politicians on Twitter (Rossi et al., 2020). Further adding to the "arms race," highly cited papers presenting new bot detection models began to claim outlandishly high performance measures with near-perfect classification rates, such as an AUC[4] of 0.99 (Kudugunta & Ferrara, 2018; Sayyadiharikandeh et al., 2020).

As another feature of this period, more critical voices regarding the true capabilities and level of sophistication of social bots began to emerge, contradicting the notion that bots are constantly evolving and increasingly advanced (Assenmacher et al., 2020; Rauchfleisch & Kaiser, 2020, 2020). The critics' main concerns were the mystification of social bots and their capabilities as well as researchers' trust in tools such as the Botometer. Dissertation paper II also notes this contradiction between empirical evidence and the existing literature. Paper II focuses on bots involved in spreading COVID-19 conspiracies via misinformation posts, which Twitter was supposedly monitoring. Most of these bots turned out to be simple spambots retweeting posts repeatedly, rather than sophisticated bots posting original content or engaging with other users

[4] AUC = Area under the ROC Curve. An aggregate measure of classification performance commonly used in machine learning.

(Rossi, 2022). Despite the discussion regarding the validity of the Botometer's classification accuracy, papers that relied on it for bot detection kept being published—since the tool was being regularly updated, comparing its performance between papers is difficult without establishing the version of the service that was used (Sayyadiharikandeh et al., 2020). Further adding to this difficulty of reliably evaluating the accuracy of the Botometer, the outputs of the model varied from version to version. Moreover, the Botometer did not claim to perform binary classification between bots and humans but instead provided a score between zero and five, with higher scores indicating bot-like behavior but not necessarily indicating whether the account is a bot.[5]

While in computer science, social bot research had matured by this point, in information systems, the topic was only emerging, especially prior to 2020. Most bot research published in IS outlets appeared during this period, and while some papers followed the trend of focusing on bot detection and case studies (Hofeditz et al., 2019; Marx et al., 2020; Onuchowska et al., 2019; Rossi et al., 2020), many also contained novelty in terms of their methods or goals. The journal publications ranged from a novel contribution to bot detection methods augmented by crowd-generated labels (Benjamin & Raghu, 2023) to theory on how social bots disseminate information (Salge et al., 2022) to an investigation of how a very small community of bots can tip the opinion climate of polarized online discussions (Ross et al., 2019). Considering the prevailing "theory fetish" of information systems research, as argued by Iivari (2020), where "excessive emphasis is on theory and theory building," social bot studies so far have been very light in terms of theoretical contributions.[6] This is possibly explained by the dearth of journal publications on social bots within the IS field, which will change as conference proceedings eventually become journal publications. One of the theories used in more than one social bot paper in IS is the *spiral of silence* (Noelle-Neumann, 1974), which is discussed extensively by Ross et al. (2019) and lightly covered in dissertation papers II and III (Rossi, 2022; Rossi et al., 2023). Other theories linked to social bot research include the algorithmic conduit brokerage theoretical framework proposed by Salge et al. (2022), dual process theory in relation to combating extremism online through social bots (Blasiak et al., 2021), and speech act theory in a framework used as part of a bot detection method (Benjamin & Raghu, 2023). Overall, many of these papers focused on empirical observations and on providing information to broader audiences beyond other scholars within the field. Given the importance of social bot research to society, I argue that this trend is beneficial, and it would therefore be preferable for social bot research to maintain its connection to practice rather than prioritizing heavily theoretical contributions.

Putting into motion the event marking the end of this period of social bot research, in April 2022, the controversial billionaire Elon Musk initiated the process of acquiring Twitter. While he

[5] The section "Can I use a threshold to classify bots?" from https://botometer.osome.iu.edu/faq further elaborates on this.

[6] It should be noted that we faced issues regarding the lack of theorizing during the review processes of paper II, paper III, and paper V. While all papers received praise for the method, relevance, and writing from a majority of reviewers, all three also faced pushback from always exactly one reviewer for not being proper IS contributions if there was no theory development.

attempted to back out of the acquisition by claiming that the social networking site had more bots than previously disclosed, he eventually ended up acquiring Twitter later the same year (Wile, 2022). Initially, the direct impact on researchers was limited. Presumably due to mass layoffs as well as voluntary turnover, the stability of the platform and its API was threatened, but the site remained mostly functional.[7] However, soon thereafter, Musk announced that Twitter would be changing its current pricing scheme in the attempt to more aggressively monetize the site's API. In contrast to the date separating the second and third periods of social bot research, the end of the third period has a precise date—February 9, when Twitter removed its API's free tier (Weatherbed, 2023). On this day, third period of social bot research ended, leading to the fourth and current period.

2.4 The end of Twitter as a source of data and the uncertain future (2023 –)

In March 2023, Twitter unveiled the new pricing scheme of its API. The change rendered using the API too expensive for academic users, who previously were able to download hundreds of thousands of tweets for free (Calma, 2023). Collecting datasets that are close to the sizes typically used in studios prior to the changes to the API would be challenging, as the new pricing scheme at the basic tier only provides access to 10,000 read requests (tweets) per month. In comparison, Bessi and Ferrara's (2016) study on how social bots distorted the 2016 US presidential election was based on over 20 million tweets, while Shao et al.'s (2018) *Nature Communications* article on the spread of misinformation by social bots analyzed 14 million tweets. While accessing such volumes of data is still possible, in practice it is unfeasible for all but the wealthiest organizations, as the cheapest enterprise plan has a monthly fee of $42,000, providing access to 50 million tweets per month (Mehta, 2023).

An immediate consequence of this change to the pricing model was that bot detection tools that were built on top of Twitter's API, such as the Botometer, stopped working. Researchers were thus deprived of what was arguably the most accessible (in terms of ease of use) and productized bot detection tool. Thus, future studies will once again depend on researchers being able to either reuse code and models shared by others or able to build their own bot detection model. Considering the criticism that the Botometer faced in its last years, some might argue that moving away from reliance on a single tool will result in higher quality research, but it may also make comparing studies more difficult if every study relies on a different bot detection method.

As this period has just started, the full consequences of Twitter data becoming largely inaccessible are not yet known. As this literature review focuses on the past, further discussion of the implications appears in the dissertation epilogue.

[7] One of the research projects I was involved in was collecting data with Twitter's API during the winter of 2022–2023, and we noticed that crashes and other bugs were more frequent than before. More concerningly, API calls would occasionally go through successfully but retrieve only partial data, with many attributes missing from all collected rows.

2.5 Concluding remarks

Social bots and social bot research have existed for a relatively brief but eventful period of time. This essay summarized some of the major milestones and trends that we have seen thus far in this area of research. In computer science, the peak of interest in social bots may have already passed, while in information systems, it may still lie ahead due to longer review cycles in IS journals. At the time of writing, the debates regarding the true influence of social bots on humans and how well humans can detect them are still ongoing. Given that social networking sites are working against rather than with researchers to answer these questions, it may take time before these contradictions in the social bot research findings are resolved.

While the disappearance of Twitter as a data source may slow down social bot research for the time being, this is still an exciting time to be studying bots. Methodologically, social bot research is diverse and may even become more open to new methods now that many previously used approaches are no longer viable. Moreover, in the past few years, in addition to traditional lines of research focusing on malicious use cases and detection, we have seen a growing number of publications investigating more diverse applications of bots on social networking sites, from handling sales on social media to deradicalizing people online. While it is important to maintain and develop our knowledge about malicious social bots, the study of benevolent bots could help broaden the interest in social bots and reduce the risk of social bot research stagnating and becoming saturated with similar research. Finally, given that the intelligence and capabilities of social bots have previously been questioned, research focusing on sophisticated social bots (that are hopefully created with benign or benevolent intentions) that can make use of the developments in large language models and other forms of generative AI would certainly be welcome. Thus, I reiterate, exciting times are still ahead of us social bot researchers!

3 Summary of articles and contributions

This chapter starts with a summary of each of the dissertation's five papers. More specifically, it presents the key information of each paper in the form of extended abstracts, including the research questions, methodology used, key findings, limitations, and contributions. This is followed by a discussion of how all the papers are linked together and a description of the overall contribution of the dissertation. The chapter concludes with a discussion of how future research could extend the research presented in the dissertation.

3.1 Paper I: Detecting political bots on Twitter during the 2019 Finnish parliamentary election

Paper I was published in the proceedings of the 53rd Hawaii International Conference on System Sciences in 2020. The paper is a traditional political bot study, which proposes a new machine learning model for detecting Twitter bots and then demonstrates its power by analyzing the followers of Finnish politicians prior to the 2019 Finnish parliamentary elections. The research questions were 1) "What are important features that can be used to identify bots?" and 2) "Do the bots have an impact on Finnish politics?"

Building on the features commonly used by previous classifiers, the paper presents a machine learning model for bot detection that labels individual accounts as bots or humans using only user-level metadata. The random forest model was built using R and was trained with a mix of manually labeled training from the Finnish Twittersphere and the previously published cresci-2017 dataset[8]. The value of this approach, compared to contemporary methods, is that it is computationally light and uses the API to retrieve user data only, rather than tweet data, which would require more calls to Twitter's API. The model achieved an accuracy of 0.837, a recall of 0.846, and a specificity of 0.793, which were modest results even at the time of publication but justified by the limited number of features used by the model and by the resulting lightness in terms of needed API calls and computational resources.

The model was applied to approximately 559,000 unique accounts that were following the most popular Finnish politicians from each party on Twitter. Of these accounts, the model classified roughly 204,000 as bots, indicating that slightly above 36% of the followers of the selected accounts were bots. Most bots detected by the model were extremely simple: they had default profile pictures and blank bios and while most had no followers or tweets, they were following 20–100 accounts. Many of the bot accounts had been created with a browser that had either Arabic or Russian as the language, which is a data point that Twitter provided in user-level metadata, although the profile names were anything from random strings to typical Finnish-sounding names. At the time, Twitter estimated that only 5–10% of accounts on the platform were bots. While the

[8] The dataset is available at the Botometer's repository https://botometer.osome.iu.edu/bot-repository/datasets.html

results of this study reflected much higher numbers, they were in line with findings from other research of the same time period.

The paper posed five research questions. To answer the first question, we determined the five most important features for identifying bots (in the context of accounts following Finnish politicians) as follows, in order of importance: 1) the number of accounts the profile followed, 2) the ratio of following to the age of the account, 3) the age of the account, 4) the ratio of followers to following, and 5) the ratio of likes to the number of accounts the profile followed. While it may seem counterintuitive that the most important feature was the "number of accounts followed by the profile," this feature resulted from the design of the most common bots found in the dataset, which followed exactly 21 other accounts. We assumed that this was either because the profiles were generated using the same script or that this number was an effect of standard Twitter recommendations at the time. It is possible that Twitter suggested popular accounts for new users to follow and that it always suggested 21 accounts. The other important features were in line with previous studies—ratio features are viewed as an effective way to capture information related to the behavior of bot vs. human profiles. Moreover, other sources have suggested that the age of the account may be indicative of bots when coupled with the other features, as many of the suspected bot profiles were recently created.

Answering the second research question proved more difficult given the way the model performed, which is a limitation of the paper. The main limitation of the model is that while the detection methods were able to identify simple bots, they struggled to identify more elaborate social bots whose behavior more closely matches that of genuine human profiles. This is based on the characteristics of the accounts that were labeled as bots, as most of them seemed to belong to a follower farm and were not actively engaging with other accounts through tweets or retweets. As a further limitation, there was relatively little investigation into the accounts following the politicians, and the study could have been further augmented by a longitudinal study of the suspected bot profiles. Thus, determining the impact of the bots was difficult, but based on the limited evidence, the impact of bots seemed minimal based on the low activity of the suspected bots. Moreover, further analysis showed that the number of bots was highly correlated with the number of accounts that were following each Finnish politician overall. This suggests that the politicians were "honeypots" that the profiles were following due to automation rather than intention, given that the profiles were likely recommended by Twitter when a new account was created because they were among the most popular Finnish Twitter profiles. As stated in the paper, this finding was later further confirmed by the Finnish Security Intelligence Service, which concluded that there was no evidence of foreign entities attempting to influence the 2019 elections.

Initially, the primary contribution of this paper was the methodology, as it presented a light metadata-based alternative to Twitter bot detection using machine learning. Since the paper was written in late 2019 and published in January 2020, the proposed method is no longer usable as originally intended due to changes to Twitter's API. Furthermore, as mentioned in the limitations of the paper, the proposed bot detection method was suitable for identifying basic bots but not sophisticated enough to detect more advanced social bots. Thus, the primary contributions of the paper were the following. First, it showed that the Finnish Twittersphere contained bots at a rate far above what Twitter was suggesting at the time. Therefore, it was one of the many studies of the time showing the existence of bots around the world and was cited as evidence of bots in Finland in the article "A Decade of Social Bot Detection" (Cresci, 2020a) and in a report by the NATO Strategic Communications Centre of Excellence (Van Sant et al., 2020). The paper also showed that while advanced social bots were of primary interest to research, crude bots belonging to "following farms" were still prevalent and Twitter's efforts to remove even the most basic bot profiles were inadequate.

3.2 Paper II: The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation

Paper II was published in the proceedings of the 55th Hawaii International Conference on System Sciences in 2022 and is a single-author paper. This paper is a bot case study, where network analysis and the Botometer were used to detect suspicious accounts that were distributing COVID-19-related conspiracy theories on Twitter during the spring of 2021. The word "scamdemic" was a term used primarily by those distributing misinformation on vaccines and the COVID-19 pandemic. Its origin is possibly related to Twitter placing a soft ban[9] on the more popular related term "plandemic," which led to the use of alternative terms to continue distributing misinformation about the pandemic. The research objectives[10] of the paper were, first, to determine whether the COVID-19 conspiracy keyword "scamdemic" was being distributed by bots or organically by humans and, second, to evaluate whether Twitter was enforcing its COVID-19 misinformation policies by banning accounts that were continuously violating the platform's policies.

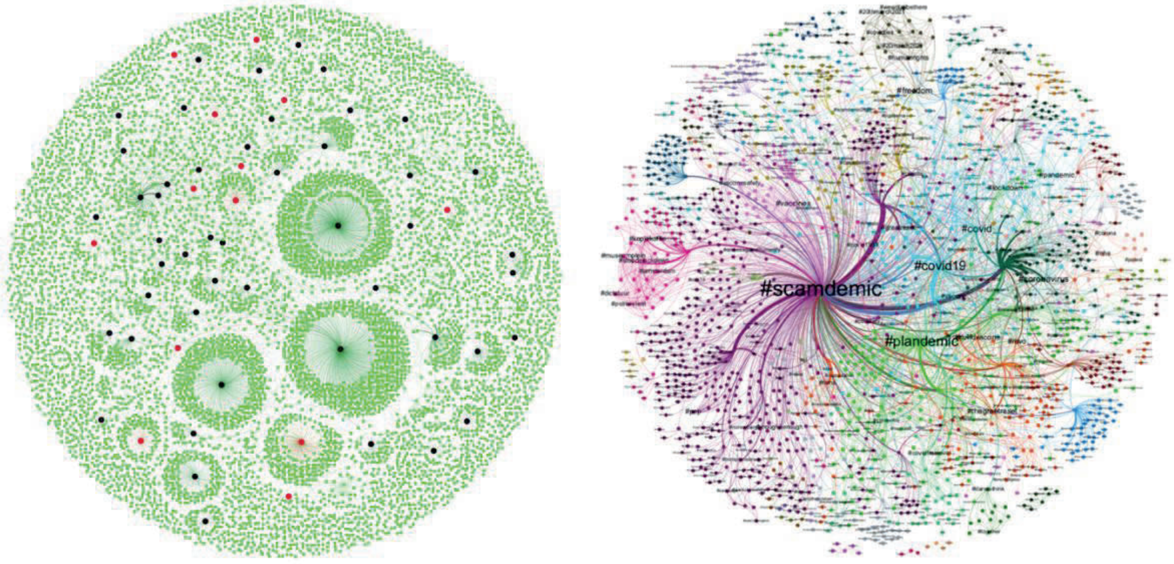
The study was based on a sample of 8263 tweets and 8540 related users that interacted with the tweets by replying, quoting, or being mentioned in the tweets. The data contained tweets found with the keyword "scamdemic" during the week of March 8–March 15, 2021. This time period was selected because Twitter updated its COVID-19 misinformation policy at the beginning of March and had supposedly started enforcing it. The data was collected in a manner that supported creating two types of networks. The first network consisted of accounts (nodes) and interactions between accounts, represented by weighted directed edges. In other words, if there were two accounts, accounts A and B, and B mentioned A twice during the monitoring period, there would be a directed edge from B to A with a weight of 2 in the network. The second network was multimodal and similar to the previously described network, with the exception that nodes could

[9] When attempting to search for content using the keyword "Plandemic" on Twitter, the page would first provide a link to a regional government page on the pandemic and redirect the search to the keyword "pandemic."

[10] Unlike in the previously presented paper I, in paper II there are no explicit research questions and the paper only lists two "goals."

also be posts instead of only accounts, allowing for the mapping of how different accounts interacted with specific posts.

The first network was used to determine influential accounts in the networks, which were then followed and analyzed for a duration of two months. More specifically, selecting the top 25 accounts based on three network characteristics (25 with highest betweenness centrality, 25 with highest indegree, and 25 with highest outdegree), created a list of 61 influential accounts. Since some accounts were at the top in multiple characteristics, there were 61 rather than 75 accounts. These accounts were first checked with the Botometer to determine which accounts were likely to be bots. Furthermore, I qualitatively inspected them, checked their recent tweeting history, and then coded them based on their characteristics into the categories of "conspiracy theorist," "spammer," "antivax," "celebrity," and "non-believer." Finally, I used the second network consisting of accounts and hashtags to determine which keywords in addition to "scamdemic" were popular and to see if clear communities could be identified within the network that used keyword clusters. Among the over eight thousand tweets, there were 3127 hashtags. The 10 most popular hashtags accounted for 34.9% of the total number of hashtags used. The two networks are shown in Figure 3.



Red nodes are influential bots, black nodes influential humans, and green nodes other accounts.

The colors in the network above represent communities of accounts.

Figure 3: The network of accounts (left) and network of accounts & hashtags (right)

The main findings of the paper were that a relatively small share of the influential accounts that were using COVID-19 conspiracy terms were clearly identifiable as bots (scoring 4.0 / 5 or higher on the Botometer). Further manual inspection indicated that up to 13 of the 61 influential accounts (21%) may have been bots, based on their behavior, which for the majority, consisted of merely retweeting and spamming unoriginal content. This type of behavior would suggest that the accounts were following a simple script and thus could be described as spambots rather than social bots. Moreover, only 12.7% of the accounts were suspended after two months, even if they were repeatedly tweeting content that was clearly labelable as misinformation and in violation of Twitter's COVID-19 policies. The hashtags used by these accounts referenced a plethora of different conspiracy theories, such as "The Great Reset," "The New World Order," and the "Plandemic." Ultimately, the paper concluded that a large share of the conspiracy content was being distributed organically by humans rather than bots and that Twitter was not enforcing its own COVID-19 policies.

The main limitation of the study was the number of data points collected due to restrictions posed by the approach used to collect data. Collecting network data is not supported naturally by Twitter's API and thus required making significantly more calls to the API than required when retrieving user-level or tweet-level data only. Moreover, the Python library twarc that was used did not support accessing Twitter's API with an academic license, further limiting the number of calls that could be made to the API. However, for the purpose of demonstrating Twitter's lack of removing content that clearly violated the platform's terms of use, this sample was deemed to be sufficient. Lastly, while the reliability of the Botometer was already being questioned when paper II was being written, the paper used a particularly conservative approach to using the tool. The threshold for assuming that an account was a bot was high: to be labeled a bot, each account needed a Botometer bot score of more than 4.0 out of 5 and had to pass a manual review three times over the course of two months.

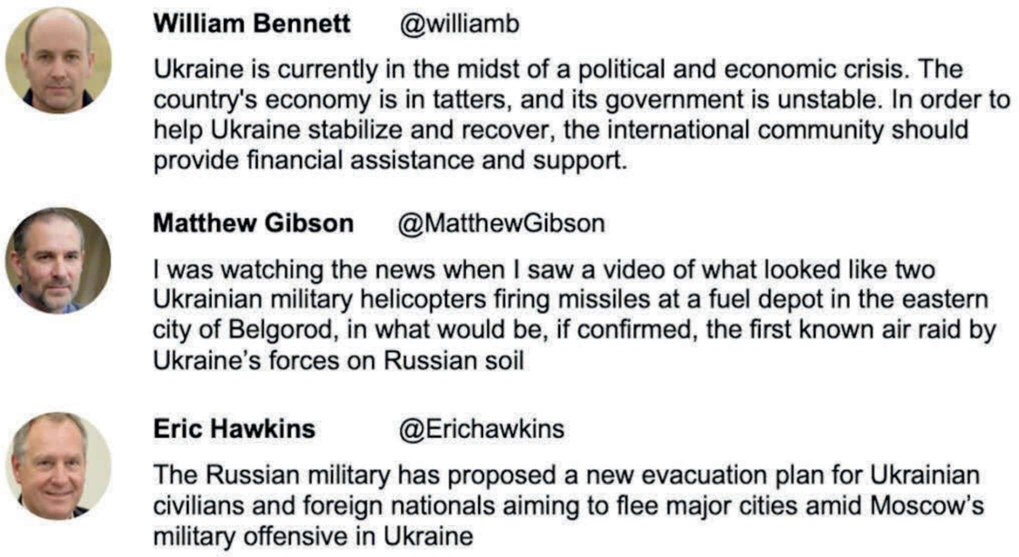
Similar to paper I, this paper contributed to the body of literature suggesting that simple spambots are still prevalent on social networking sites such as Twitter, even though the current research primarily focuses on social bots. Furthermore, paper II provided evidence that repeatedly sharing malicious content on Twitter via bots, trolls, or humans during the time frame of the pandemic was feasible, despite Twitter's claims that they were actively removing content that violated their misinformation policies. Thus, the main contributions of paper II were empirical and provided information to researchers, policy makers, and the general population.

3.3 Paper III: Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?

Paper III was published in the proceedings of the 56th Hawaii International Conference on System Sciences in 2023. This paper focuses on the human ability to detect advanced social bots created using generative AI. More specifically, the paper presents the results of a pilot study where participants were asked to label accounts as bots or humans after seeing a tweet and the basic profile information of the account that posted the tweet. This was designed to simulate situations where people scroll through social media feeds and see posts written by accounts they do not recognize or follow. The research questions of the study were 1) "Can humans distinguish social media profiles with DL-generated profile pictures and DL-generated posts from real ones in the feed of a social networking site?" and 2) "Which components of a profile are more likely to make humans suspect that the profile is fake?" The main hypothesis of the paper was that humans would not be able to distinguish modern bot profiles created using generative AI from human profiles when given limited information similar to what is visible in a social media feed.

To answer the research questions and test the hypothesis, the paper used an experiment with human subjects. Participants were shown a random sample of accounts and posts and then asked to first label the account as bot or human and then evaluate how confident they were with this label. They also had to rate which features of the post and account (profile picture, tweet, name of the account, handle of the account) made them suspect that the account might be a bot. The accounts and posts that participants were shown were drawn from a sample of 18 accounts—nine genuine Twitter posts and profiles and nine bot profiles and posts that were created for the experiment. The topic of all the posts used in the experiment was the war in Ukraine because, given its political nature, it thematically fit an experiment related to bots and because it was a trending topic on Twitter.

The process for creating the bot profiles was based on mimicking practices used by real bot developers. The bot profiles had pictures that were created using a script that repeatedly visited the website thispersondoesnotexist.com, which uses NVIDIA's StyleGAN, a pre-trained deep learning model that can generate realistic images of human faces to produce a unique image every time the page is visited. The bots' posts were written with GPT-3 by feeding the language model articles about the war and prompting it to create short responses. The profile names and handles (short unique identifiers next to the name used by Twitter) were generated using a simple Python script that took first and last names from a list of common English names. Figure 4 contains examples of the generated profiles and posts.



William Bennett @williamb

Ukraine is currently in the midst of a political and economic crisis. The country's economy is in tatters, and its government is unstable. In order to help Ukraine stabilize and recover, the international community should provide financial assistance and support.

Matthew Gibson @MatthewGibson

I was watching the news when I saw a video of what looked like two Ukrainian military helicopters firing missiles at a fuel depot in the eastern city of Belgorod, in what would be, if confirmed, the first known air raid by Ukraine's forces on Russian soil

Eric Hawkins @Erichawkins

The Russian military has proposed a new evacuation plan for Ukrainian civilians and foreign nationals aiming to flee major cities amid Moscow's military offensive in Ukraine

Figure 4: Three of the bot profiles used in the experiment

The experiment was conducted in May 2022 and participants were recruited via the crowdsourcing service Amazon Mechanical Turk (MTurk). Based on an initial screening phase involving 1292 subjects, 478 were invited to complete the full experiment and 375 participants completed it successfully. The participants were from the United States, with 56% identifying as male and 43.7% as female. To ensure high-quality responses, the experiment included attention checks to exclude participants providing low-quality responses. Furthermore, to incentivize high-quality responses, the participants were told that they would be rewarded based on their performance and that failing an attention check or being caught rushing through the experiment would result in no reward. Lastly, to check whether the instructions for the task given in the experiment were sufficient and to determine how the participants perceived the difficulty of labeling accounts, we asked two questions about the design of the experiment. The first question asked if "the tasks and instructions were clear"—96% of participants agreed or strongly agreed. The second question asked if "the given task was easy to do"—83% of participants agreed or strongly agreed.

The main finding of the experiment was that participants cannot distinguish bots from humans: the accuracy for detecting bots ranged from 10% to 27.4%, while for the human profiles, it was between 58.5% and 91.4%. This resulted in an overall accuracy of 48.9% for all accounts. Since each of the 375 participants saw only four of the accounts, each account was labeled approximately 80 times. Furthermore, no statistically significant relationships were found between how the participants labeled the accounts and how they rated the suspiciousness of each of the four features of the posts and profiles. The results are summarized in Table 2.

Table 2. Classification accuracy

Accuracy	Generated	Genuine
Mean	18,2 %	79,7 %
95% CI	14.5% - 21.9%	73.7% - 85.6%
Highest	27,4 %	91,4 %
Lowest	10,0 %	58,5 %

This study had several limitations. First, the number of accounts and visible pieces of information per post was limited. In the real world, a user would be able to investigate accounts more thoroughly, but since the study sought to emulate a setting where users come across posts in a feed and since most people would not go through every profile they come across, this limitation was deemed acceptable. However, as a result, the findings cannot be generalized outside a setting where users see only limited information, such as when scrolling through a social media feed. The second limitation was that the participants were quite homogeneous and since they were recruited via MTurk, they were most likely not representative of the general population. This limitation was considered acceptable because the paper presented a pilot study and was mainly used to guide the development of an experiment with higher ecological validity, which is presented in paper IV.

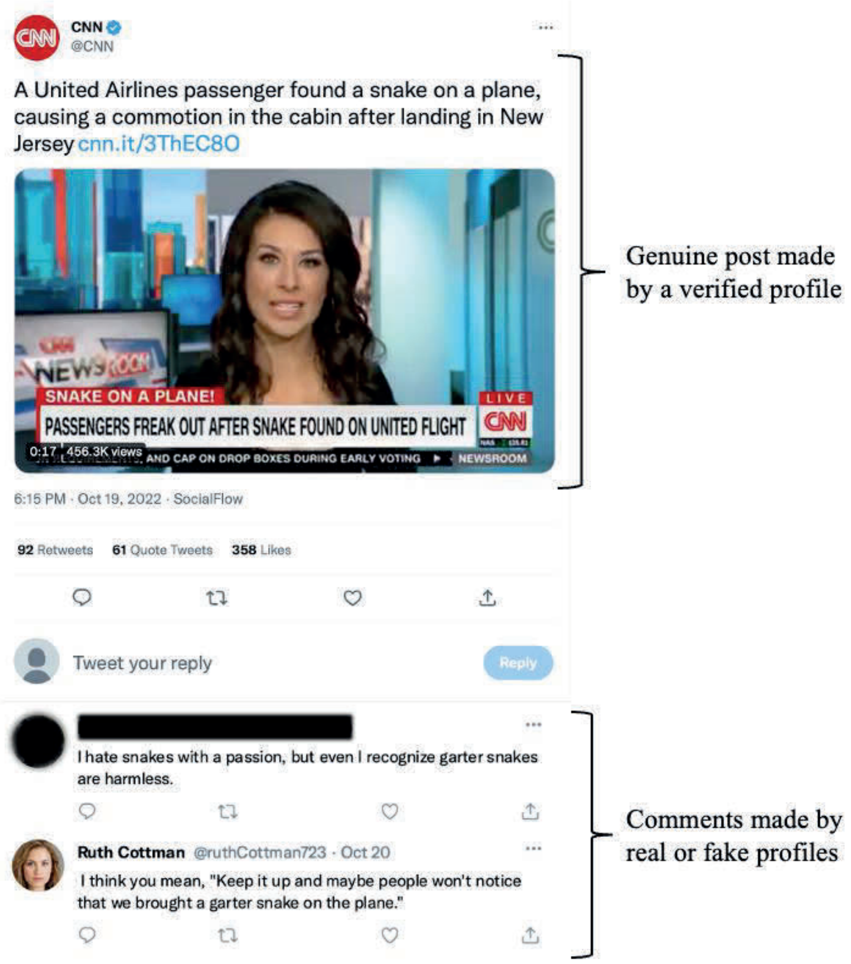
In late 2022 when the paper was written, previous research had already shown that large language models were capable of producing short texts that human evaluators could not distinguish from texts written by humans (Clark et al., 2021; Dugan et al., 2023). Furthermore, multiple studies had already shown that images generated using generative adversarial networks, a deep learning architecture, could produce images that humans could not distinguish from genuine photos (Nightingale & Farid, 2022). The contribution of this paper was combining both generated text and images into a profile that represented social bots in a social media feed and showing that such profiles can no longer be distinguished from genuine human profiles when shown to a human evaluator. Given the limitations of the study, the extent to which this finding is generalizable is still uncertain, but the initial results are concerning and if further validated would have a significant effect on, for instance, producing training data for bot detection models, which often rely on training data that has been manually labeled by humans.

3.4 Paper IV: AI-Generated Profiles Are Indistinguishable from Real Profiles in Social Media Feeds

Paper IV is currently under review at the *Journal of Information Technology*. Building on the pilot study that was presented in paper III, for paper IV, we conducted a more robust set of experiments to further confirm the hypothesis that humans cannot distinguish AI-generated bot profiles from genuine human profiles. To increase ecological validity, rather than showing isolated posts, as in paper III, the experiments in this study used full discussions that consisted of multiple tweets that were replying to one tweet. The research questions of the paper were 1) "Can humans distinguish AI-generated bot profiles from genuine profiles that are commenting on a social media post?" and 2) "Which features on a bot profile can humans detect are AI-generated and not genuine?"

The setup of the experiment was very similar to the process described in section 3.3. Again, participants were first given instructions on how a Twitter discussion works and explanations of the different visible components in the posts (the tweets, profile pictures, names, and handles). Then, participants were shown a randomly drawn sample of Twitter feeds one at a time and were asked to label each of the accounts that were participating in the discussion as bots or humans. There were six variations of the feeds shown during the experiment, with discussions containing a varying number of bots and humans. Thus, some participants saw posts that contained only genuine human discussants, others saw discussions with mixed bot and human participants, and still others saw discussions populated only by the generated bot accounts. One of the discussion threads shown during the experiment is shown in Figure 5. In addition to labeling the accounts as humans or bots, the participants were asked to rate their confidence in the label they gave and evaluate how much each of the four features contributed to their confidence in the account being bot or human.

The process used to generate the profile pictures, names, and handles was identical to the process described in the previous section (3.3). Thus, the profiles again used names and handles that were created with a basic Python script, and the profile images were again created using the website thispersondoesnotexist.com, which generates a unique profile picture using NVIDIA's StyleGAN deep learning model every time the page is visited. However, as the bot profiles' tweets should have been replying to specific discussions at this point, the process for generating the post content was different. In this paper, the main post and the existing comments under it were fed into GPT-3 and it was asked to respond or add to it with a tweet-length post.



CNN
@CNN

A United Airlines passenger found a snake on a plane, causing a commotion in the cabin after landing in New Jersey cnn.it/3ThEC8O

NEWSROOM
SNAKE ON A PLANE!
PASSENGERS FREAK OUT AFTER SNAKE FOUND ON UNITED FLIGHT
LIVE
CNN
0:17
456.3K views
AND CAP ON DROP BOXES DURING EARLY VOTING
NEWSROOM

6:15 PM · Oct 19, 2022 · SocialFlow

92 Retweets 61 Quote Tweets 358 Likes

Tweet your reply

[redacted]

I hate snakes with a passion, but even I recognize garter snakes are harmless.

Ruth Cottman @ruthCottman723 · Oct 20

I think you mean, "Keep it up and maybe people won't notice that we brought a garter snake on the plane."

Genuine post made by a verified profile

Comments made by real or fake profiles

Figure 5: An example of a thread used in the experiment[11]

We performed the data collection by conducting the experiments in parallel at three universities and via MTurk during 2023. At the three participating universities (Aalto University, CBS, and Thammasat University), the participants were primarily graduate students of various backgrounds, while the participants recruited via MTurk were from the United States. The purpose of using a diverse set of participants recruited via different channels was to address issues related to solely using MTurk and because previous related studies on the human ability to detect AI-generated content had partially contradictory results depending on where the participants were

[11] Note that in the actual experiment, the black boxes were not there and the participants saw a profile similar to the one at the bottom of Figure 5. As the first account commenting on the post by CNN was a genuine Twitter user, to preserve the user's anonymity the profile picture, name, and handle were redacted from the dissertation and paper IV.

recruited from. The experiment was completed successfully by 231 university participants and by 252 MTurk participants.

In the experiment conducted on university students, the average accuracy for classifying accounts as bot or human was 56%. Notably, one of the bot accounts was labeled as a bot account by only 6.3% of participants, whereas one of the accounts belonging to a human was labeled as a bot by slightly more than half (53.9%) of the participants. In the statistical analysis of the findings, a generalized linear mixed-effects model (GLMM) was used. The lower rate of identifying bot accounts when compared to human profiles was statistically significant. In the second experiment, which was conducted with MTurk participants, the results seemed more random and the average accuracy for labeling the accounts was only 48%. In general, the MTurk participants were more suspicious of the accounts than the human participants, which was also in line with their higher rate of mislabeling human profiles as bots. Moreover, a statistical effect similar to the effect in the experiment using students was not found; the MTurk participants did not show a statistically significant difference in their ability to correctly identify AI-generated or human profiles, although their overall accuracy was low.

Similar to the limitations presented in paper III, in paper IV, the main limitation was the limited amount of information visible on each profile. Once more, this limitation was justified by the argument that the study aimed to replicate a situation where a user is scrolling through a social media feed rather than intentionally attempting to scrutinize each profile in detail. Moreover, paper IV realized an improvement in terms of ecological validity compared to the previous study; the feed was much more authentic in that it consisted of related posts commenting on a main post rather than a series of unrelated posts.

The main contribution of this paper was further demonstrating that bots using modern generative AI are becoming challenging if not impossible to detect on social media feeds. Moreover, this paper showed that humans have difficulties labeling human and bot accounts even when the accounts are participating in a discussion thread on a social networking site such as Twitter. This would suggest that generative AI-powered social bots are indeed capable of evading detection, as suggested by many publications; however, given the limitations of the study, it is not yet clear whether this would hold if the people evaluating accounts were given access to more data points, such as full profile information and multiple posts for each account.

3.5 Paper V: Augmenting Research Methods with Foundation Models and Generative AI

Paper V was published as a peer-reviewed editorial in the *International Journal of Information Management* (Rossi et al., 2024). The paper discusses the opportunities and challenges related to using foundation models[12] and generative AI in research. More specifically, the paper outlines

[12] Foundation model is a term used to describe large pre-trained machine learning models that can be used as is or after further fine-tuning in tasks such as image/text generation/summarization/classification. Examples include BERT, Llama 2, GPT-n.

how these models can be used to develop content and data to support conducting different forms of research such as experiments. Unlike papers I-IV, which focus on studying different aspects of social bots, the goal of paper V was to formalize some of the methods that were used in papers III and IV and propose how they can be used to augment existing research methods. Many of the proposed use cases for generative AI in this paper were demonstrated by examples of generating content to social media experiments and are thus well suited for social bot research, thus linking the paper thematically to the dissertation.

Paper V begins by defining the terms "foundation model" and "generative AI" and summarizing the recent developments in deep learning, specifically highlighting the possibilities engendered by recent advances, such as the generation of text and images that are indistinguishable from genuine photos or human-written text. This is followed by a more critical review, recalling the risks associated with and the limitations of these deep learning methods. After presenting the background and the current state of the art, the paper proceeds to its main contribution, which is a discussion of how these new technologies can be adopted by researchers. As an editorial, the paper does not have research questions or findings but instead presents a number of potential uses for foundation models and generative AI in information systems research, drawing on examples from recent studies both within and outside of IS.

The first proposed use case is using generative AI to create content such as text or images for experiments, building heavily on the research presented in dissertation papers III and IV. The main benefit of using generative AI is that it allows for the production of realistic and controlled variations of content that are used in experiments, such as images of a person of varying ages or ethnicities, or, alternatively, writing short texts with subtle differences in tone. For example, if a variable in a study[13] is the age or ethnicity of a person whose photo is shown to participants, finding or taking real photos of humans would be significantly more labor intensive than using generative AI to produce the images.

The second proposed use case is using foundation models to produce synthetic data for quantitative research, with examples from both text and images. The paper discusses two main benefits of creating and using synthetic data—namely, creating a privacy-preserving but characteristically similar dataset and increasing the number of data points by creating slight variations of data points. It should be noted that both of these have strict limitations regarding their appropriate use. One example of a situation where synthetic data can be used is with training machine learning models for well-defined and well-understood problems, such as image classification and recognition.

While the examples of how to use generative AI and foundation models are applicable to those using specific research methods such as experiments, the paper has a more general set of guidelines that are applicable to all research making use of generative AI and foundation models.

[13] Examples of such studies that are still under review include: 1) a study by Youngjin Kwon, which investigates how ethnicity visible in LinkedIn profiles affects perceptions of recruiters, and 2) a study by Yuting Jiang, which investigates bias in gig economy hiring trough generated variants of worker profile pictures.

In addition to the methodological suggestions, the paper outlines ethical considerations related to using these tools and proposes four principles that should be followed: First, humans should be kept in the loop, which means that the qualitative and or quantitative inspection of generated content and data is always needed before using generated materials in a study. Second, the use of foundation models and generative AI should always be disclosed, even when their use is minimal or seemingly inconsequential. Third, sufficient documentation should be provided to allow for the evaluation of the use of foundation models and generative AI. This includes listing the model version, the parameters, and the prompts that are used. Fourth, authors should always store and be able to provide access to the generated data upon request.

Due to the recency of the topic, there were few relevant studies at the time of writing that had adopted foundation models or generative AI. It should be noted that many of the proposed methods will require further validation and analysis to determine their practical usability. Therefore, the primary contribution of this paper is ultimately found in the proposed ethical principles and the examples of good practices when using foundation models and generative AI rather than in the suggested approaches to augment existing research methods. Moreover, as an editorial, the full contribution of the paper will be seen only after sufficient time has passed and it can be evaluated if the practices suggested by the paper have been adopted or further developed.

3.6 Contribution of the dissertation

The five papers of this dissertation differ from each other significantly in terms of the research questions as well as the research designs. Initially, the focus was on case studies and bot detection (papers I and II), as was common in early social bot research. However, the focus of the papers then shifted to experiments on the human ability to qualitatively detect AI-generated bots (papers III and IV), concluding in the attempt to suggest how the methods used in papers III-IV to study social bots could be used more broadly in other domains of information systems research (paper V).

While the five papers are methodologically distinct, in terms of their contributions they are more unified. The first four papers contain empirical contributions to bot research, with papers I and II providing further evidence of the pervasive nature of bots and, more specifically, simpler spambots. In contrast, papers III and IV shift to advanced social bots, demonstrating that humans are incapable of distinguishing which posts and profiles in a social media feed are genuine humans and which are social bots that have been created using generative AI. As empirical contributions, these four papers provide general information on social bots, which could be used to guide future research and could lead to other forms of contributions as well, such as building novel theories on human-bot interaction.

Arguably the most important empirical contributions of this dissertation are in the findings of papers III and IV, as recent literature has questioned the capabilities of social bots and no other papers have yet experimented with generative AI-driven social bots. If the findings, which suggest that humans cannot spot generative bots using generative AI, are further validated by future research and are shown to hold even in more generalized settings, this will have broad implications to the ability of humans to manually label social media data. More specifically, producing human-labeled training data for machine learning-based bot detection will become more difficult, and even qualitatively reviewing the performance of a bot classifier will become challenging.

In addition to the empirical contributions, several of the papers contribute to the research methods used to study bots. Originally, the first paper was primarily a methodological contribution, as it proposed using metadata alone as features in machine learning-based bot detection, but due to Twitter removing affordable access to the API, the methodology presented is now expensive to use. While the contributions of papers III and IV are presented as empirical contributions only, it can be argued that the process for using generative AI to produce content for the experiments in them is a methodological contribution. This contribution is formalized in paper V, which presents how foundation models and generative AI could be used in experiments and quantitative studies in relation to both social bot research and other topics. Moreover, in paper V, the principles for using foundation models and generative AI could contribute to best practices for using these technologies.

As a deviation from the current trends in information system research, the papers do not contain theoretical contributions. This omission of theory and theorizing is justified by the novelty of the topics under investigation and by the goals of the papers, which were to validate the existence of a phenomenon empirically or to provide methodologies to study a phenomenon rather than to build theories of how a phenomenon works. Thus, the papers lay the groundwork for future contributions, which may be theoretical. The following point further supports the emphasis on empirical rather than theoretical contributions. In a recent editorial on artifactual and empirical contributions in information systems research Ågerfalk and Karlsson (2020) state:

"Empirical statements can be used to validate theoretical statements. Empirical statements can also constitute an empirical contribution if they provide "a novel account of an empirical phenomenon that challenges existing assumptions about the world or reveals something previously undocumented. It follows that researchers could make contributions other than theoretical ones and journal editors and reviewers need to reconsider their single-minded focus on contribution to theory."[14]

Following this line of reasoning, the empirical contributions of this dissertation, particularly regarding the human inability to detect AI-generated social bots in papers III and IV, should be sufficient on their own without contributions to theory. However, the importance of building theories related to social bots should not be downplayed and will be further discussed in section 3.7 as an opportunity for future research.

Besides contributing to the academic study of social bots, the papers were intentionally written with non-academic audiences in mind, and one of the goals of the studies was to have an impact outside as well as inside academia. Assessing the real-world impact of the papers is difficult, but

[14] The original full quote contains multiple references, some with page numbers that were removed from this quote for clarity and brevity.

evidencing their relevance, the studies presented in this dissertation have been cited in magazine and news articles both locally (Kailio, 2023; Magnussen, 2023) and internationally (Casillas, 2023; Hopkin, 2023). Moreover, the findings of paper I were cited in a NATO report on how bots and trolls on social media are being used to influence Finnish politics (Van Sant et al., 2020). Thus, one of the primary contributions of the dissertation's research is an increase in public awareness of bots in social media through dissemination via print and online media.

3.7 Future research

This final section of the chapter outlines the future research that could build upon the findings of the dissertation papers. As Twitter is no longer viable as a data source, and because of the uncertainty this has introduced to the future of social bot detection research, the potential for future research related to papers I and II will not be addressed here. Indeed, since the most forward-looking and thus the timeliest contributions to social bot research can be found in the three most recent papers (III–V), the discussion will focus on them. More specifically, I will first discuss how the experiments on the human ability to detect generative AI-based social bots could be extended and will then discuss how the proposed use of generative AI and foundation models as part of existing research methods can be further developed and how their usability can be evaluated.

Papers III and IV demonstrate that in a constrained context where an evaluator can see only what is visible in a social media feed,[15] it is not possible to distinguish genuine profiles from social bot profiles that have AI-generated profile pictures and posts. However, in the real world, it is possible to find much more data about an account one encounters in a social media feed by going to the profile and viewing additional information as well as previous posts made by the account. Thus, it should be possible to base the classification of an account on this much larger set of information. Therefore, a natural extension to increase ecological validity would be to build a more in-depth social media environment into an experiment, allowing participants to view additional information on profiles such as the bio[16] or even a list of tweets made by the same account. While similar mock social networking sites have been used in other research (Wang et al., 2013), the profiles were not created using AI. Ultimately this type of study would reveal if a social bot whose posts and biographic information are generated with a large language model could be stylistically consistent enough to avoid suspicion.

Another extension to the experiments of papers III and IV would be to study if bot profiles in a social media feed can effectively change the perceived majority opinion in online discourse. In other words, an experiment could be conducted where participants are shown a feed consisting of

[15] "What is visible in a social media feed" is illustrated in Figure 5.

[16] Typically, a social networking site allows users to write a short bio, which is visible when visiting the profile, or on some platforms, it is even shown when hovering with the cursor over the name of a profile. The bio itself can contain anything from personal information to emojis and professional or political affiliations of the person behind the account.

a large number of both humans and bots discussing a specific topic, and the participants could then be asked to state the majority opinion. While papers III and IV focus on empirical contributions, this approach would be suited for contributing to theory, with one possible example being spiral of silence theory (Noelle-Neumann, 1974). Previous studies have shown via agent-based modeling (or in other words through simulation) that a very low number of bot profiles can tilt the opinion climate in online discussions (Ross et al., 2019) and that testing this with human subjects through an experiment could further validate the claims. This would further strengthen the evidence showing that social bots can worsen the spiral of silence, leading to genuine users no longer expressing their opinions online if they differ from the perceived majority—even if this perception of the majority's opinion is actually created by bots. Furthermore, such an experiment could be conducted both with social bots similar to those described previously as well as simpler bots to determine if even crude and easily identifiable bots can shift the perceived majority opinion.

As a final extension to the research contained in papers III and IV, it could be insightful to determine if humans could be trained to detect social media profiles that contain elements generated using AI. Similar experiments have been conducted on the human ability to detect deep learning-generated images and LLM-produced texts. Recent studies have shown that LLM-generated texts are practically undetectable even by trained evaluators (Clark et al., 2021; Dugan et al., 2023), but with images, the situation is less clear. The study that was the main inspiration for papers III and IV concluded that humans cannot detect deep learning-generated images even after training (Nightingale & Farid, 2022), but other works have suggested that generated images are in some cases identifiable, and in academic and non-academic studies, bot communities have been detected based on profile pictures and other cues (Bond, 2022; Goldstein & DiResta, 2022; Strick, 2021). Thus, care would need to be taken to design an experiment in a way that goes beyond merely an exercise related to spotting GAN images. One possible approach would be to train participants to look for consistency between posts in addition to checking profile pictures and biographic information.

A second clear area of future research related to the dissertation research stems from paper V. Paper V is an editorial proposing multiple ways that foundation models and generative AI can be used to augment existing research methods, but due to the novelty of these deep learning methods, many of them are still underdeveloped, and there are few examples of papers where they have been used. Paper V proposes that generative AI could be used to create realistic content for experiments such as texts or images that would replace the use of real photos. While previous studies have shown humans cannot distinguish between human- and computer-generated content, it would be beneficial to validate these studies with further studies showing that replacing content with computer-generated content does not affect or bias the results of studies where they are used. For instance, as generated content tends to represent highly "typical" or average examples of the data on which they have been trained, would they have less variance and thus be perceived differently than more heterogeneous genuine data and thus ultimately lead to different results than similar experiment conducted without generated content? As a more concrete example, if an experiment contains images of humans (e.g. a study on how the gender of a customer service chat worker’s avatar affects people’s satisfaction), generated human faces may have less variation and radical features than genuine photos of human faces, and it is uncertain whether such differences could affect results.

Finally, further studies are needed to ensure the safety of paper V’s proposal that foundation models and generative AI could be used to augment and or mask sensitive data by creating new data points that closely resemble existing data points. This technique has been already used as a way to extend training data—for instance, in the context of computer vision—to create larger training datasets (Trabucco et al., 2023), and recent papers have similarly suggested that text data used to train models can be augmented through generating more variations with a large language model (Chung et al., 2023; Hartvigsen et al., 2022). However, there is still scarce research on issues that may arise when using generated data alongside genuine data or as a complete substitute when augmenting image training data for machine learning models. Moreover, the viability of generative AI as a masking tool needs to be further evaluated to ensure that it can sufficiently change the points to make them difficult to connect to the exact original training data.

The proposed future research discussed in this section focuses purely on studies that would build upon the research in dissertation papers III–V. Broader and more speculative suggestions for future research on social bots will be discussed in in the epilogue, which is the final essay of this dissertation.

4 Conclusion

This chapter concludes the dissertation by first outlining the main insights of each of the three preceding chapters and by highlighting what new knowledge has been gained on social bots from the research presented in this dissertation. Then, the final section of this chapter provides a brief glimpse at the PhD project behind this dissertation and to how it may be further developed in the future.

4.1 Summary of chapters

Chapter 1 started with a brief introduction summarizing what social bots are, how have they been studied, and why they matter to society. The introduction provided an overview of social bots and the different roles that studies have attributed to them, ranging from being sinister autonomous agents that are capable of influencing opinions online to being mere benign or even benevolent curators of information. This was followed by a more precise definition of social bots and an explanation that this thesis concerns only social bots found in social networking sites, omitting corporate social bots and chatbots. The rest of the first chapter focused on providing practical information on the research objectives and the structure of the dissertation.

Chapter 2 was written in the form of a commentary essay and reviewed the history of both social bots and social bot research. The essay proposed that the history of social bot research can be divided into four approximate periods from 2010 to the present. The first period started around 2010, when the first publications on bots in social media started to appear, and lasted until 2014, when public interest started growing. This period can be described as the "early days" of bot research, when the term "social bot" was not yet widely used. The second period, called the rise of social bot research, lasted from 2015 to 2018. During this time period, several seminal works in social bot research were published and public and academic interest in social bots grew rapidly. This resulted in a surge in the number of publications describing bots attempting to influence elections, which in turn fueled perhaps unwarranted concerns regarding the capabilities of social bots. The hype surrounding social bots resulted in the "arms race" period of social bot research, when bots were constantly evolving alongside proposed detection methods, which became increasingly intricate as time went on. Meanwhile, criticism regarding some of the prevailing assumptions regarding the sophistication of social bots as well as the quality of bot detection tools such as the Botometer started to appear, resulting in contradictions that have yet to be resolved. The third period started around 2019 and ended in February 2023, when Twitter removed the primary source of social bot data by making the platform's API prohibitively expensive for researchers. This changed the course of social bot research and, given that the fourth period has only recently begun, it is still too early to assess how the field is moving forward.

Chapter 3 presented the contributions of the five articles of the dissertation and discussed the overall contribution this dissertation makes to social bot research. At the beginning of the chapter, each of the five articles were presented separately through extended abstracts that highlighted the background, research objectives, main findings, and limitations of the studies. The discussion on the overall contribution points out that the dissertation is heavily focused on empirical contributions, with a limited number of methodological contributions, but notably did not seek to develop theory, which should be done in future research once the existence of a phenomenon has been established through empirical studies. Furthermore, due to the rapidly evolving nature of social bot research and the issues related to Twitter's API, some aspects of papers I and II have become prematurely dated; therefore, at this point, the most important contributions can be found in the more forward-looking papers III, IV and V. This chapter ended with the discussion of future research that could build upon the findings of papers III–V.

The remainder of the dissertation contains the current chapter, which concludes the formal part of the dissertation. However, a fifth chapter, the epilogue, follows, providing a short speculative essay on the future of social bot research.

4.2 Contributions to social bot research

Chapter 1 presented three research questions. The first research question asked: "How can bots be detected and studied in specific contexts with computationally light approaches?" Paper 1 demonstrated that a light metadata-based approach was sufficient for detecting simple bots. The use of this approach revealed a large community of simple bots following Finnish politicians on Twitter. Similarly, paper II presented a mixed methods approach relying on the Botometer and network analysis to study bots spreading COVID-19-related misinformation and showed that bots in this domain were seemingly simple and not describable as advanced social bots. The methods of both studies became challenging to reuse or build upon when Twitter stopped supporting free API access to the degree required by the papers' methods. Both papers' findings suggested that Twitter has a much higher share of bots than was publicly claimed when the papers were published. Thus, the first two papers ultimately contain empirical contributions that add to the general understanding of the pervasiveness of bots on Twitter.

The second research question asked: "Can humans detect AI-generated bot profiles on social media?" Answering this question was the main objective of papers III and IV. Most academic research on social bot detection has focused on proposing machine learning-based solutions to classifying accounts, largely ignoring the study of how well humans can detect bots via qualitative assessment. Thus, this is one of the most important contributions of the dissertation, as it addresses an identified gap in the literature. Based on the experiments conducted in the two papers, AI-generated bot profiles have become indistinguishable from genuine profiles when shown individually or within a social media feed to a human evaluator. The main implication of this finding for social bot research is that creating training datasets for machine learning via human labeling models is becoming unreliable. Furthermore, qualitatively evaluating the outputs of a bot detection model will become more difficult if social bots that use generative AI are too difficult to distinguish from genuine profiles belonging to humans.

The third and final research question asked: "How can generative AI be used to augment existing research methods?" Based on the methods used in papers III and IV and other recent publications and preprints, paper V proposed ways that foundation models and generative AI can be integrated into existing research methods.

Overall, the dissertation predominantly focuses on empirical contributions and also provides limited methodological contributions to the study of social bots. The most important empirical contribution is the finding suggesting that humans cannot detect bots that have profiles and posts generated by the latest generative AI methods. The significance of the methodological contribution is difficult to determine, as paper V was published in January 2024; thus. It remains to be seen whether the proposed methods will be adopted or further developed.

Because of the recent news media interest in social bots, as well as the relevance of the findings of papers I–IV for the general public, the dissertation research has already received a fair amount of media attention in Denmark and Finland as well as internationally. In particular, paper I has been cited in a NATO report and paper III has been picked up by multiple online magazines around the world. Thus, in addition to the academic contributions made by the dissertation research, another important contribution of this thesis is that it has raised public awareness of bots through dissemination of the findings via news media and magazines.

4.3 The past, present and future

The groundwork for dissertation paper I began in the fall of 2018 as I began working on my master's thesis and experimenting with bot detection methods. The master's thesis was eventually further developed into a conference paper, which became dissertation paper I. While the paper was published in the HICSS proceedings in January 2020, my PhD project at CBS began 11 months later in November. Two years later, dissertation paper II was also published in the HICSS proceedings. At this point, it was becoming clear that the remainder of the PhD project needed to be devoted to something more forward-looking, rather than more backward-looking case studies on bots in a particular context featured in papers I and II.

Fortunately, the developments in foundation models and generative AI brought an interesting new angle to study, which was the human ability to detect bots that harness these new and advanced technologies. The first test builds of bot profiles populated with deep learning-generated content were created early in 2022. We ran a pilot study with these newly created fake profiles and the results were promising: most participants failed to identify the bots, and in some cases, our evaluators mislabeled some human profiles as bots more often than they identified the best-performing bots as bots. Thus, we conducted the first proper experiments in the spring of 2022, shortly after these trials were completed. The results were then published once again in the HICSS proceedings in January 2023. To address the limitations and to further confirm the results, work on the revised version of the experiments started while paper III was still under review. The newer and more robust experiments again resulted in the same conclusion: humans seem to have lost the ability to distinguish genuine photos from generated ones and to identify whether a post was written by a human or by a large language model.

While I believe that the majority of social bot detection research will, for the time being, continue to focus on machine learning-based detection methods, there seems to be much to still explore in terms of human capabilities to perform the same task. Based on the findings of papers III and IV, we now know that when only a limited amount of information is available, detecting bots is difficult if not impossible for human evaluators. But we do not know whether this would be the case if the evaluator were given access to the full history of an account's posts and additional biographic information. Further, given the significant amount of time I have spent looking at social media profiles and identifying bots, would it be possible to turn this experience into training materials that could be administered as a treatment to experiment participants to potentially enable them to label accounts as bot or human with more accuracy? I hope to find answers to these questions in my future work following the completion of my PhD. At this point, I have been researching social bots for over five years. I believe there are easily another five more years to go before my research interests take me elsewhere.

5 Epilogue

This chapter is a speculative essay on the future of social bot research in an era where Twitter is no longer a data source and "we are hurtling towards a glitchy, spammy, scammy, AI-powered internet," as the headline of a recent magazine article suggests (Heikkilä, 2023).

I believe many information technology researchers are proponents of open source software and would frown upon the thought of our research being dependent on a handful of corporations, especially when we must trust that these corporations would go against their own interests to provide unaltered raw data—data that may indeed reveal inconvenient truths of the state of social media, which has consistently proven to be darker than what companies such as Meta and X are likely willing to admit (De Guzman, 2022; Gayle, 2021; Swaine, 2018). Yet when studying social bots, we have been doing exactly that, opting to use Twitter almost exclusively and building an ecosystem of tools around its API, essentially putting ourselves at the mercy of the platform. Indeed, rarely has a field of study been so dependent on one privately controlled source of data as social bot research was on access to Twitter's API. And this reliance backfired spectacularly when a mercurial billionaire purchased the platform and decided to cut off academic access, unraveling years of work on research methodology and sending researchers back to the drawing board (Calma, 2023).

Retrospectively, it is not difficult to understand how we got ourselves into this situation. As a researcher with a limited amount of time to produce a convincing number of publications to reach the next career stage, following the surrounding research trends of using Twitter's API or other productized tools such as the Botometer would have appeared to be the most prudent decision. Attempting to do something more ambitious such as building a robust scraper or going after a less studied platform would have been a time-consuming and potentially fruitless endeavor (Bobrowsky, 2021; Brandom, 2021) whereas incrementally building on existing research would have seemed like a safe way of pursuing a valuable contribution to the understanding of social bots.

Relying on the Twitter API and other tooling built on top of it also had other benefits. It created stability and made reviewing work easier, as it was reasonable to assume that a tool used by thousands had been rigorously tested and worked as intended—and one could safely assume that any problems would be noticed, reported, and eventually patched. In the post-Twitter era, if we all return to using custom tools built for each individual researcher or research group, finding errors such as small mistakes in the logic of a script will be much more difficult and time-consuming and could potentially skew results. Moreover, comparing results between papers and conducting meta-analysis will be much more challenging if everyone has to build their own tools for collecting data.

Despite the temporary shock brought by the death of Twitter as a data source and the ecosystem of tools built around it, I believe we are far from the death of this field of research. Instead, this may even be a turning point that revitalizes social bot research, resulting in more methodological richness and diversity in terms of the social networking sites studied. Given the challenges related to simply collecting data, this will hopefully result in more careful consideration and research, compared to the practice of simply continuing the "arms race" and making incremental improvements to prior bot detection research.

Moreover, as there is no competitive advantage to focusing on Twitter, in the near future, we will ideally see more papers that present findings on understudied but popular platforms. Although Twitter has been extremely popular among researchers, it has never been the social media of choice for most people, and the platform never reached the userbase size of some of its competitors. Especially among younger demographics, platforms such as Instagram and TikTok are much more Popular. While researching them would thus perhaps be more relevant (Vogels et al., 2022), research on bots (and other forms of computational propaganda) on these platforms is sorely lacking.

Focusing on new platforms will have its challenges. The type of content that serves as the primary form of communication on social media platforms can vary significantly. For instance, Twitter-related research was based heavily on analyzing text data, as most tweets contained text, although sharing images, links, and videos was also possible. But on Instagram, the primary form of content is images, which not only convey messages visually but may also contain text embedded within them rather than in a more easily accessible text field. Moreover, apps have increasingly been offering short videos or "reels" as a feature, which are a highly popular form of content on Instagram and other Meta products. Other popular social networking platforms such as TikTok has always focused on short videos, while Snapchat has its own quirks, as it was meant for content that is only temporarily visible. Thus, analyzing content that is not in text form in a text box, as is the case with images or videos, will clearly be more difficult.

However, beyond the initial hurdle of collecting and analyzing image or video data is the potential for a wide range of valuable insights and information on both bot and human behavior as well as on the patterns of human-bot interaction—for example, regarding what type of content is distributed by bots, what type of content spreads well, and what types of bots pass undetected by humans as well as platforms. While researchers have been gathering this type of information for years, there is little research beyond research focusing on Twitter and occasionally Facebook. Moreover, some of the previously identified methods, such as those used in bot detection, could potentially be carried over to newer platforms. For example, features such as social media account metadata (e.g., follower-to-following ratios, age of the account, and so on), posting rhythms, and patterns in profile information or posts can all suggest orchestration and the likelihood that an account is a bot. Furthermore, by developing methods to analyze image and video data from social media, researchers could also help tackle a looming problem predicted to cause problems in the near future. This problem is the hypothesized proliferation of deep fakes and other content created using generative AI (Hsu, 2023).

Bot detection will remain a central issue in social bot research because, to study the matter, we need to be able to detect it in the wild. While studying non-text-based social media content will prove challenging and require getting to know new technologies and methods from different areas of computer science, such as computer vision, detecting bots using previously proposed bot detection methods will hopefully not be as difficult. There is a plethora of research on bot detection from the Twitter era, some of which contains design choices that can most likely be ported over to make developing new bot detection models faster than when starting from scratch. Thus, it may even be possible to rerun studies using previously proposed methods or replicate previous findings with existing methods that have been adapted to a new social media platform. Although the efforts needed to rebuild code and come up with new ways to collect data without an API should not be understated, it might still be easier to conduct a study by porting a research design from previous literature to fit a new platform rather than creating everything from scratch.

Given that collecting large volumes of data is difficult without robust methods and tooling such as access to an API, we might also need to accept that future studies will be based on a smaller number of data points. One area of social bot research where not being able to collect data on large numbers of external accounts may not be a major limitation is the study of bot designs, where researchers build and deploy their own bots. This approach has been used previously, especially in earlier social bot studies, but it never became particularly popular, possibly due to the technical expertise needed to build a functioning bot. However, thousands of publicly available scripts posted on Github already exist for various types of social media bots (Kollanyi, 2016), and some of them could likely be adopted for research purposes, again reducing the time and effort needed.

Another potential avenue of future research is a focus on benign and benevolent bot designs. We are already seeing LinkedIn sales bots (Bond, 2022; Goldstein & DiResta, 2022), and researchers have proposed that social bots could be used to deradicalize extremists in online networks (Blasiak et al., 2021). However, a majority of bot research has focused and still focuses on malicious bots. Now that bot detection research will be less viable, there could be increased interest in bots that are not engaged in malicious activities. As different use cases for bots become more common due to advances in the large language models they can use to produce their posts, new research opportunities will arise—from marketing and sales to summarizing information. For example, how does an LLM-powered LinkedIn bot fare when compared to a salesperson doing cold messaging? Does using social bots for marketing or advertising affect a consumer's perception of the company employing social bots?

Social bot research has had its growing pains as an area of research, and the latest setback, the loss of Twitter as a data source, will take time to recover from. However, there remains much to be studied regarding social bots, and in this essay, I provided some possible areas of future research. Due to the lack of cooperation from social networking sites, in the foreseeable future, researchers might be forced to work with methods such as scraping or using other means of data collection that violate the terms of use of the platforms. As social media platforms are not sharing information with scholars, we scholars should at least share our methods with each other and aim to build open-source tooling.

Lastly, since social media platforms have proven to be bad at auditing themselves and disclosing issues to the general public, the role of academics as watchdogs remains important for society.

Thus, it is vital that social bot and other social media research continue to flourish and that researchers continue to develop new methods to study social media phenomena.

6 Reference list

Ågerfalk, P. J., & Karlsson, F. (2020). Artefactual and empirical contributions in information systems research. In *European Journal of Information Systems* (Vol. 29, Issue 2, pp. 109–113). Taylor & Francis.

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media + Society*, *6*(3), 2056305120939264. https://doi.org/10.1177/2056305120939264

Benjamin, V., & Raghu, T. (2023). Augmenting social bot detection with crowd-generated labels. *Information Systems Research*, *34*(2), 487–507.

Bessi, A., & Ferrara, E. (2016). Social Bots Distort The 2016 U.S. Presidental Election. *First Monday*, *21*(11).

Blasiak, K. M., Risius, M., & Matook, S. (2021). "Social Bots for Peace": A Dual-Process Perspective to Counter Online Extremist Messaging. *ICIS 2021 Proceedings*.

Bobrowsky, M. (2021, August 4). Facebook Disables Access for NYU Research Into Political-Ad Targeting—WSJ. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-cuts-off-access-for-nyu-research-into-political-ad-targeting-11628052204

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bond, S. (2022, March 27). That smiling LinkedIn profile face might be a computer-generated fake. *National Public Radio*. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?t=1654174474533

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The Socialbot Network: When bots socialize for fame and money. *ACM International Conference Proceeding Series*, 93–102. https://doi.org/10.1145/2076732.2076746

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2012). Key challenges in defending against malicious socialbots. *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 12)*.

Brachten, F., Mirbabaie, M., Stieglitz, S., Berger, O., Bludau, S., & Schrickel, K. (2018). Threat or opportunity?-examining social bots in social media crisis communication. *ACIS 2018 Proceedings*.

Brachten, F., Stieglitz, S., Hofeditz, L., Kloppenborg, K., & Reimann, A. (2017). Strategies and influence of social bots in a 2017 German state election—A case study on twitter. *ACIS 2017 Proceedings*.

Brandom, R. (2021, August 13). *Facebook shut down German research on Instagram algorithm, researchers say*. The Verge. https://www.theverge.com/2021/8/13/22623354/facebook-instagram-algorithm-watch-research-legal-threat

Calma, J. (2023, May 31). *Scientists say they can't rely on Twitter anymore*. The Verge. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

Casillas, D. (2023, April 25). La IA dificulta la detección de bots en las redes sociales. *Metro World News*. https://www.metroworldnews.com/noticias/2023/04/25/la-ia-dificulta-la-deteccion-de-bots-en-las-redes-sociales/

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on twitter: Human, bot, or cyborg? *Proceedings - Annual Computer Security Applications Conference, ACSAC*, 21–30. https://doi.org/10.1145/1920261.1920265

Chung, J. J. Y., Kamar, E., & Amershi, S. (2023). Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. *arXiv Preprint arXiv:2306.04140*.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). *All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text* (arXiv:2107.00061). arXiv. http://arxiv.org/abs/2107.00061

Cresci, S. (2020a). A Decade of Social Bot Detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

Cresci, S. (2020b). Detecting malicious social bots: Story of a never-ending clash. *Disinformation in Open Online Media: First Multidisciplinary International Symposium, MISDOOM 2019, Hamburg, Germany, February 27–March 1, 2019, Revised Selected Papers 1*, 77–88.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972. https://doi.org/10.1145/3041021.3055135

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, *80*, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

Davis, C., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *Proceedings of the 25th International Conference Companion on World Wide Web*, 273–274. http://dx.doi.org/10.1145/2872518.2889302

De Guzman, C. (2022, September 29). Report: Facebook Algorithms Promoted Anti-Rohingya Violence. *Time*. https://time.com/6217730/myanmar-meta-rohingya-facebook/

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(11), 12763–12771.

Elyashar, A., Fire, M., Kagan, D., & Elovici, Y. (2013). Homing socialbots: Intrusion on a specific organization's employee using socialbots. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1358–1365.

Emily A. Vogels, Risa Gelles-Watnick, & Navid Massarat. (2022). *Teens, Social Media and Technology 2022*. Pew Research Center. https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

Fernquist, J., Kaati, L., & Schroeder, R. (2018). Political bots and the swedish general election. *2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018*, *January*, 124–129. https://doi.org/10.1109/ISI.2018.8587347

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Gallwitz, F., & Kreil, M. (2021). The Rise and Fall of'Social Bot'Research. *Available at SSRN 3814191*.

Gayle, D. (2021, September 14). Facebook aware of Instagram's harmful effect on teenage girls, leak reveals. *The Guardian*. https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals

Goldstein, J. A., & DiResta, R. (2022). Research Note: This Salesperson Does Not Exist: How Tactics from Political Influence Operations on Social Media are Deployed for Commercial Lead Generation. *Harvard Kennedy School Misinformation Review*, *3*(5), 1–15.

Gorwa, R., & Guilbeault, D. (2020). Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy and Internet*, *12*(2), 225–248. https://doi.org/10.1002/poi3.184

Grimme, C., Assenmacher, D., & Adam, L. (2018). Changing Perspectives: Is It Sufficient to Detect Social Bots? *International Conference on Social Computing and Social Media*. https://doi.org/10.1007/978-3-319-91521-0_32

Grimme, C., Preuss, M., Adam, L., & Trautmann, H. (2017). Social Bots: Human-Like by Means of Human Control? *Big Data*, *5*(4), 279–293. https://doi.org/10.1089/big.2017.0044

Guilbeault, D., & Woolley, S. (2016, November 1). How Twitter Bots Are Shaping the Election. *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv Preprint arXiv:2203.09509*.

Hatmaker, T. (2021, August 4). Facebook cuts off NYU researcher access, prompting rebuke from lawmakers. *TechCruch*. https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/

Heikkilä, M. (2023, April 4). We are hurtling toward a glitchy, spammy, scammy, AI-powered internet. *MIT Technology Review*. https://www.technologyreview.com/2023/04/04/1070938/we-are-hurtling-toward-a-glitchy-spammy-scammy-ai-powered-internet/

Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F., & Stieglitz, S. (2019). Meaningful Use of Social Bots? Possible Applications in Crisis Communication during Disasters. *ECIS 2019 Proceedings*.

Hopkin, G. (2023, March 16). World faces flood of fake people now AI dodges bot spotters. *AI Magazine*. https://aimagazine.com/articles/world-faces-flood-of-fake-people-now-ai-dodges-bot-spotters

Howard, P. N., & Kollanyi, B. (2017). *Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum*. https://doi.org/10.2139/ssrn.2798311

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, *15*(2), 81–93.

Hsu, T. (2023, January 22). As Deepfakes Flourish, Countries Struggle With Response. *The New York Times*. https://www.nytimes.com/2023/01/22/business/media/deepfake-regulation-difficulty.html

Hwang, T., Pearce, I., & Nanis, M. (2012). Socialbots: Voices from the fronts. *Interactions*, *19*(2), 38–45. https://doi.org/10.1145/2090150.2090161

Iivari, J. (2020). A critical look at theories in design science research. *Journal of the Association for Information Systems*, *21*(3), 10.

Kailio, A. (2023, November 6). Näin "kyborgit" levittävät koronavalheita – skriptejä löytyy Githubista. *Tivi*. https://www.tivi.fi/uutiset/nain-kyborgit-levittavat-koronavalheita-skripteja-loytyy-githubista/e529e733-d6ca-4e37-9bb5-ec40c746fac0

Kollanyi, B. (2016). Automation, Algorithms, and Politics| Where Do Bots Come From? An Analysis of Bot Codes Shared on GitHub. *International Journal of Communication*, *10*, 20.

Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*, 312–322. https://doi.org/10.1016/j.ins.2018.08.019

Kupferschmidt, K. (2023, February 16). Twitter's plan to cut off free data access evokes `fair amount of panic' among scientists. *Science*. https://doi.org/10.1126/science.adh0813

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering social spammers: Social honeypots+ machine learning. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 435–442.

Lee, K., Eoff, B., & Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 185–192.

Magnussen, M. (2023, April 3). *Næsten umuligt at spotte en Twitter-bot, viser ny forskning | Radar*. https://radar.dk/artikel/naesten-umuligt-spotte-en-twitter-bot-viser-ny-forskning

Marx, J., Brünker, F., Mirbabaie, M., & Hochstrate, E. (2020). Conspiracy Machines–The Role of Social Bots during the COVID-19 Infodemic. *ACIS 2020 Proceedings*.

Mehta, I. (2023, March 30). Twitter announces new API with only free, basic and enterprise levels. *TechCrunch*. https://techcrunch.com/2023/03/29/twitter-announces-new-api-with-only-free-basic-and-enterprise-levels/

Meske, C., & Amojo, I. (2018). Social bots as initiators of human interaction in enterprise social networks. *ACIS 2018 Proceedings*.

Mitter, S., Wagner, C., & Strohmaier, M. (2014). A categorization scheme for socialbot attacks in online social networks. *arXiv Preprint arXiv:1402.6288*.

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, *24*(2), 43–51.

O'Connor, G., & Schneider, A. (2017, April 3). How Russian Twitter Bots Pumped Out Fake News During The 2016 Election. *NPR*. https://www.npr.org/sections/alltechconsidered/2017/04/03/522503844/how-russian-twitter-bots-pumped-out-fake-news-during-the-2016-election

Oentaryo, R. J., Murdopo, A., Prasetyo, P. K., & Lim, E.-P. (2016). On Profiling Bots in Social Media. In E. Spiro & Y.-Y. Ahn (Eds.), *Social Informatics* (pp. 92–109). Springer International Publishing.

Onuchowska, A., Berndt, D. J., & Samtani, S. (2019). Rocket Ship or Blimp?–Implications of Malicious Accounts Removal on Twitter. *ECIS 2019 Proceedings*.

Pepitone, J. (2010, July 21). Facebook 500 million users reached Wednesday—Jul. 21, 2010. *CNN Money*. https://money.cnn.com/2010/07/21/technology/facebook_500_million/index.htm

Rao, L. (2010, October 31). Twitter Added 30 Million Users In The Past Two Months. *TechCrunch*. https://techcrunch.com/2010/10/31/twitter-users/

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2010). Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv Preprint arXiv:1011.3768*.

Ratkiewicz, J., Meiss, M., Conover, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and Tracking Political Abuse in Social Media. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 297.

Rauchfleisch, A., & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PLoS ONE*, *15*(10). https://doi.org/10.1371/journal.pone.0241045

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Rossi, S. (2022). The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation. *The 55th Hawaii International Conference on System Sciences: HISS 2022*.

Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2023). Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?

Proceedings of the 56th Hawaii International Conference on System Sciences, 134–143. https://hdl.handle.net/10125/102645

Rossi, S., Rossi, M., Mukkamala, R. R., Thatcher, J. B., & Dwivedi, Y. K. (2024). Augmenting research methods with foundation models and generative AI. *International Journal of Information Management*, 102749. https://doi.org/10.1016/j.ijinfomgt.2023.102749

Rossi, S., Rossi, M., Upreti, B., & Liu, Y. (2020). Detecting Political Bots on Twitter during the 2019 Finnish Parliamentary Election. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, *February*. https://doi.org/10.24251/hicss.2020.298

Salge, C. A. D. L., & Karahanna, E. (2018). Protesting Corruption on Twitter: Is It a Bot or Is It a Person? *Academy of Management Discoveries*, *4*(1), 32–49. https://doi.org/10.5465/amd.2015.0121

Salge, C. A. de L., Karahanna, E., & Thatcher, J. B. (2022). Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter. *MIS Quarterly*, *46*(1). https://doi.org/DOI: 10.25300/MISQ/2021/15598

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2020). Detection of Novel Social Bots by Ensembles of Specialized Classifiers. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://doi.org/10.1145/3340531.3412698

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. *arXiv Preprint arXiv:1707.07592*, *96*, 104.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06930-7

Stella, M., Ferrara, E., & Domenico, M. D. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(49), 12435–12440. https://doi.org/10.1073/pnas.1803470115

Stieglitz, S., Brachten, F., & Kissmer, T. (2018). Defining bots in an enterprise context. *ICIS 2018 Proceedings*.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A. K. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. *ACIS 2017 Proceedings*, 1–11.

Strick, B. (2021). *Analysis of the Pro-China Propaganda Network Targeting International Narratives*. Center for Information Resilience. https://www.info-res.org/post/revealed-coordinated-attempt-to-push-pro-china-anti-western-narratives-on-social-media

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting Spammers on Social Networks. *Proceedings of the 26th Annual Computer Security Applications Conference*, 1–9. https://doi.org/10.1145/1920261.1920263

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016). The DARPA Twitter bot challenge. *Computer*, *49*(6), 38–46.

Swaine, J. (2018, January 20). Twitter admits far more Russian bots posted on election than it had disclosed. *The Guardian*. https://www.theguardian.com/technology/2018/jan/19/twitter-admits-far-more-russian-bots-posted-on-election-than-it-had-disclosed

Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv Preprint arXiv:2302.07944*.

Van Sant, K., Fredheim, R., & Bergmanis-Korāts, G. (2020). *ABUSE OF POWER: COORDINATED ONLINE HARASSMENT OF FINNISH GOVERNMENT MINISTERS*. NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/cuploads/pfiles/abuse_of_power_online_harassment_of_fin_ministers_16-03-2021.pdf

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *1151*(March), 1146–1151.

Wagner, C., Mitter, S., Körner, C., Strohmaier, M., & others. (2012). When social bots attack: Modeling susceptibility of users in online social networks. *#MSM2012 Workshop Proceedings*, 41–48.

Wald, R., Khoshgoftaar, T. M., Napolitano, A., & Sumner, C. (2013a). Predicting susceptibility to social bots on twitter. *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, 6–13.

Wald, R., Khoshgoftaar, T. M., Napolitano, A., & Sumner, C. (2013b). Which users reply to and interact with twitter social bots? *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, 135–144.

Wang, A. H. (2010). Detecting spam bots in online social networking sites: A machine learning approach. *IFIP Annual Conference on Data and Applications Security and Privacy*, 335–342.

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2013). Social turing tests: Crowdsourcing sybil detection. *Proceedings of The 20th Annual Network & Distributed System Security Symposium (NDSS)*.

Wang, T., Chen, L., & Genc, Y. (2018). An N-gram-based Approach for Detecting Social Media Spambots. *Proceedings of the 2018 Pre-ICIS SIGDSA Symposium*.

Weatherbed, J. (2023, February 2). *Twitter is replacing free access to its API with a new paid tier*. The Verge. https://www.theverge.com/2023/2/2/23582615/twitter-removing-free-api-developer-apps-price-announcement

Wile, R. (2022, November 17). A timeline of Elon Musk's takeover of Twitter. *NBC News*. https://www.nbcnews.com/business/business-news/twitter-elon-musk-timeline-what-happened-so-far-rcna57532

Woolley, S. C. (2016). Automating power: Social bot interference in global politics. *First Monday*.

Yang, K.-C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. *arXiv Preprint arXiv:2307.16336*.

7 Appendix

The appendix contains the full versions of the five papers of this dissertation. The papers are identical to the form that they have been published in (papers I–III and V) or to the form that they have been submitted for review in the case of unpublished work (paper IV). The only changes have been made to the papers is in typesetting so that they match the rest of the dissertation.

Paper I

The following paper has been published in the proceedings of the 53rd Hawaii International Conference on System Sciences in 2020.

Detecting Political Bots on Twitter During the 2019 Finnish Parliamentary Election

Sippo Rossi
Aalto University

Matti Rossi
Aalto University

Bikesh Raj Upreti
Aalto University

Yong Liu
Aalto University

Abstract

In recent years, the impact of bots used for manipulating public opinion has become an increasingly prevalent topic in politics. Numerous sources have reported about the presence of political bots in social media sites such as Twitter. Compared to other countries, the influence of bots in Finnish politics has received little attention from media and researchers. This study aims to investigate the influence of bots on Finnish political Twitter, based on a dataset consisting of the accounts following major Finnish politicians before the Finnish parliamentary election of 2019. To identify the bots, we extend the existing models with the use of user-level metadata and state-of-art classification models. The results support our model as a suitable instrument for detecting Twitter bots. We found that, albeit there is a huge amount of bot accounts following major Finnish politicians, it is unlikely resulting from foreign entities' attempts to influence the Finnish parliamentary election.

1. Introduction

Nowadays, many organizations and individuals attempt to influence people by spreading propaganda in social media through large networks of bot accounts [1, 2]. There are multiple examples of bots being used to distort political discussions on Twitter. One of the most notable cases is the 2016 US presidential election, where an organization linked to the Russian government has been accused of striving to manipulate the elections by spreading fake news or biased content via Twitter bot accounts [2, 3]. In this light, a number of studies have delved into the detection of bot accounts through developing and testing new bot detection methods. Based on synthesizing key factors for bot detection reported in previous studies, the study developed an integrated framework for bot detection.

Specifically, this study aims to demonstrate how bots that are being used to influence politics on Twitter can be identified using machine learning approaches. To demonstrate the application of the method, we identified the bots that existed before the Finnish parliamentary election in April 2019 using user-level metadata. Noticeably, recent publications have found evidence of bots being used to influence opinions in countries such as the United States [3], Japan [4], Brazil [5] and

Russia [6]. Similar studies have not been conducted in Finland, albeit there is already evidence of at least one large but inactive Finnish Twitter botnet according to a researcher at F-Secure [7, 8]. In other words, our study seeks to answer the following two research questions, including:

RQ1: What are the important features that can be used to identify bots?

RQ2: Do the bots have an impact on Finnish politics?

To answer the research questions, we first develop a model that can predict bots using machine learning methods. Once the bots are identified, we assessed the impact in terms of visibility and popularity of politicians followed by these bots.

This paper contributes to the growing information systems science and political data science literature on the use of bots and information systems to influence voters. The study also adds to bot detection literature by evaluating the feasibility of using a limited set of profile metadata features in a supervised machine learning bot detection model. As a part of the research project to detect bot's effect on ongoing European elections, we deem the study addresses a timely and important topic, as there is evidence of attempts to use bots to influence voters during recent European elections [9, 10].

2. Related research

We analyze the related research in three parts. The first part looks at how previous research has classified bots and provides a clear definition of key terms and concepts. The second part analyzes methods that have been used to detect bots in Twitter-related research and provides a background and benchmarks for the bot detection model proposed. The third and last part covers literature on the use of bots in political influencing during recent years to support the findings and assumptions made.

2.1. Terminology and the definition of a bot

A bot can be defined as an account that is operated fully or partially by a program. Thus, at least some parts of a bot account's activities are automated. Examples of these include bots belonging to like farms that are used on social media to increase the number of followers of an account or likes of a particular post. However, they are prone to detection and thus, deletion. More advanced bots adjust their content dynamically based on the behavior of other accounts, making them more difficult to detect even if the bot is still operated solely by a program. The most sophisticated bots are such that humans control parts of their activities, such as content creation, which blurs the line between the bot and a human user. When properly operated, these hybrid bots are almost invisible to automatic detection mechanisms, according to Grimme et al. [11]. Some bot accounts are inactive, also known as sleeper bots [12]. The accounts are 'quiet' most of the time before being activated e.g. to spread spam.

On Twitter, bots can be divided into benign and malicious bots [13]. The benign bots adhere to Twitter's rules and guidelines and are clearly distinguishable from human accounts usually by name or description. Conversely, malicious bots participate in activities that are not permitted by Twitter and rarely disclose the fact that they are operated by a program. Typical use cases include artificially boosting the number of followers, likes or retweets and directing or blurring discussions as well as spreading spam or content that supports a certain cause. Both types include bots ranging from simple content sharing accounts to human-like social bots that participate in discussions and create original content. The phrase of social bot here refers to a bot that is meant to mimic human behavior [12] by communicating and interacting with human users [14].

2.2. Detecting bots on Twitter

2.2.1. Simple versus complex models. As algorithms that control bots become more advanced, so do the bot detection algorithms. In literature, bot detection models range from the very simple ones that are based on analyzing one piece of metadata to those that use ensemble methods to analyze large feature sets including a mix of metadata, tweeting behavior, and content data.

Past studies on bot detection have been to some extent restricted to bots with a specific feature. For instance, Beskow and Carley [15] managed to identify specific automatically generated bot accounts based on a single piece of metadata, the profile name, with approximately 95%-99% accuracy depending on the algorithm used. However, this type of approach results in a very narrow use and the aforementioned model could only detect bot accounts that have an account name consisting of a randomly generated string of 15 characters and more than likely to miss out the bots with different characteristics. However, as Beskow and Carley [15] propose, a tool-box approach where multiple different models are combined can make even the simple models an important contribution to more advanced bot detection models.

A number of bot detection models looked into various characteristics of accounts by combining metadata and behavior features to identify bots (e.g. [16, 17]). One notable issue hinders the reusability of these models. Many of these models rely on some form of natural language processing, sentiment analysis techniques [14] or a specific list of keywords [6, 16, 18]. This restricts their applicability to a particular language and region as well as an event such as an election, due to certain themes and hashtags being important only in that specific context. Bots have been evolving rapidly during the past few years to a point that they may be difficult even for a human to distinguish them from real users [19]. There is a need to update bot detection algorithms, since a workable algorithm today may prove to be ineffective after a couple of years.

2.2.2. Feature space selection. Machine learning methods represents the key approach used in early bot detection literature, in which an essential aspect of work is to determine the optimal feature space that boost bot prediction performances. There are two main considerations in the selection of bot detection features. Firstly, the features should be added only if they improve the accuracy of bot detection. Secondly, the features must not make the data collection phase overly time-consuming, since Twitter’s API has strict rate limits.

In previous studies, the most common classes of features used in bot detection include metadata-based features and tweeting characteristics-based features [6, 8, 18]. User profile provides a large amount of metadata while tweets offer useful information, albeit being limited to a certain number of characters (280). Based on these features, the amount of analyses that can be performed is vast. Other classes of features, such as keywords, are not included in the analysis, because these features restricted the applicability of the model to a specific event.

Metadata-based features can be divided into two different branches. Intuitively, metadata extracted from a profile gives information on the account, while metadata from tweets gives a combination of information from the profile posting it as well as the tweet itself [20]. Metadata that can be extracted from Twitter include basic profile information such as name, description, and number of friends. An examination on whether different pieces of profile information are blank or at default contributes to a collection of binary features, such as a variable of whether or not the profile picture has been added [6, 16]. The more fields are left at default, the more likely the account is a bot [6]. Data on the number of users that the profile is following, the number of followers and ratios of these are also often used in prior bot detection studies [6, 8, 18, 20]. Profiles that have none or a few followers, but follow many profiles are suspect [8]. Lastly, the contents of the textual metadata can be analyzed and used to classify bots for instance by inspecting the length or frequency of certain keywords in the description or name [8, 15, 18].

Earlier findings suggest that a combination of both metadata and content features yields optimal results [3, 18]. Hundreds of different features can be derived from Twitter's metadata and content data, making it a matter of preference on which ones to choose. Examples include counting the number of hashtags, URLs, and instances of specified keywords in the name or description of an account.

The model proposed in this study utilizes metadata-based features only and therefore, they are examined more thoroughly than content-based and other types of features. Further, unlike tweet content-based features, metadata-based features are more generalizable across different linguistic context. Table 1 illustrates some of the features that have been used in previous papers [6, 17, 18]. Unsurprisingly, the most common features are the ones that are directly related to how Twitter functions, default profile values, with the number of followers, friends, tweets, and retweets being examples of these.

Table 1. Summary of key features for bot detection used in prior literature

Binary features	Profile information features	Ratio features	Metadata content features
Defaults: - Profile image - Background image - No user description **Other:** - Profile verified - Location specified - No friends - No tweets	**General:** - Number of followers - Number of friends - Number of tweets - Number of likes - Age of account - Account language **Length:** - Profile name - Profile description	**Activity:** - Ratio of following and followers (FE/FI) - Reputation (FE/(FI + FE)) - Given likes per friend - Given likes per follower **Account age:** -Friends/Account age - Following rate (FI/AU)	**Bot check:** - Name contains bot - Description contains bot **Other content:** - Number of # in description - Keywords in description - URL(s) in description

In Table 1, the features are grouped into four types. Some of the most commonly used features are found in the first group as binary features. Based on the popularity, it can be assumed that they are appropriate for bot accounts detection despite their simplicity. Binary features are designed to check whether profile customization options, such as the profile image and background image, are left at default [6, 17, 18].

The second group also contains many of the prevalent features in bot detection models. These features are often numerical variables, many of which are related to how popular a Twitter account is and how actively it is used. Particularly, the numbers of followers, friends, tweets, retweets, and likes were often investigated [6, 17, 18]. Another commonly used feature is the length of the description text, which cannot be obtained directly from Twitter but can be calculated easily from the metadata [6, 17, 20].

The third group of features is ratios that can be obtained from the same metadata. When compared to the two previous groups, the ratio features offer more variety as they are not based on Twitter's built-in attributes. Followers-to-friends ratio is a common ratio feature used in many previous studies [6, 18, 20]. In the model created by Fernquist et al. [18], the top features for bot detection include multiple of ratios, with examples being given likes per friend, followers-friends ratio, and number of likes per followers.

The last group consists of the features deriving from the contents of different attributes. Features in this group are occasionally used in earlier studies. Two of the features in this list simply check whether an account is a bot according to the profile description or name by looking if the fields contain the word "bot" [15]. The rest of the features relate to an examination of URLs, hashtags ,or other keywords [20].

Because ratio features were among the best performing features widely used in early studies [6, 18, 20], we include several of them in the study alike.

2.2.3. Classification methods. Because Twitter, like most of the social media sites, actively attempts to detect and disable bot accounts, the creators of bots have responded by making bots behave more like humans. Consequently, the selection of features as well as preparing the training data has become more demanding and for a model to stay up to date, feature engineering and adding new training datasets is needed [20].

Both supervised [6, 15] and unsupervised [16, 21] machine learning models have been used in bot detection research. The drawback of supervised learning is that creating a labeled dataset for training the model either requires a large amount of manual labeling [6] or using a pre-labeled dataset, which may limit the applicability of the model as the datasets most likely represent only a fraction of the possible behavior of bot accounts in Twitter. Unsupervised learning models can detect novel bot behavior that may get past a supervised model [16], as the supervised models can only detect bots that are similar enough to the dataset that was used to train it. However, the results of unsupervised models are more difficult to validate due to the absence of labeled data.

Past studies indicated that supervised models are better suited for analyzing topical datasets that are collected from Twitter's streaming API [6]. Twitter's API allows performing searches and collecting the data on tweets that contain for certain keywords or hashtags, which is particularly useful when analyzing political discourse that is related to a specific topic, such as an election [18, 22]. Since campaigns, political parties, candidates and users use hashtags to make their tweets visible when commenting on specific topics, it is more efficient to mine data on a topical level with the keyword search instead of first collecting a large dataset of Twitter accounts and then analyzing the content of their tweets.

2.3. Use of bots in political influencing

Previous studies illustrated that Twitter-based computational propaganda has been used by organizations and governments across the world [12]. There are several hypothesized goals of the creators of bots. These range from increasing the partisanship of a population or advancing a cause that the creator of the bots supports. [23] noted that "it is an effective non-military means for achieving political and strategic goals." Measuring the successfulness of political bots is difficult as it is hard to quantify the impact that they have had for example, on voting behavior [23]. Nevertheless, the prevalence of computational propaganda campaigns would suggest that they are viewed as a functional tool that does have an effect on the target audience [23].

More measurable and easily achievable targets include manipulating the popularity and visibility of tweets by liking, following, and retweeting content with a botnet. These methods can cause a particular hashtag to trend thus, pushing it higher into the feeds of other Twitter users. Other goals may be to make an opinion seem more popular than it actually is or to bury actual discussions or factual information by making it difficult to follow. Concrete examples include spamming pro-government tweets or flooding search results related to protests with meaningless content making it more difficult for human users to find and participate in discussions [24].

Two earlier studies monitored bot activity in Germany during a state parliament as well as Federal presidential election [10] and federal election during 2017 [9]. Bots represented around 7 – 11% of the accounts and bot-driven content represented 7.4 – 9% of all traffic during the German elections [9, 18]. These are modest numbers and in line with Twitter's estimate of bots accounting for approximately 10% of Twitter activities. The main reason for concern is that the bot activity was skewed towards supporting the alt-right movement and was possibly produced by accounts outside of Germany [9]. As an extreme example, in Russia, Stukal et al, [6] reported that up to 85% of the daily tweets containing political keywords were posted by bot accounts during 2014-2015. Obviously, there are regional differences in the prevalence of bot accounts [12].

Finland, the focus of this research, may also be affected by bot account, considering i) the current growth of Euroscepticism, ii) rise of right-wing political movements alongside Finland's historical relations and iii) proximity to Russia that may provide a more fertile foundation for bot activity than for example Sweden.

3. Methodology

This study employed an unorthodox approach to collect Twitter data as the dataset is compiled from individual accounts' followers, which differs from the more traditional methods by collecting all tweets (and associated account metadata) that use specific hashtags or keywords are gathered through Twitter's Streaming API. This method is appropriate since the proposed bot detection model only requires metadata. The benefits of this approach include that it allows detecting both dormant bots as well as those that do not use specific hashtags or words that the streaming method queries for.

The primary tools used in data collection and formatting phase were the statistical programming language R and its "rtweet" package, which is an "R client for accessing Twitter's REST and stream APIs."

The study analyzed the Twitter accounts of several politicians and their followers on Twitter. The profiles were selected based on several heuristically chosen criteria to ensure that as many political parties as possible were represented and that a sufficient amount of data was collected. At least a member of parliament was taken from each of the current coalition parties as well as from all parties that have support of over 5%, albeit a maximum of two per party. Furthermore, only accounts with over five thousand followers were picked. Lastly, some prominent politicians with over 10 thousand followers were selected even if they do not match the other criteria as several influential political figures would otherwise be excluded. Table 2 shows the selected politicians.

Table 2. Summary of selected politicians

Name	Political party	Username	Followers (K)*
Alexander Stubb	National Coalition Party	@alexstubb	370
Sauli Niinistö	National Coalition Party	@niinisto	159
Juha Sipilä	Centre Party	@juhasipila	126
Anne Berner	Centre Party	@AnneBerner	21.7
Pekka Haavisto	Green League	@Haavisto	130
Ville Niinistö	Green League	@VilleNiinisto	84.3
Paavo Arhinmäki	Left Alliance	@paavoarhinmaki	109
Li Andersson	Left Alliance	@liandersson	76.5
Antti Rinne	Social Democratic Party	@AnttiRinnepj	25.6
Sanna Marin	Social Democratic Party	@MarinSanna	14.3
Jussi Halla-aho	Finns Party	@Halla_aho	14.5
Laura Huhtasaari	Finns Party	@LauraHuhtasaari	13.7
Sampo Terho	Blue Reform	@SampoTerho	7.6
Paavo Väyrynen	Seven Star Movement	@kokokansanpaavo	10

*Number of followers at March 2019

The sample consists of 14 politicians from 8 different parties ranging from liberal to conservative and left-wing to right-wing, including the current president and three ministers as well as 6 party leaders. Many small parties were left out by this approach, of which respective politicians did not have accounts or had much fewer followers. As a result, over 1.1 million Twitter accounts were collected as followers of these politicians, but the number was reduced to approximately 550,000 after filtering out duplicate accounts. The duplicates were a result of the fact that many Twitter users were following multiple selected politicians.

The binary features were calculated from corresponding attributes where a setting left at default or blank equals 1 and a nondefault 0. The ratio features were created similarly by calculating the values from the profile information metadata and then placed into new columns. Ratio calculations that resulted in NaN (not a number) or Inf (infinite), were replaced with a zero.

3.1. Creating training data

Based on the findings of previous research and datasets available for training the model, a selection of 11 features was picked for testing the first version of the model. The feature space consists of four binary features, four profile information features, and three ratio features.

Initially, the model was trained with the cresci-2017 dataset [25], which contains over 13,000 labeled accounts divided into groups of social spambots, traditional spambots, fake followers, and genuine accounts. The training dataset was balanced to include 3,000 randomly sampled bot accounts and 3,000 randomly sampled genuine accounts from the cresci-2017 dataset.

To find a suitable algorithm for the prediction, the Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), K-nearest neighbors (KNN), Support-vector machine (SVM) and Random Forest algorithms were tested. Out of these Random Forest performed the best, although there were signs of either the training data not representing the variety of real data or that the model being overfitted as the accuracy was over 97% or 98% on most runs. This issue was ignored, as the model was deemed sufficiently accurate for the first phase where the goal was mainly to speed up the training data creation by manually validating list of potential bot accounts from the prediction results. The model was then tested on a sample of 5,000 accounts from the dataset that was collected for this study.

After manually inspecting on Twitter the accounts that the model labeled as bots, it was evident that the model had difficulties distinguishing bots and genuine accounts. Particularly, accounts that were apparently created by people trying out Twitter without becoming active users were prone to be labeled as bots due to the similarity in the account behavior. In most cases, the easily distinguishable bots were following approximately 20-100 accounts, had 0-2 followers and little to no tweets, retweets or likes.

Based on the performance of the first version of the model, it was apparent that the cresci-2017 dataset was unsuitable for training a model that could accurately distinguish bots from humans based on metadata. One possible explanation for such performance is the fact that the training data used in the model had only very clear examples of bots and genuine accounts, where the behavior in terms of tweets, retweets, likes and ratios of followers and following differed widely depending on whether the account was a bot or not. However, this does not reflect the actual behavior of accounts where in some cases even with quantitative and qualitative assessment it is difficult to label an account accurately as either a bot or a human.

By manually labeling a set of accounts from the dataset consisting of followers of the Finnish politicians, a new training dataset that represents the actual distribution and behavior of the accounts of the target dataset was created. The training data was created by checking and verifying the accuracy of 2,000 accounts predicted to be bots by the first model. The results were that out of these accounts 1,336 were accurately labeled as bots, as they were either bots or accounts exhibiting extremely bot-like behavior while 664 were actually humans or accounts that were impossible to determine as belonging to either group.

A qualitative approach was employed for classifying the accounts as either bots or humans. The classification started by inspecting the profile information of the account. Common signs of a bot were the name or description of the account, which often included Russian or Arabic and or a seemingly random string of characters and numbers coupled with the account following 21 other Twitter users, which is the default number of recommended users to follow given by Twitter when creating a new account. Other possible predictors included in this step are the profile image and banner as well as the age of the account. As a second step, the tweets and retweets were checked when available to see what kind of activity the account has and what other accounts it interacts with. As the third step, the accounts that the possible bot was following were inspected to find discrepancies. For example, a user following mainly seemingly random foreign accounts coupled with one Finnish politician or if it was following exactly 21 very popular Finnish accounts were usually the best predictors of an accurate classification as a bot even though the machine learning model did not look for these. If after the three first steps the account was still too ambiguous for classification, the likes and followers were checked for bot-like behavior.

During this process, several interesting findings were made, which can be used later in the analysis of the whole dataset. Firstly, most of the bot accounts were dormant as well as possibly a part of a follower boosting operation. Secondly, most of the bots were difficult to label as political bots as it is not sure whether they were created to boost the followers of a particular politician or if it followed them by coincidence based on Twitter's recommendations. Commonly, shared characteristics among bots included that they barely engaged with content or interacted with other users and that they followed a random group of 21 accounts, which most likely are those suggested by Twitter during the creation of the account [7, 8]. Peculiar accounts that they often followed included less well-known US politicians, an obscure game called Growtopia and a niche Finnish newspaper called Markkinointi and Mainonta.

3.2. Building the bot prediction model

The second version of the bot detection model differed from the previous one mainly in how the splitting of the training and validation data was done, what parameters and algorithms were used as well as how many features were included.

New features could be added to the second version of the model as the training data was no longer a limiting factor. By including the age of the account, and two ratio features derived from comparing the profile information to the age of the account, the number of features was increased from 11 to 14.

Several variants of the Random Forest algorithm were tested, but the standard version still performed optimally and was selected for the final model. The model was trained with a randomly sampled set of 500 bots and 500 humans from the new manually labeled dataset. The remaining

1000, with 836 bots and 164 humans, were used in the validation of the performance. The final version of the bot detection model has an accuracy of 83% with only slight changes after multiple runs and small variations in parameter settings. Table 3 lists the most important statistics for assessing the performance.

Table 3: Performance of the bot detection model

Metric	Value
Accuracy	0.837
Recall	0.846
Specificity	0.793

In terms of feature importance, the top features were a mix of profile information and ratio features, while the binary features were all in the bottom half of the feature ranking. Table 4 contains the full ranking of the features. Based on this, the model gives much weight to the number of accounts that an account is following, since the two top features are related to the following attribute. This is somewhat problematic for our overall goal of political bot detection as it implies that the model is best at detecting dormant bots and bots belonging to follower farms. These accounts can be political bots, but in many cases determining if they are following politicians on purpose or by coincidence is difficult. This is because the popular politicians often appear on the top of the recommended accounts to follow in Finland.

Table 4: Features ranked

Rank	Feature	Importance
1.	Following	100.00
2.	Following to age of account	61.00
3.	Age of account	56.57
4.	Followers to following	22.87
5.	Likes to following	15.68
6.	Tweets	15.18
7.	Likes to age of account	11.95
8.	Followers	10.05
9.	Default profile image	7.11
10.	Likes	6.34
11.	No description	4.72
12.	No location	3.61
13.	No banner	2.56
14.	Likes to followers	0.00

4. Findings and discussion

4.1. The proposed bot detection model

The bot detection model proposed in this study demonstrated that metadata alone is sufficient for classifying at least spambots and bots that belong to follower farms. The primary benefit of a model based on metadata is that the data collection is much quicker as 90,000 accounts' information can be retrieved every 15 minutes. Therefore, a model that uses metadata works particularly well when studying countries that have a small population, since then even the most popular Twitter users are likely to have a manageable number of followers. In other words, due to the limited number of users in these countries, it is possible to gather comprehensive datasets for analysis in short periods. Furthermore, analyzing entire populations instead of samples is feasible with a purely metadata-based model, contrary to models that use tweet data, where the number of accounts to analyze is restricted by Twitter's streaming API's rate limits.

Regarding the selection of the feature space and algorithm, most of the results were in line with the reviewed literature, although some of the results were surprising. Random forest was the optimal classification algorithm, which was the result in several other models as well [18]. While ratio features had high feature importance as suggested by previous research, the binary features did not despite their popularity in earlier models. Overall, the performance of the model was below most of those listed in the literature review, but as stated earlier direct comparison is difficult due to the differences in the goals of the models.

4.2. Bots Counts in Finnish political Twitter

Based on our bot detection model, we predicted the total number of bots in the dataset consisting of the 558,983 followers of the 14 Finnish politicians was formatted to match the training dataset. The model predicted that out of the dataset approximately 36.6% are bots. Since the model's accuracy is 83%, out of the 204,426 accounts classified as bots it can be assumed that 169,673 should be the real number of bots when not taking into consideration the accounts labeled as humans that in reality, are bots. Therefore, the percentage of bots in reality is likely to be closer to 30% based on the results and the accuracy of the model.

4.3. Influence of Bots in Finnish political Twitter

Overall, the findings of the study do not support the notion that Finland and Finnish politics would be the target of internal or external bot influencing campaign, due to most of the bots having almost no activity besides following popular accounts. This finding is line with a recent announcement made by Supo, the Finnish Security Intelligence Service, which stated that it has not found evidence of foreign entities attempting to influence the elections [26].

Despite few political bots, over 150,000 bot accounts following Finnish politicians on Twitter were identified. Although these bot accounts do not interact much with other accounts, they still help the politicians that they follow by two ways. Firstly, they artificially inflate the number of followers a politician has making them possibly more popular than they actually are. Secondly, they help increase the visibility of politicians, since being followed by many promotes an account over other less popular accounts in Twitter's "who to follow" suggestions. Consequently, bot accounts that were created for an entirely different purpose may unintentionally follow politicians when they follow their accounts based on Twitter's recommendations.

The primary impact that the bots have on Finnish political Twitter is related to increase the visibility and perceived popularity of the politicians' accounts. Considering a low utilization of Twitter as a medium for political debate in Finland, the possible effects the bots that may have had on voters may be negligible. Nevertheless, one metric for measuring a politician's popularity that can be used to predict election results is how many followers they have on different platforms and how much their audience engages with them [27]. Therefore, even if the impact on actual voting behavior is minimal, the presence of bots may manipulate perceptions, influence predictions and damage the validity of social media engagement as an indicator of actual popularity.

When inspecting the scores of individual politicians, Pekka Haavisto and Alexander Stubb had the highest percentages of bot followers, with both at above 30%, which is beyond Twitter's own estimates of 5-10% accounts being bots. The strong bot presence in Haavisto's Twitter follower base was subject to debate already in 2017 during his presidential election campaign [28]. Previous analysis attributed the bot followers to a result of a sudden increase in bots promoting the game Growtopia and Twitter's recommendations boosting Haavisto, which is similar to the findings of this study.

Alexander Stubb, the other notable example of a politician benefitting from the added visibility, has acquired the largest absolute number of bot followers. Many of the bots did not follow any other politicians besides Stubb, which is likely due to his strong presence in Twitter as the 3rd most followed account in Finland.

Contrary to findings elsewhere [4, 9], the candidates most likely to be linked to the Finnish alt-right movement Laura Huhtasaari and Jussi Halla-aho had the lowest percentage of bot followers. However, this is not surprising when taking into consideration that they also have the lowest number of followers from the sample of accounts inspected, which means that they do not attract bots that follow accounts by default based on Twitter's recommendations.

5. Conclusions

The goals of this study were to develop a new supervised machine learning bot detection model to investigate if Twitter bots were used to influence the 2019 Finnish parliamentary election and to test a new approach for Twitter bot detection. The developed model was used to estimate the number of bot followers that a sample of the most popular Finnish politicians have in their follower base.

The dataset used in the study consisted of 550,000 unique accounts out of which roughly 169,600 were classified as bots. The metadata-based model was found to be feasible for classifying bots on Twitter and the predictions of the model were used to assess if bots were utilized during the 2019 Finnish parliamentary election. The findings imply no evidence of attempts to influence the elections via Twitter bots. Although the bots increased the visibility of some politicians and made them seem more popular, the bots are unlikely to have had much effect due to their passive behavior.

This study holds important implications for both the researchers as well as practitioners. Our study explores a number of primary meta data-based features as well as ratio-based profile features to predict bots in Twitter. This approach provides better coverage of the profile characteristics, and it is generalizable to a wide variety of context due to the linguistics independence.

To the best of our knowledge, this study is the first in studying the presence and influence of bots in a Finnish context. Our results imply that the bots are surfacing in the Finnish domain. Even though we did not find the bots to have a significant impact, we cannot predict how this could change in the future. These results should be of interest not only to researchers, but also to politicians and users of social media in Finland.

Lastly, our results also highlight the influence that Twitter's suggestions can have on the number of followers that popular accounts have. These results indicate that profiles followed by bots are likely to attract more bots, further inflating their number of followers and perceived popularity.

5.1. Limitations

Like all studies, our study also has limitations. First, the approach used in the selection of politicians and data collection phase as well as the choice of features in the machine learning model introduced some constraints to the analyses that could be performed. Our sampling approach ignores the user accounts that do not follow politician yet remain politically active. Although it was possible to determine if an account is a bot based on metadata, the collected data did not enable examining the content that they interacted with or spread via tweets, retweets, and likes. However, it is worth noting that most of the bots detected are not actively creating or distributing content. Lastly, politicians with much higher or lower percentages of bot followers may have been omitted from the sample.

5.2. Suggestions for further research

To further understand the use of bots in the Twittersphere, the model could be reused during future elections by collecting new datasets. This would be particularly interesting due to the Finnish Security Intelligence Service's suggestion that the EU elections are likely to be a more attractive target for external influencing attempts than the Finnish parliamentary election [26].

To analyze the efficiency of Twitter's own bot detection and removal practices, the rate at which accounts labeled as bots are removed from the social media site can be followed. In addition, changes in the activity of the bots can be monitored by inspecting how the attributes such as a number of tweets and likes changes over time. Especially interesting would be to find evidence if some of the accounts were sleeper bots waiting for activation.

6. References

[1] D. W. Nickerson and T. Rogers, "Political campaigns and big data," *Journal of Economic Perspectives,* vol. 28, no. 2, pp. 51-74, 2014.

[2] Twitter. "Update on Twitter's review of the 2016 US election." https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html.

[3] A. Bessi and E. Ferrara, "Social bots distort the 2016 US Presidential election online discussion," 2016.

[4] F. Schäfer, S. Evert, and P. Heinrich, "Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzō Abe's Hidden Nationalist Agenda," *Big data,* vol. 5, no. 4, pp. 294-

309, 2017.

[5] C. A. D. L. Salge and E. Karahanna, "Protesting Corruption on Twitter: Is It a Bot or Is It a Person?," *Academy of Management Discoveries,* vol. 4, no. 1, pp. 32-49, 2018.

[6] D. Stukal, S. Sanovich, R. Bonneau, and J. A. Tucker, "Detecting bots on Russian political Twitter," *Big data,* vol. 5, no. 4, pp. 310-324, 2017.

[7] E. Gallagher. "Visualizations of the Finnish-themed Twitter botnet." https://medium.com/@erin_gallagher/visualizations-of-the-finnish-themed-twitter-botnet-bfc70c6f4576.

[8] A. Patel. "Someone Is Building A Finnish-Themed Twitter Botnet." https://labsblog.f-secure.com/2018/01/11/someone-is-building-a-finnish-themed-twitter-botnet/.

[9] F. Morstatter, Y. Shao, A. Galstyan, and S. Karunasekera, "From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election," in *Companion of the Web Conference 2018 on The Web Conference 2018*, 2018: International World Wide Web Conferences Steering Committee, pp. 621-628.

[10] L.-M. N. Neudert, "Computational propaganda in Germany: A cautionary tale," *Computational Propaganda Research Project, Paper,* vol. 7, p. 2017, 2017.

[11] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, "Social bots: Human-like by means of human control?," *Big data,* vol. 5, no. 4, pp. 279-293, 2017.

[12] S. C. Woolley and P. N. Howard, "Computational propaganda worldwide: Executive summary," *Working Paper,* no. 11. Oxford, UK, p. ProjectonComputationalPropaganda, 2017.

[13] R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, "On profiling bots in social media," in *International Conference on Social Informatics*, 2016: Springer, pp. 92-109.

[14] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016: International World Wide Web Conferences Steering Committee, pp. 273-274.

[15] D. M. Beskow and K. M. Carley, "Its all in a name: detecting and labeling bots by their name," *Computational and Mathematical Organization Theory,* pp. 1-12, 2019.

[16] A. Minnich, N. Chavoshi, D. Koutra, and A. Mueen, "BotWalk: Efficient adaptive exploration of Twitter bot networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017: ACM, pp. 467-474.

[17] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.

[18] J. Fernquist, L. Kaati, and R. Schroeder, "Political Bots and the Swedish General Election," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2018: IEEE, pp. 124-129.

[19] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM,* vol. 59, no. 7, pp. 96-104, 2016.

[20] B. Wang, A. Zubiaga, M. Liakata, and R. Procter, "Making the most of tweet-inherent features for social spam detection on twitter," *arXiv preprint arXiv:1503.07405,* 2015.

[21] N. Chavoshi, H. Hamooni, and A. Mueen, "DeBot: Twitter Bot Detection via Warped Correlation," in *ICDM*, 2016, pp. 817-822.

[22] B. Kollanyi and P. N. Howard, "Junk news and bots during the German parliamentary election: What are German voters sharing over Twitter," ed: Oxford University: Comprop Data Memo, 2017.

[23] C. Bjola, "Propaganda in the digital age," ed: Taylor & Francis, 2017.

[24] P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer, "On the influence of social bots in online protests," in *International Conference on Social Informatics*, 2016: Springer, pp. 269-278.

[25] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017: International World Wide Web Conferences Steering Committee, pp. 963-972.

[26] N. Simojoki. "“Ei se ulkopuolelta suunnatulta kampanjalta vaikuta” in Finnish." https://demokraatti.fi/ei-se-ulkopuolelta-suunnatulta-kampanjalta-vaikuta-asiantuntijat-eu-vaalit-houkuttelevampi-vaikutusyritysten-kohde/.

[27] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PloS one,* vol. 8, no. 11, p. e79449, 2013.

[28] F.p.s.b.c. (Yle). "Pekka Haavisto campaign concerned over suspected Twitter bots." https://yle.fi/uutiset/osasto/news/pekka_haavisto_campaign_concerned_over_suspected_twitter_bots/9988551.

Paper II

The following paper has been published in the proceedings of the 55th Hawaii International Conference on System Sciences in 2022.

The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation

Sippo Rossi
Copenhagen Business School

Abstract

During the past few years, social media platforms have been criticized for reacting slowly to users distributing misinformation and potentially dangerous conspiracy theories. Despite policies that have been introduced to specifically curb such content, this paper demonstrates how conspiracy theorists have thrived on Twitter during the COVID-19 pandemic and managed to push vaccine and health related misinformation without getting banned. We examine a dataset of approximately 8200 tweets and 8500 Twitter users participating in discussions around the conspiracy term Scamdemic. Furthermore, a subset of active and influential accounts was identified and inspected more closely and followed for a two-month period. The findings suggest that while bots are a lesser evil than expected, a failure to moderate the non-bot accounts that spread harmful content is the primary problem, as only 12.7% of these malicious accounts were suspended even after having frequently violated Twitter's policies using easily identifiable conspiracy terminology.

1. Introduction

We may be living in a golden age of conspiracy theories [1, 2]. In everyday life, you can hear terms such as QAnon and Pizzagate, which previously belonged to the vocabulary of fringe groups but have increasingly been adopted by wider audiences and normalized as part of the vernacular by the news media and elected officials [3]. This "normalization" of conspiracy theories started well before the COVID-19 pandemic as several politicians, particularly in the United States, began promoting them during the 2016 and 2020 presidential elections. However, the coronavirus resulted in an eruption of new theories, ranging from 5G causing the virus [4] to the Great Reset that suggests the pandemic is being used as an excuse to take control of the world economy [5].

The role of the social media platforms in spreading conspiracy theories has been tied to their loose regulation as well as to conspiracy theorists taking advantage of their algorithms to amplify the spread of content, bringing it from the obscure corners of the internet to the mainstream feeds of the general public. Some evidence supports the idea that bad actors have also become better at exploiting vulnerabilities of the algorithms that power modern content recommendation systems [6].

Furthermore, recent studies have suggested that the conspiracy theories and misinformation related to the COVID-19 pandemic have been amplified by suspected bot accounts [7, 8]. For example, bots have been used to distribute conspiracy theories related to the pandemic alongside references to QAnon and the Great Awakening as well as to share links to other low credibility content and fake news sites [7]. Although the existence of bots in this context has been proven, the estimates of their prevalence range wildly.

To reduce the reach of conspiracy theories, social media sites such as Twitter, Facebook and YouTube have begun moderating content more aggressively, suspending accounts that spread misinformation and shutting down groups that are devoted to or helping spread conspiracy theories [9]. For instance, many social networking sites have attempted to remove links and references to the Plandemic, which was a viral video that promoted several conspiracy theories related to COVID-19 and has become a term used to refer to the pandemic as an orchestrated epidemic or hoax [10]. While some social media sites like Facebook have seen a decline in the number of interactions with content containing misinformation during the recent years, in Twitter interactions with such content have been steadily growing [11].

Moderating content on conspiracy theories is challenging. The propagators' methods have evolved quickly, adapting to restrictions, developing new terms and adding nuance to content that is harder to identify as misinformation. When searching on Twitter with the term Plandemic, the social media site redirects the search to the word "pandemic" and provides a link to official information about the COVID-19 pandemic. But, when using the search term Scamdemic, Twitter does not attempt this redirection and users can even see tweets where the word or hashtag Plandemic is used in unison with the word Scamdemic. The difference between the words is minor, but the new term is unmistakably related and has been widely used without consequences. This is surprising considering Twitter's policy update which banned sharing conspiracy theories about the pandemic or COVID-19 vaccines [12].

Due to the possible negative effects on society and public health that conspiracy theories around the COVID-19 pandemic may have, it is important to evaluate the effectiveness of these content bans and to understand what is driving the conspiracy theories so that they can be addressed accordingly. The possible involvement of bots in the COVID-19 conspiracy theory discussions further complicates the study of the topic as well as the analysis of policies, as moderation of computer-generated content is far less ambiguous than the moderation of content made by humans. This is due to the use of bots for manipulation being clearly banned and the removal of such accounts will not result in dissatisfied users, whereas moderating genuine human accounts can lead to accusations of stifling free speech as well as users fleeing to competing social networking sites. Analysis of the level of bot involvement is thus needed as it will increase our understanding of whether the primary issue is genuine or inauthentic accounts. Furthermore, the policy recommendation will vary depending on what types of accounts are the primary source of misinformation.

At the time of writing only few publications addressed the effectiveness of Twitter's misleading information policy that was adjusted multiple times during 2020 and 2021 to reduce COVID-19-

related misinformation. The policy changes adjusted the criteria for suspensions and content removal and introduced a strike system for accounts, where repeatedly tweeting content containing misinformation would eventually lead to permanent suspension [12]. To address this research gap, and to provide a basis for future research, an exploratory study was conducted on a sample of COVID-19 conspiracy theory tweets that are using the conspiracy term "Scamdemic," which is commonly used both as a hashtag or as a keyword in the tweet to signal that the pandemic is a conspiracy or hoax.

The goal of this paper is two-fold. Firstly, it will investigate who are using the Scamdemic term on Twitter in order to determine whether the conspiracy theories are being pushed by coordinated attacks by bots and trolls or organically by users that believe in the conspiracy theories. Based on the findings on what type of accounts and content are evading the bans, the effectiveness and level of enforcement of current policies could be evaluated and policy changes suggested. Secondly, it will evaluate how well Twitter's COVID-19 misleading information policy is enforced using the tweets containing the word "Scamdemic" as a case example.

2. Related research

The literature review is divided into three parts. The first part will summarize research related to the spread of misinformation in social media and explain what is currently known of the characteristics of COVID-19 conspiracy content on Twitter. The second part will discuss what is known of bots and their ability to influence discussions on Twitter and provides a motivation for the method that will be used for bot detection in this study. Lastly, articles that are methodologically similar to this paper and use network analysis to study misinformation on Twitter will be reviewed. Overall, the goal is to establish and describe what has previously been observed in misinformation research and to justify the design choices outlined in the methodology section.

2.1 Misinformation and conspiracy theories on social media

The spread of conspiracy theories and misinformation has been widely studied [13] and the role of social media in distributing such content is a well-known issue [7]. New conspiracy theories and adaptations of earlier ones have appeared during the COVID-19 pandemic [14] and quickly reached wide audiences through social media platforms [4]. The conspiracy theory on the "Plandemic" which trended in multiple social media sites as a result of a viral video, is an example of COVID-19 influencing and modifying existing conspiracy theories related to vaccines [10].

One distinguishable characteristic of tweets involving conspiracy theories is that certain groups of hashtags and keywords are prevalent in them. For example, 5G is commonly mentioned due to the popularity of the related conspiracy theory that suggests the technology's emergence is linked to the disease [4]. Plandemic tweets also often include hashtags or mentions of other indirectly related or unrelated conspiracy theories and common examples include QAnon and the Deep State [7, 10]. Furthermore, many tweets containing misinformation include links to both YouTube videos as well as fake news websites [4].

Social engagement metrics such as likes and retweets have been shown to increase the susceptibility of users to posts containing misinformation [15]. This suggests that the virality of conspiratorial content and other misinformation is a threat precisely because people are unlikely to critically evaluate the source. Based on this, posts made by influential accounts that are retweeted and or liked in large numbers are a bigger threat than those made by less popular accounts, thus suggesting the focus of the analysis should be on them.

Influential accounts are not the only issue behind the propagation of misinformation, as the design of the platforms and the way they promote content is a major part of the problem as well. One of the theories that is used to both describe behavior in online social networks as well as to support or oppose restrictive policies is the concept of “echo chambers” or alternatively “filter bubbles” [16, 17]. These echo chambers are formed as a result of recommendation systems that aim to maximize interaction by providing users of the social networking site with content such as tweets that matches their views and suggestions on which accounts to follow that share similar content [16]. The danger according to this theory is that an individual will be eventually exposed mainly to material that aligns with their world view giving a false sense of unanimity while only a small minority of individuals supports the belief. Due to not seeing material that challenges for example the conspiracy theories, the individuals become more entrenched in their bubble or echo chamber [17]. However, too aggressive moderation of content or the banning of entire communities is a risk as it may result in the users abandoning the platform and moving to an alternative social networking site that may further increase the divide and drive individuals to the fringe.

2.2 The role of Twitter bots in the distribution of misinformation

When using the term bot or bot account, this paper refers to basic spambots as well as social bots that are either fully or partially automated and engaging in distributing controversial content without self-identifying as non-human. This is based on the definition given by Ferrara et al. [18].

Many papers have discussed the role of bots in distributing fake news and misinformation [6, 19]. It is argued that at least their role in distributing content from low-credibility sources is disproportionately big [19]. The percentage of bots in the entire Twitter population has been estimated to be around 10% - 20%. However, in the case of accounts pushing the United States to reopen the country and to reduce COVID-19 restrictions it has been up to 50% [8]. Furthermore, there is evidence of social bots that are interacting predominantly with COVID-19 content, suggesting that their purpose is to spread or amplify misinformation related to the pandemic [7].

One of the negative effects of bots, which further demonstrates their potential in the context of disseminating misinformation, is their assumed role in strengthening the spiral of silence [20]. The spiral of silence theory suggests that individuals monitor and attempt to understand the general opinion on a given topic and if they perceive themselves to be supporting the stance of the minority, they are likely to refrain from expressing their opinion [21, 16]. This ultimately affects other people’s perception of the topic and can lead to a setting where a silent majority accepts that the opposing view is the prevailing opinion of the population, while in fact it is supported by a vocal minority [21]. One recent study suggests that even a relatively small percentage of bots can affect online discussion and tip the perceived public opinion [20].

A major issue in studies that investigate the role of bot accounts in the spread of misinformation is the difficulty of reliably detecting modern social bots. Recent papers focusing on Twitter bot detection rely increasingly on machine learning [22, 23, 24] and ensemble methods combining multiple classifiers due to the level of sophistication of bots as well as hybridization where both humans and programs control the accounts [25]. One particularly widely used example has been the Botometer (or originally BotOrNot), which has been featured in many of the most cited publications on social bots [26, 27]. However, fully automated bot detection may not be realistic [25, 26] because the results of such techniques have been shown to vary with new datasets. Therefore, relying on existing tools such as the Botometer alone and drawing conclusions without critical qualitative inspection appears no longer sufficient and thus in this paper a hybrid approach combining algorithmic and qualitative labeling is employed.

2.3 Networks on Twitter

Social network analysis has been widely used to study social media [28], how information spreads in networks [29] and which accounts are most influential in facilitating information spread [30]. Network models based on Twitter data can be built in multiple ways with the simplest examples being models where connections represent accounts following each other or mentioning each other in tweets. The networks can also represent relationships between content that is being shared such as tweets containing the same hashtag, and in that case the nodes can be the hashtags.

Research has shown that in the case of misinformation, unverified accounts that do not belong to any well-known public figures influence the spread of conspiracy theories [31, 32]. However, the way in which these influential accounts are defined varies a lot and influence can be measured in many ways. For example, when defining influence algorithmically, the most influential accounts can be those that are surrounded by highly retweeted accounts who commonly share the content of the less well-known account [31]. Simpler approaches rely on using different metrics related to the Twitter accounts such as the number of followers [30] and betweenness centrality [4].

3. Methodology

3.1 Data collection

The dataset consists of Twitter usernames, tweets and a mapping of the relationships between the different objects, which will be described in more detail under the network analysis section. The data was collected using Twarc, a Python library for accessing and retrieving data from the Twitter API.

The data contains 8263 tweets and 8540 users interacting with or being related to these tweets. The data was gathered from the Twitter API with the tweets/search command and search term "scamdemic." Users are considered related to a tweet if the tweet mentions the user, retweets or quotes the user or if it is a reply to a tweet made by the user. The dataset contains tweets posted during a one-week period starting on the 8th of March and ending on the 15th of March 2021. The time was chosen based on Twitter having updated their COVID-19 misinformation policy at the beginning of the month. The script used to collect and process the data is based on a tool described in [33].

3.2 Network analysis

The data was mapped so that two separate network graphs can be created. The first one is labeled the account-interaction network, which is a weighted directed network where nodes are accounts while the edges represent interactions towards other accounts with tweets. Weights are determined by how many times during the analysis period an account interacted with the other account by for example retweeting or mentioning them. The second is labeled the account-hashtag network and is a directed multimodal network where nodes are both hashtags as well as accounts and the edges indicate which hashtags an account interacted with. From now on, the first network will be referred to as the account-interaction network and the latter as the account-hashtag network. The networked data was analyzed both quantitatively, with standard network analysis metrics, as well as qualitatively by manually inspecting the most important nodes' Twitter profiles. Overall, the purpose of this network analysis was to determine how the average account using the Scamdemic word behaved, which hashtags were used together, and by whom, and to identify which accounts were most prominent in the network.

To make inferences on the effectiveness of Twitter's policies, a population of influential accounts were selected based on the three node characteristics: betweenness centrality, indegree and the outdegree. A high indegree indicates that the account is often referred to in other tweets, while a high outdegree would indicate a spammer whose content are likely to be seen by individuals searching with the right keywords. Lastly, users with a high betweenness centrality are the accounts that act as a bridge between communities and discussions. Figure 1 illustrates these different node characteristics with the teal "A" node representing a node which has a high indegree, while the red node "B" has a high outdegree and the yellow node "C" a high betweenness centrality as it acts as a link between the two communities around the node "A" and node "B."

Heuristically, the 25 accounts with the highest betweenness centrality, in and outdegrees were determined to be influential, and consequently their activities reviewed twice during the two months following the collection of the dataset. As some influential nodes were in the top 25 of several characteristics, the final list of influential nodes consists of only 61 accounts. The rationale behind focusing on these accounts is that they should be among the first to be deleted due to their prominence assuming that they are in fact supporting the conspiracies and not attempting to debunk them.

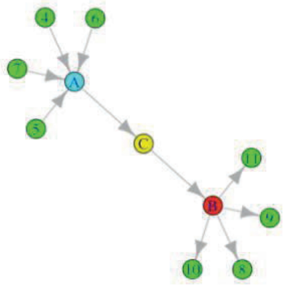


Figure 1. Node characteristics example

3.3 Bot detection and classifying accounts

Several methods were used in conjunction to determine what types of accounts were participating in the discourse and if bots are amplifying the Scamdemic conspiracy. Firstly, the 61 most influential nodes were checked with the Botometer which provides a rating on the likelihood of the account being a bot rather than a classification. Secondly, manual inspection and coding was done to further validate the scores provided by the Botometer. This two-step classification of accounts should reduce the risk of misclassification tied to the Botometer's scores [26]. All influential accounts were checked even if the Botometer suggested that they are not suspicious.

In addition to labeling accounts as humans or bots, the manual inspection was used to bin the accounts into overlapping categories. The labels for these categories are conspiracy theorists, spammer, antivax, celebrity and non-believer. Conspiracy theorists are accounts that seemed to authentically believe and participate in the discussions. Spammers are accounts that solely push content through liking or retweeting. Antivaxxers are a subset of conspiracy theorists, mainly engaging with content questioning the safety of COVID-19 vaccines. Lastly, celebrity indicated prominent politicians and non-believers accounts that are participating in the discussions in order to debunk conspiracies. This manual inspection was conducted twice. First, a month after the collection of the dataset and a second time a month later to see on both occasions which accounts had been banned during the monitoring period and to follow-up on whether the coding was still accurate.

Figure 2 shows the account-interaction network with influential human accounts being marked as black, and influential nodes suspected of being bots as red. Due to the metrics used to determine influence, the influential nodes are mostly in the center of a cluster of accounts or acting as a link between several clusters.

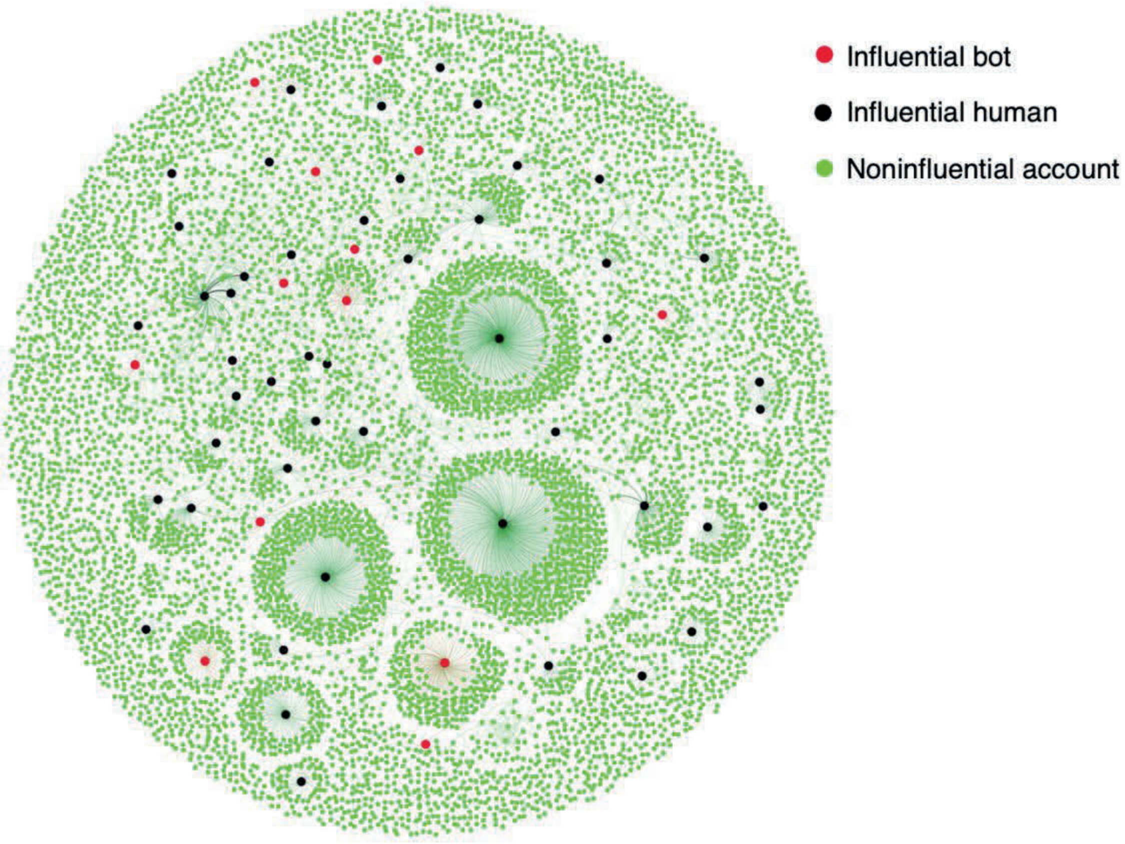


Figure 2. The account-interaction network with influential nodes highlighted

3.4 Limitations

Accessing Twitter's API with Twarc does not guarantee that all tweets related to the search term are collected. This is due to the tool not supporting Twitter's academic product track's full-archive search. However, even small samples instead of full datasets have been successfully used in previous studies [4] and especially considering the relative niche status of the Scamdemic, a low volume of tweets can be expected. Furthermore, the dataset is small when compared to typical Twitter studies, but for the purposes of demonstrating how individual influential accounts can avoid bans while repeatedly posting content that is against the rules, it should be sufficient.

4. Findings

4.1 Accounts

The account-interaction network which consisted of all the 8540 accounts in the discussions is very sparse. Most nodes are peripheral or separate from the main network with 83% of the nodes having an outdegree between 0 and 1, while 90% of the nodes have an indegree between 0 and 1, meaning that they have interacted with another account or been mentioned in a tweet 0-1 times. In all three previously described network characteristics that were used to define the influential accounts; the indegree, outdegree and betweenness centrality, the top 10 to 25 accounts are distinguishable from the rest by having values that are several hundred or even thousand percent higher than the mean.

Figure 3 represents the account-interaction network that was created using Gephi with the Force Atlas 2 layout algorithm. The different colors represent communities that were identified based on the modularity. There are several influential nodes that have large communities of accounts interacting with them while most are in hardly visible small clusters of 2-3 accounts.

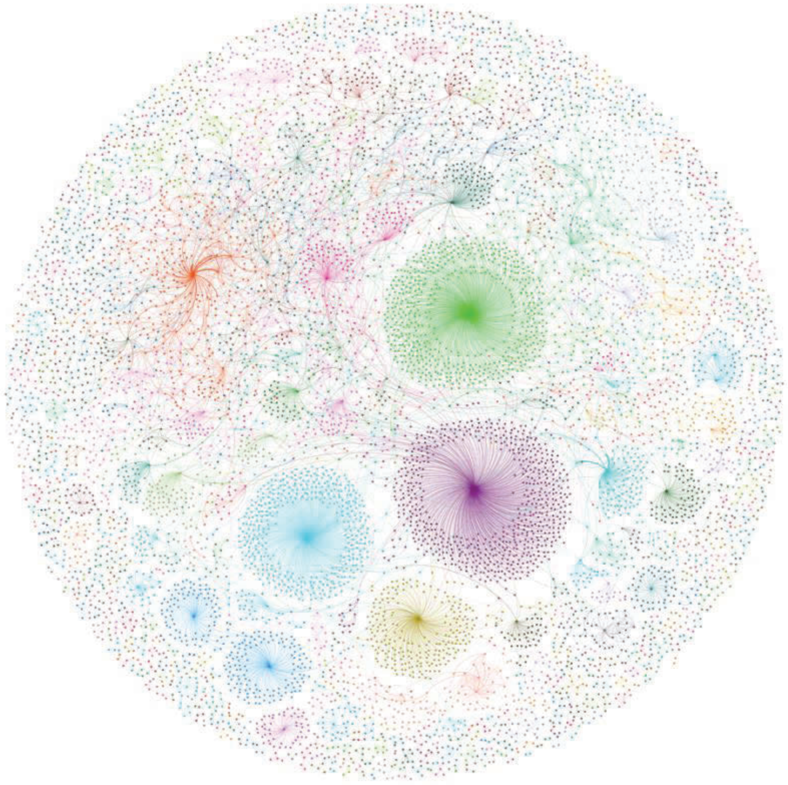


Figure 3. The account-interaction network clustered into communities

From the list of 61 influential accounts, only five were identified as public organizations, well-known individuals or politicians that commented on the conspiracies or who were mentioned in the discussions. Furthermore, only one of these five accounts was a supporter of COVID-19 conspiracy theories. Additionally, in the case of one of the participants, it was unclear whether they were debunking or promoting the conspiracies.

This is in line with previous research which suggests that with misinformation, most of the influential accounts are not verified users or public figures. The remaining 55 accounts were a mix of trolls, bots and users assumed to be authentic conspiracy theorists or believers and thus can be referred to as malicious accounts. Only six accounts had been banned after a month and three had been renamed and thus were no longer characterizable. On the second inspection two months after the data was collected, one additional account had been suspended and two more renamed making them untraceable.

Defining the exact type of the malicious accounts proved to be difficult as their goals were not clear in most cases. By looking at the profile descriptions and content that they tweeted and retweeted, in some cases the motives as well as their assumed country of origin were identifiable. Surprisingly, over 40 percent of the influential accounts were interacting with COVID-19 content originating or related to Great Britain and British politics, which suggests that the word Scamdemic is popular in the British Twitter conspiracy theory circles, or that the sample was taken during a time in which the word was trending in the United Kingdom.

According to the Botometer, from the list of most influential accounts only five were given a score above 4 on a scale of 1-5 by the Botometer. The universal, or language independent score was used as not all accounts were tweeting in English. Two of these assumed bot accounts were banned at the time of writing and one was manually reclassified as a human on closer inspection. During the manual labeling, thirteen accounts were labeled as likely to be bots based on their behavior, which usually included mass retweeting and spamming of hashtags and mentions. The thirteen suspected bot accounts included all except one of the five accounts labeled by the Botometer as highly likely of being a bot. This would indicate that the qualitative labeling was more aggressive than the Botometer, which tends to be conservative with its estimates.

Overall, these thirteen suspected bot accounts represent a quarter of the influential accounts which is on the high end of the assumed share of bot accounts on Twitter, but on the low end of the estimates of the analyses that looked into the share of bots in COVID-19 misinformation tweets [6, 8].

4.2 The content

The content analysis focused on the different hashtags that were used since they play an important role in making a particular topic recognizable and easy to find on social networking sites such as Twitter. The capitalization was removed in order to combine some of the otherwise identical hashtags, such as Scamdemic and scamdemic which were initially treated as unique hashtags. A total of 3127 hashtags were used in the 8263 tweets and the ten most used ones represent 34.9% of all hashtags.

Figure 4 shows the account-hashtag network, where nodes are a mix of hashtags as well as the accounts that used them in their tweets. The communities are less distinct as in the previously discussed account-interaction network, which is due to the common practice of including multiple hashtags in posts. Several somewhat separate communities can be seen at the edges of the network, such as the pink cluster of non-English hashtags on the left side of the graph and calls to participate in rallies in the gray cluster at the top.

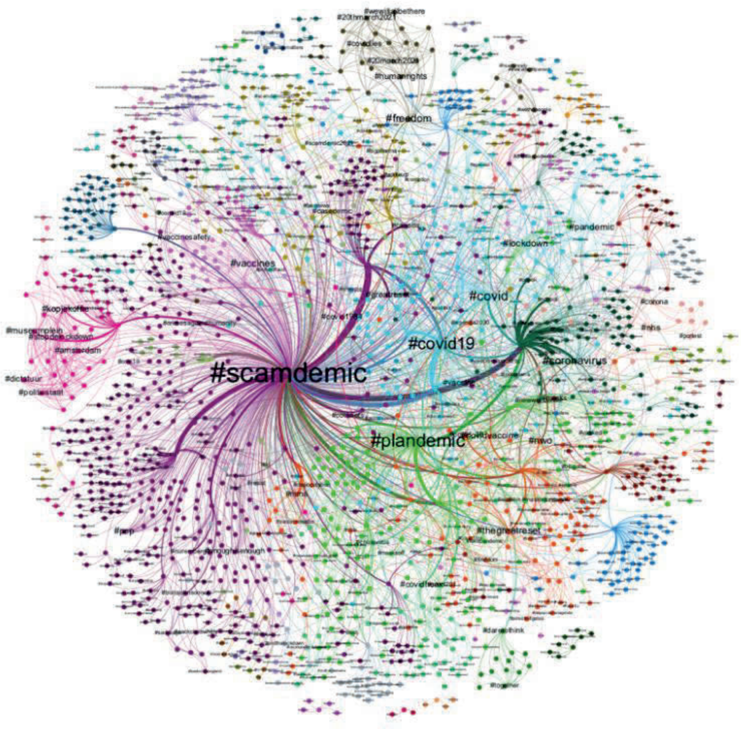


Figure 4. The account-hashtag network

Unsurprisingly, the most frequently used hashtag was #scamdemic which was used over 500 times in the dataset, followed by the over 200 mentions of #covid19 in various ways of writing, which were merged with fuzzy matching. At third place was #plandemic which had been used over 100 times despite being a particularly scrutinized word. Other much used hashtags included the popular Great Reset (#thegreatreset) and New World Order (#nwo) conspiracy theories, as well as a large variety of different references to the COVID-19 vaccine. The table below shows the top ten hashtags, which includes generic pandemic related words such as lockdown and vaccines in addition to the terms linked to conspiracy theories. Most hashtags are in English, although German and other minor European languages were used in small numbers as well. Table 1 shows the ten most popular hashtags and how many times they were used in the dataset.

Table 1. Most popular hashtags

Hashtag	Count
#scamdemic	524
#covid19	220
#plandemic	114
#thegreatreset	41
#nwo	35
#coronavirus	31
#freedom	28
#vaccines	24
#lockdown	21
#covidvaccine	19

5. Discussion

5.1 Implications

One of the main objectives of the study was to determine the nature of the accounts that were participating in the distribution of the Scamdemic conspiracy term by looking closely at a sample of 61 highly active and influential accounts. Furthermore, by following these influential accounts for a duration of two months the research aims to highlight the lack of moderation and enforcement of Twitter's policies against misinformation. The design of the study makes it difficult to draw conclusion on the implications that the findings have on existing theories used in misinformation research. During the qualitative inspection of the influential accounts, the lack of critical comments against the COVID-19 conspiracy theories can however suggest that the active participants are within an echo chamber and or that the spiral of silence is making it difficult for the participants to voice critical comments, but this will be verified with more thorough analysis during future studies. Thus, the discussion in this paper will be centered on the empirical evidence and based on the key implications, a critical commentary on the current status is provided. Lastly, recommendations on how to adjust Twitter's misleading information policy are given.

Interestingly, a majority of the influential accounts that are using the Scamdemic word and participating in the spread of other related conspiracy theories, seem to be legitimate users rather than bots. Moreover, most of the suspected bot accounts were merely retweeting conspiracy theories constantly without producing any original tweets, indicating that they are operating with crude scripts rather than more sophisticated programs found in modern social bots. Of the 55 accounts defined as malicious and influential the rate at which they were banned is surprisingly low at 12.7%, with only 6 bans during the first month and one additional ban after two months. Previous research has focused predominantly on how modern misinformation is spread by advanced social bots and coordination but based on the sample used in this study, both the bots and humans could merely continuously retweet and post malicious content without consequences. Therefore, the level of sophistication of the accounts avoiding suspension is likely lower than previously assumed.

In order to analyze Twitter's moderation and how well they follow their new policy, we looked at the content produced and shared by the influential accounts. Content wise it seems that using indirect or novel words and hashtags to avoid suspension is not needed on Twitter. This is based on the observation that using words and hashtags known to be associated with misinformation, or directly implying for example that the pandemic is a hoax (e.g., #plandemic and #scamdemic) are not being removed.

Only one instance of a tweet being flagged as against Twitter's rules was detected during the review of the influential accounts. Based on this, the content is not getting actively flagged and censored even in obvious cases. Considering that flagging a tweet rather than deleting it is a much lighter approach and is already employed by Twitter, it is questionable why it is not used more actively.

From the findings on the accounts and content that they engage with, Twitter's ability or interest to enforce its COVID-19 misinformation policy seems very weak. Almost 90% of the inspected accounts were openly tweeting or retweeting using easily identifiable words and hashtags related to popular conspiracy theories without getting suspended. Approximately a quarter of the influential accounts were also spreading anti-vaccine content, which is another topic when discussed in the context of COVID-19 that violates Twitter's misinformation policy. It is especially surprising how these accounts that are posting multiple types of content that should automatically raise alarms do not get removed or filtered from public searches. Interestingly even names and bios containing the word covid and mentions of Scamdemic or other conspiracy words had managed to not be suspended.

5.2 Policy recommendations

Lastly, two suggestions on how to mitigate the further spread of misinformation are provided. Due to the simplicity of the accounts involved, relatively basic changes to policy would reduce the visibility of the misinformation and conspiracy theories.

Firstly, more aggressively suspending accounts according to the current misinformation policy based on repeated use of known conspiracy theory terms is suggested. Particularly accounts involved in the distribution of conspiracies such the Plandemic as well as other vaccine related misinformation have a clear lexicon and should be targeted similarly as the accounts spreading the Plandemic are on Facebook, where suspensions are given more frequently. Considering that Twitter already attempts to filter content by requiring an additional click to access the tweets when querying with the search term Plandemic, it is clear that they are already capable of identifying the misinformation but abstaining from removing it. In other words, this recommendation simply suggests that Twitter should enforce its own current policies.

Secondly, considering that the most incriminating hashtags and vocabulary such as the Plandemic and Scamdemic are used by the malicious accounts to make content easy to find, filtering the tweets containing them from the search results would reduce their visibility even without the need of removing the content or associated accounts. This would also avoid false positives leading to bans of accounts that are not promoting conspiracies but in fact attempting to debunk them.

6. Conclusion

The different misinformation, fake news as well as conspiracies surrounding the COVID-19 pandemic have been studied from many angles despite of the recentness of the topic. The goal of this paper was to contribute to the understanding of what types of accounts are distributing the conspiracy theories and misinformation related to the pandemic, as well as demonstrate that Twitter is not highly successful at mitigating the spread of misinformation. The study found limited evidence of bot accounts dedicated to spreading misinformation related to the COVID-19 pandemic as the share of assumed bots when compared to human operated accounts was lower than expected. However, the findings were in line with previous research that cites humans as the most likely cause of misinformation spreading. Lastly, the study suggests that stricter enforcement is needed, and that the situation could be improved by merely removing or filtering content that contains certain keywords or hashtags such as #scamdemic and #plandemic.

This paper highlighted how it is possible for influential accounts to repeatedly share content that is against Twitter's policies during a short time without having the content removed or the associated accounts suspended. Future studies would benefit of having a longer monitoring period than the two months used for this study, as this could provide insights on whether in the long-term enforcement of the policies is more successful. Furthermore, by expanding the list of keywords and conducting the longitudinal study on a larger group of accounts, more inferences could be made on which type of behavior and terminology in tweets manages to evade suspension.

7. References

[1] D. Freeman and J. Freeman, "Are we entering a golden age of the conspiracy theory?," *The Guardian*, Mar. 28, 2017. [Online]. Available: https://www.theguardian.com/science/blog/2017/mar/28/are-we-entering-a-golden-age-of-the-conspiracy-theory

[2] Z. Stanton, "You're Living in the Golden Age of Conspiracy Theories," *Politico*, 2020.

[3] A. Willingham, "How the pandemic and politics gave us a golden age of conspiracy theories," *CNN*, Oct. 03, 2020. [Online]. Available: https://edition.cnn.com/2020/10/03/us/conspiracy-theories-why-origins-pandemic-politics-trnd/index.html.

[4] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, "COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data," *Journal of Medical Internet Research*, vol. 22, no. 5, pp. 1–9, 2020.

[5] J. Goodman and F. Carmichael, "The coronavirus pandemic 'Great Reset' theory and a false vaccine claim debunked," *BBC*, Nov. 22, 2020.

[6] D. M. J. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[7] E. Ferrara, "What Types of Covid-19 Conspiracies Are Populated By Twitter Bots?," *First Monday*, vol. 25, no. 6, 2020.

[8] K. Hao, "Nearly half of Twitter accounts pushing to reopen America may be bots," *MIT Technology Review*, May 2020. [Online]. Available: https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/.

[9] A. Hern, "Tech giants join with governments to fight Covid misinformation," *The Guardian*, Oct. 20, 2020.

[10] M. D. Kearney, S. C. Chiang, and P. M. Massey, "The Twitter origins and evolution of the COVID-19 'plandemic' conspiracy theory," *Harvard Kennedy School Misinformation Review*, vol. 1, no. October, pp. 1–18, 2020.

[11] Allcott, H., M. Gentzkow, and C. Yu, "Trends in the diffusion of misinformation on social media," *Research and Politics 6*(2), 2019.

[12] Twitter, "COVID-19 misleading information policy," *Twitter Help Center*, 2021. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

[13] T. R. Tangherlini, S. Shahsavari, B. Shahbazi, E. Ebrahimzadeh, and V. Roychowdhury, *An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web*, vol. 15, no. 6. 2020.

[14] S. Shahsavari, P. Holur, T. R. Tangherlini, and V. Roychowdhury, "Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news," *arXiv*, pp. 1–21, 2020.

[15] M. Avram, N. Micallef, S. Patil, and F. Menczer, "Exposure to social engagement metrics increases vulnerability to misinformation," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 5, pp. 1–11, 2020.

[16] M. Nelimarkka, S.-M. Laaksonen, and B. Semaan, "Social Media Is Polarized," *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 2018, pp. 957–970.

[17] M. D. Vicario, A. Bessi, F. Zollo, et al., “The spreading of misinformation online,” *Proceedings of the National Academy of Sciences of the United States of America 113*(3), 2016, pp. 554–559.

[18] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[19] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, no. 1, 2018.

[20] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, “Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks,” *European Journal of Information Systems 28*(4), 2019, pp. 394–412.

[21] E. Noelle-Neumann, “The Theory of Public Opinion: The Concept of the Spiral of Silence,” *Annals of the International Communication Association 14*(1), 1991, pp. 256–287.

[22] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, “Malicious accounts: Dark of the social networks,” *Journal of Network and Computer Applications*, vol. 79, no. November 2016.

[23] D. M. Beskow and K. M. Carley, “Bot conversations are different: Leveraging network metrics for bot detection in Twitter,” *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 825–832, 2018.

[24] S. Kudugunta and E. Ferrara, “Deep neural networks for bot detection,” *Information Sciences*, vol. 467, pp. 312–322, 2018.

[25] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, “Social Bots: Human-Like by Means of Human Control?,” *Big Data*, vol. 5, no. 4, pp. 279–293, 2017,

[26] A. Rauchfleisch and J. Kaiser, “The False positive problem of automatic bot detection in social science research,” *PLoS ONE*, vol. 15, no. 10, 2020.

[27] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “Detection of Novel Social Bots by Ensembles of Specialized Classifiers,” Jun. 2020.

[28] J. Cao, K. A. Basoglu, H. Sheng, and P. B. Lowry, “A systematic review of social networks research in information systems: Building a foundation for exciting future research,” *Communications of the Association for Information Systems*, vol. 36, pp. 727–758, 2015.

[29] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman, and C. Espina, “Classifying Twitter Topic-Networks Using Social Network Analysis,” *Social Media and Society*, vol. 3, no. 1, 2017.

[30] I. Anger and C. Kittl, “Measuring influence on Twitter,” in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 2011, pp. 4–7,

[31] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[32] B. Huang and K. M. Carley, “Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak,” *arXiv*, pp. 1–19, 2020.

[33] A. Patel, “Searching Twitter With Twarc,” *F-Secure blog*, 2018. https://blog.f-secure.com/searching-twitter-with-twarc/.

Paper III

The following paper has been published in the proceedings of the 56th Hawaii International Conference on System Sciences in 2023.

Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?

Sippo Rossi
Copenhagen Business School
sr.digi@cbs.dk

Youngjin Kwon
Temple University
youngjin.kwon@temple.edu

Odd Harald Auglend
Copenhagen Business School
oha.digi@cbs.dk

Raghava Rao Mukkamala
Copenhagen Business School
rrm.digi@cbs.dk

Matti Rossi
Aalto University
matti.rossi@aalto.fi

Jason Thatcher
Temple University
jason.thatcher@temple.edu

Abstract

In recent years, deep learning methods have become increasingly capable of generating near photorealistic pictures and humanlike text up to the point that humans can no longer recognize what is real and what is AI-generated. Concerningly, there is evidence that some of these methods have already been adopted to produce fake social media profiles and content. We hypothesize that these advances have made detecting generated fake social media content in the feed extremely difficult, if not impossible, for the average user of social media.

This paper presents the results of an experiment where 375 participants attempted to label real and generated profiles and posts in a simulated social media feed. The results support our hypothesis and suggest that even fully-generated fake profiles with posts written by an advanced text generator are difficult for humans to identify.

Keywords: Social Bots, Social Media, Experiment, Deep Learning, GAN Images

1. Introduction

Disinformation, conspiracy theories, links to phishing sites and even legitimate sales pitches are being sent with the help of fake social media profiles to unsuspecting users of social networking sites (Bond, 2022; Shafahi et al., 2016; Shao et al., 2018). These fake profiles might be automated bots belonging to a network of such or be operated by a human that manages multiple accounts. Previously maintaining realistic avatars and producing convincing content at scale was difficult as stolen images could be found with a reverse image search. Similarly, even the best generated text was unable to consistently pass for something written by a human.

As a result of recent rapid advances in deep learning (DL) methods that can be used to generate realistic images and due to the proliferation of advanced pre-trained natural language processing (NLP) models, creating synthetic profile pictures and producing human-like texts is easier than ever before (Brown et al., 2020; Köbis & Mossink, 2021; Nightingale & Farid, 2022). Consequently, producing large numbers of fake profiles with individual or even multiple synthetic components such as photorealistic profile pictures and generated but human-like posts is now technologically and economically viable. Thus, these advances may have made detecting sophisticated fake profiles near impossible for the average user of a social networking site. From the perpetrator's point of view, this has made information operations and trolling cheaper and much less labor-intensive. If these fake profiles are not distinguishable by humans and large quantities of them can be produced easily, we could see an explosion of bot accounts being used for marketing, phishing or political campaigns among other purposes.

It is not yet known whether fully synthetic profiles and social media posts are in fact able to pass the Turing test and go unnoticed by humans. Therefore, the primary goal of this paper is to study how well humans can detect deep learning-generated social media profiles and posts in a social networking site's feed, as well as to assess the ability of modern deep learning to produce humanlike content. Our first research question is:

RQ1: Can humans distinguish social media profiles with DL-generated profile pictures and DL-generated posts from real ones in the feed of a social networking site?

While it is known that in isolation these individual generated components are no longer distinguishable from real ones, there is a possibility that in some cases when combined within one profile they become suspicious. As an example, consider the situation when the text in a post contains vocabulary used by a certain demographic group, but the profile picture belongs to clearly different group. Therefore, our second research question is:

RQ2: Which components of a profile are more likely to make humans suspect that the profile is fake?

We hypothesize that we have crossed the boundary where generated social media profiles can no longer be consistently detected by humans. To test this hypothesis and to answer the research questions, we conducted an experiment where participants were shown both genuine and fully generated bot profiles in a simulated social media feed and asked to classify the accounts and assess different components of the profiles on whether they are suspicious or not. The feed contained both the basic profile information as well as one post made by each account.

This paper begins by briefly synthesizing the findings of recent literature related to image and text generation and fake content on social media. Next, the experiment and methodology used to produce the social media profiles are explained. We then describe the results of the experiment and discuss the main findings and implications. Lastly, we conclude by considering the limitations of this study and provide an overview of the planned future work that will address them.

2. Background and related research

In this section, we first briefly discuss recent trends in fake account detection. Then, we summarize the state of the art in image and text generation using deep learning methods and lastly review recent experiments involving fake social media content.

2.1. Bots and fake content on social media

During the recent years bots on social media as well as methods to detect them have been studied in both information systems (Ross et al., 2019; Salge et al., 2022; Stieglitz et al., 2017; Williamson & Scrofani, 2019) and computer science research (Cresci, 2020; Ferrara et al., 2016). There are also a number of studies on the prevalence and impact of bots under various topics such as elections (Brachten et al., 2017) or the COVID-19 pandemic (Marx et al., 2020; Rossi, 2022). Fake content such as misinformation and fake news (Kim & Dennis, 2019; Moravec et al., 2019), as well as the role of bots in distributing them have also been investigated (Lazer et al., 2018; Shao et al., 2018).

However, academic research on the human ability to detect deep learning-generated content is scarce, based on our search. We assume this is most likely due to the technology only recently having matured enough to produce believable fake content. Meanwhile, algorithmic detection of fake images and bot profiles on social media have been studied more, despite the relative recentness of the topic (Cresci, 2020; X. Wang et al., 2022; Yu et al., 2019). From other non-academic sources, there are documented examples of cases where fake profiles were caught using GAN images on Twitter (Strick, 2021) and LinkedIn (Bond, 2022) to pass as real humans. While some examples were more benign, groups of accounts with GAN images impersonating real humans have been said to be employed for promoting computational propaganda (Da San Martino et al., 2021; Strick, 2021).

2.2. Text and image generation using deep learning

Deep learning methods for text generation have become more accessible due to powerful pre-trained models such as GPT-2 and GPT-3 (Brown et al., 2020; Li et al., 2021). In the past language models required significant computational resources and large dataset sets to train them for each individual topic. In contrast, pre-trained models are trained with massive amounts of training data before being released to the public. Therefore, they do not have this limitation as they can be used immediately or after being fine-tuned with much more manageable datasets and computing power (Li et al., 2021).

These powerful pre-trained models for text generation have been researched and developed primarily by leading technology companies and private research organizations in recent years. Therefore, there is still a limited number of peer-reviewed academic research on, for example on, how well humans can distinguish human-written texts from the auto-generated text. Early works have shown that GPT-3 is, for example, capable of producing poems that are indistinguishable from genuine ones (Köbis & Mossink, 2021) and that humans have difficulties even after training to detect machine-generated text (Clark et al., 2021). It has also been suggested that GPT-2 has been used to produce texts for malicious accounts (Da San Martino et al., 2021), although it is difficult to prove and the effectiveness of it is still unknown.

For image generation, Generative Adversarial Networks (GANs), a type of deep learning architecture, have been demonstrated to be able to produce synthetic images that are algorithmically detectable, but for humans seemingly photorealistic (Karras et al., 2019; Yu et al., 2019). A recent experiment with facial images generated with StyleGAN (Karras et al., 2019), an advanced and alternative architecture for GANs, demonstrated how synthetic images were deemed on average more trustworthy than real faces and nearly undetectable (Nightingale & Farid, 2022). Figure 1 contains examples of GAN-generated images of human faces.

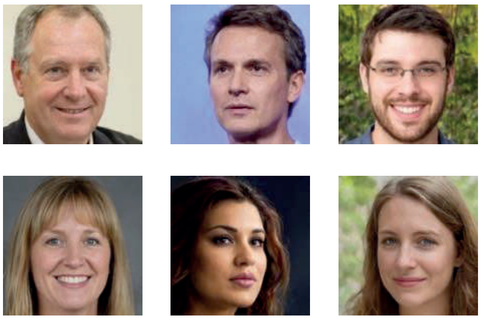


Figure 1. GAN-generated images

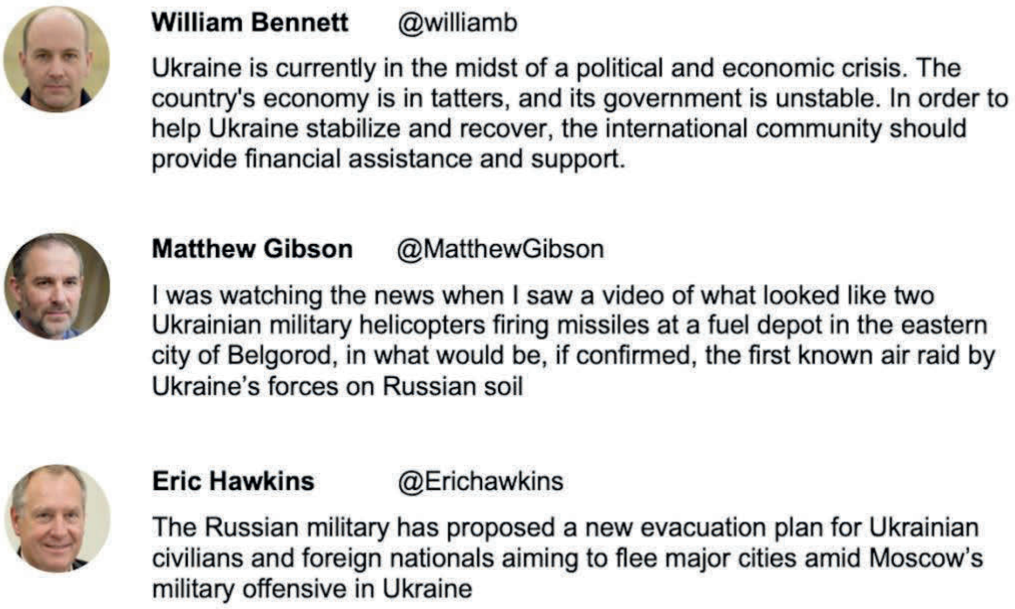
2.3. Experiments and fake content on social media

Experiments are a common method in studies related to deception and misinformation on social media. We identified two main approaches for experiments in social media, which are either conducting experiments directly within the social media platform (Cresci et al., 2017; Freitas et al., 2015; Shafahi et al., 2016) or alternatively by using a survey or other simulated environment such as a web-based game (Moravec et al., 2019; Roozenbeek & Linden, 2019).

While conducting the experiment with a simulated social media page can limit realism, they are generally less risky due to the environment being controlled and since debriefing is possible as well as acquiring informed consent from participants. Failure to take the appropriate measures has resulted in criticism of such research in the past (Flick, 2016). Therefore, for this study we preferred simulated environments to reduce potential ethical concerns.

Experiments with similar methods and goals to this study have been conducted in social bot research, where researchers have used crowdsourcing to determine whether humans can detect different types of bots, such as social bots and spambots (Cresci et al., 2017; G. Wang et al., 2013). The main difference in these studies was that the profiles of the bots were not generated using deep learning, and that the participants had access to view complete profiles, while in our study participants are shown a view similar to social media feeds. We argue that showing only what is visible in the feed is more realistic, as the average user might not go meticulously through each profile that they come across.

Other related experiments are the previously mentioned tests on the human ability to detect generated content in the context of poems (Köbis & Mossink, 2021) and profile pictures (Nightingale & Farid, 2022), which have shown that individual components similar to those in social media profiles can fool humans, but based on our knowledge no studies have yet at the time of writing been conducted on the human ability to detect fully deep learning-generated profiles.



William Bennett @williamb

Ukraine is currently in the midst of a political and economic crisis. The country's economy is in tatters, and its government is unstable. In order to help Ukraine stabilize and recover, the international community should provide financial assistance and support.

Matthew Gibson @MatthewGibson

I was watching the news when I saw a video of what looked like two Ukrainian military helicopters firing missiles at a fuel depot in the eastern city of Belgorod, in what would be, if confirmed, the first known air raid by Ukraine's forces on Russian soil

Eric Hawkins @Erichawkins

The Russian military has proposed a new evacuation plan for Ukrainian civilians and foreign nationals aiming to flee major cities amid Moscow's military offensive in Ukraine

Figure 2. Examples of the generated profiles shown during the experiment

3. Research design

In this section, we will explain the setup of the experiment and then describe the process used to generate the fake profiles and tweets, as well as how the real profiles and posts were collected.

The experiment imitated a situation where a Twitter user is scrolling through the feed and sees a post and the limited set of profile information about the account that posted it. The manipulations were built into pages hosted by Qualtrics using images of profiles along with the posts as shown in Figure 2. Participants were recruited using Amazon's Mechanical Turk (MTurk).

This simulated approach to having a mock Twitter feed was chosen for two reasons. First, this ensured that the auto-generated profiles and posts were not seen by any non-participating individuals. This could have been an issue as some of the generated content can be described as misinformation or otherwise controversial. Second, in this way we avoided violating the terms of use of Twitter and GPT-3, the deep learning-based language model used to generate texts of the posts.

The topic of the posts shown during the experiment was the war in Ukraine. It was chosen because of the timeliness of the topic and since a vast number of suitable real tweets were readily available from verified accounts. The style of the real tweets and accounts that were selected for the experiment are described in detail in the methodology section. While the posts discussed the war and sometimes contained questionable content, the experiment did not have any mentions of violence or other forms of graphic or disturbing content. Participants were also warned of the subject before being shown the posts and given an opportunity to back out without any consequences.

Due to the controversial nature of the topic and possible differences in views of participants, we designed the experiment to include real and generated profiles with tweets supporting both sides. During the analysis, we checked that the personal view of the participant did not correlate with how they rated perceived accounts as genuine or fake.

To reduce low-quality responses and to ensure the survey worked smoothly, we initially ran a limited trial experiment with approximately 100 subjects. We then improved the survey based on feedback received during trial. The experiment had a screening phase which was used to recruit US citizens that speak English fluently and that have experience with social networking sites, including Twitter. Lastly, the experiment was designed to be completed in a short time to reduce fatigue or learning effects. To incentivize participants to respond properly, they were instructed that the experiment includes an attention check and that identified cases of rushing through the survey or failing to respond adequately would result in no reward. The survey was terminated upon failure to answer correctly on the question containing the attention check. The attention check was placed near the beginning of the survey to not waste the time of inattentive participants. Ultimately all participants who passed the screening, passed the attention check and completed the survey were given the reward regardless of their performance.

During the experiment each participant was told that they are going to be shown profiles that belong to humans or bots, with the bots being the accounts that are deep learning-generated. The participants were shown four profiles, which were drawn from a total of 9 bots and 9 humans. For each profile shown, the participants were asked to label whether the account is a "bot" or a "human" and to rate their perceived likelihood of the account being a bot as well as to score different components in terms of whether they make the account seem suspicious, on a scale of 1 (not at all suspicious) to 10 (extremely suspicious).

We opted to have only white adult male profiles with common English names both in the real and generated profiles, to eliminate the influence of gender, race or perceived nationality on the results. Furthermore, fields such as the time of the post, likes and retweets were removed to control for their effects. While the number of comments, likes and retweets can have an impact on the credibility of a post, these can also be inflated with the use of bots or bought interactions from follower farms.

The following two sections describe in detail the process used to generate the fake profiles as well as how the real profiles were collected.

3.1. Generating fake profiles and tweets

The generated profiles and accompanying posts consisted of four elements, which were created using an automated script and with as little human intervention or tweaking as possible. This was done to imitate the mass production of fake accounts. The four generated elements were the profile picture, name, handle and post (tweet). Out of these, the profile picture and post were generated using deep learning-based methods and the name and handle with a basic script written in Python.

The profile pictures were scraped from the website “thispersondoesnotexist.com”, which produces a unique image every time the page is visited using StyleGAN, an advanced generative adversarial network (GAN) model. This crude approach demonstrates how easy it is to get a large sample of fake images generated by deep learning. While the method is straightforward, the underlying GAN model itself is an innovative approach to image generation. GAN models consist of two separate neural networks, a generator that produces images and discriminator that attempts to classify real images from the synthetic given to it by the generator. The discriminator provides feedback to the generator, which adjusts its parameters until the produced images are no longer detected by the discriminator (Creswell et al., 2018).

The posts were generated using OpenAI’s Generative Pre-trained Transformer 3 (GPT-3), which is an advanced deep learning-based NLP model that can be used among other things to produce high-quality text, based on prompts (Brown et al., 2020). GPT-3 has 175 billion parameters and has been trained with several massive datasets consisting of for example Wikipedia pages, text collected by crawling the internet and two large internet-based books corpora (Brown et al., 2020). As an autoregressive transformer model, given some input it can predict very accurately for example what words would complete a sentence or what is an appropriate response to a question.

While officially GPT-3 bans its use for generating any content including posts that are shown on social media, we applied and received an exemption to use it for this purpose in our experiment. The posts were created by prompting the model to summarize news articles related to the war in Ukraine with slight randomized variations in the parameters. Examples include asking the model to write the summary as a positive or negative opinion or by requesting that it explains the given text in language understandable to children. Due to the sensitive nature of the topic, we manually checked each text before including it in the experiment, to determine whether it contained any kind of violence or any other kind of objectionable content.

The two final and related fields shown in the experiment are the username and handle. The names consisted of a first and last name that were generated with a script that used a list of common US names. The handles were then derived from the names using a stochastic process that cut the first or last name to only the initial, and occasionally added random numbers to the end. After qualitatively inspecting the produced names and handles, they were deemed sufficiently similar to those of real Twitter users.

3.2. Collecting real profiles and tweets

The real profiles shown during the experiment mainly belonged to verified profiles of journalists, celebrities, pundits and politicians, whose identity and status as real humans could be easily verified and whose posts were public. The profiles included had to have a clear profile picture containing their face as well as full names so that they were comparable with the generated profiles and would not introduce any unnecessary noise to the experiment and influence the results.

The profiles were collected by retrieving verified profiles tweeting about Ukraine. These were then manually checked and included in the experiment if they met the criteria regarding the profile picture and name described above. Lastly, to reduce the chance of the real profiles being too well-known and thus recognizable by the participants, we removed profiles that belonged to very high-ranking politicians such as ministers or leaders of states as well as accounts which multiple authors could recognize.

4. Results

In this section, we first describe the demographics of the participants in the experiment and then examine the classification accuracies and assess the participants' ability to identifying the fake/real posts. Lastly, we discuss the perceived suspiciousness of the components according to the participants.

4.1. Participants

The results presented in this section are from the experiment held in May 2022 through MTurk. Out of 1292 subjects who participated in the screening, 478 were invited to complete the experiment. Ultimately, 375 participants both completed the experiment and passed the attention check at the beginning of the survey. The results discussed are only for the 375 subjects that successfully completed these steps. The average duration that it took to complete the experiment was 10 minutes and 44 seconds.

The participants predominantly identified as white (83.5%) and with a slight skew towards males, representing 56% of all subjects. The average age was 38 years and a most subjects (85.3%) have at least a bachelor's degree, meaning that the participants are more educated than the US population on average. The demographic information is summarized in Table 1. The lack of non-white participants can introduce bias to the results, making generalizing the findings to the general population risky. Therefore, these findings will be more relevant to assessing the capability of this particular demographic's ability to detect fake content.

To check that the experiment and survey's designs and instructions were sufficient, each participant had to rate both the clarity of the instructions and the perceived difficulty of the task after completing the survey on a 5-point Likert scale. When asked if "the tasks and instructions were clear," 96% responded either agree or strongly agree. When asked if "the given task was easy to do," the result was more spread with 83% stating that they agree or strongly agree, 11% neither agreeing nor disagreeing, and the remaining 6% disagreeing or strongly disagreeing. Based on these results the instructions were adequate. Despite the poor performance in terms of classification accuracy, only a small share of the subjects viewed the task difficult.

Table 1. Demographics

Category	Sub-Category	N	%
Gender	Male	210	56,0 %
	Female	164	43,7 %
	Other	1	0,3 %
Age	19-29	41	10,9 %
	30-39	132	35,2 %
	40-49	104	27,7 %
	50-59	51	13,6 %
	60-69	39	10,4 %
	70+	8	2,1 %
Race / Ethnicity	White	313	83,5 %
	Asian	21	5,6 %
	Hispanic	19	5,1 %
	Black	11	2,9 %
	Native American	5	1,3 %
	Other	6	1,6 %
Highest Degree	Primary school	1	0,3 %
	High school	54	14,4 %
	Bachelor's	258	68,8 %
	Master's	57	15,2 %
	Doctoral	5	1,3 %

4.2. Classification accuracy and likelihood

Since each participant was shown 4 randomly drawn profiles, the number of times each profile was labeled during the experiment varied from 77 to 85. None of the eighteen profiles received unanimous labels and when inspecting the accuracy by profile (i.e., the percentage of time it was correctly labeled), for the generated accounts the accuracies ranged from 10% to up to 27.4%. The generated profiles had a mean of 18.2% and 95 percent confidence interval (CI) at [0.145, 0.219]. The genuine profiles had accuracies ranging from 58.5% to 91.4% with a mean of 79.7% and a

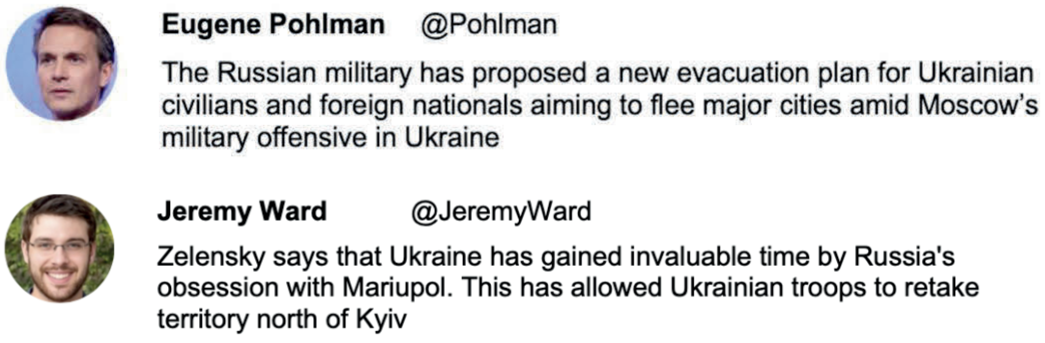
95 percent CI at [0.737, 0.856]. This suggests that the participants were unable to reliably detect the generated profiles. Interestingly, the most divisive accounts belonged to genuine humans. One of the real profiles was mislabeled by 41.5% of the participants that saw it. Meanwhile, the two best-performing fake profiles shown in Figure 3 were labeled as bots by only 10% and 11.1% respectively. The mean accuracy considering all profiles was 48.9%, which, given that there was an equal amount of genuine and generated profiles, is close to a random guess. The survey results are summarized in Table 2.

The participants were also asked to rate the likelihood of the account being a bot on a 5-Point Likert scale. This was done so that the level of certainty for the labeling could be assessed, with 1 being very unlikely and 5 very likely. The mean likelihood given to the generated profiles was 3.19 while for the genuine profiles it was 3.26. This indicates that the participants were on average uncertain of their labeling, regardless of whether accounts are bots or humans.

Table 2. Classification accuracy

Accuracy	Generated	Genuine
Mean	18,2 %	79,7 %
95% CI	14.5% - 21.9%	73.7% - 85.6%
Highest	27,4 %	91,4 %
Lowest	10,0 %	58,5 %

The difficulty of the task was brought up by some of the participants in their qualitative comments that they could write for each profile. Examples of these comments include the following: “I had to read and reread the tweet trying to understand what they were trying to say. Possibly a person, but it feels like it could be a bot.” and “Once again it is impossible to tell.” as well as “Too hard to tell.”



Eugene Pohlman @Pohlman

The Russian military has proposed a new evacuation plan for Ukrainian civilians and foreign nationals aiming to flee major cities amid Moscow’s military offensive in Ukraine

Jeremy Ward @JeremyWard

Zelensky says that Ukraine has gained invaluable time by Russia's obsession with Mariupol. This has allowed Ukrainian troops to retake territory north of Kyiv

Figure 3. The least detected generated profiles

4.3. Ratings of the components

As mentioned in the research design, participants were asked to rate each of the four different elements, the profile picture, post, name and handle in terms of how much it makes them suspect the account is a bot, with a slider ranging from 1 (not at all suspicious) to 10 (extremely suspicious). The results gave no real indication of any component being seen as a giveaway for either a human- or a generated account. For both classes, the mean results were ranging from 4 to 5 for all components. The highest scores were on the tweets for both genuine and generated profiles, with the mean value being 5.05 and 4.89, respectively. The lowest scoring component for both was the profile picture, with the means being 4.23 and 4.11, with the latter being the value for generated profiles. Table 3 summarizes the results and shows the 95% confidence intervals (CI) for the ratings.

We performed statistical analysis (ANOVA) to determine if there were any relationships between the ratings of the components and the classification but found no statistically significant results. Considering the complexity of the task and that the values had little variation as all four components were on typically given values between 4-5, this result is not surprising.

Table 3. Suspiciousness of components

Profile	Generated	Genuine
Picture	4,11	4,23
95% CI	3,88 - 4,33	3,96 - 4,51
Tweet	4,89	5,05
95% CI	4,52 - 5,25	4,60 - 5,49
Name	4,15	4,4
95% CI	3,92 - 4,38	4,24 - 4,56
Handle	4,4	4,59
95% CI	4,18 - 4,62	4,40 - 4,78

Rating scale: 1 (not suspicious) to 10 (extremely suspicious)

5. Discussion

Research question 1 asked whether humans can detect fully deep learning-generated social media profiles and posts on the feed of a social networking site. This study finds that accounts with GAN profile pictures, names drawn from a random name generator, and posts made with GPT-3 could not be distinguished from tweets and profiles created by real humans. To the best of our knowledge, this is the first time that this has been tested for a "whole" profile and not just for a generated face or text. Similar to the results of a recent experiment with GAN profile pictures alone (Nightingale & Farid, 2022), the generated profiles were viewed as more likely to be humans than the genuine human profiles. This can be explained by the fact that generated content tends to produce "average" looking data points, which in this case are the components of the profiles. At the same time, real data, i.e., the genuine profiles in this study, have more variety in components and human evaluators can make the mistake of assuming this type of noise is a sign of the generator having made an error. However, a larger sample of accounts and participants would be needed to determine if this result is generalizable or simply a result of the 18 accounts having a particular distribution of components.

The second research question, which asked whether some of the generated components can reveal that a profile is fake to a human evaluator, was left unanswered. However, the findings made in relation to RQ2 further supports the conclusion that humans are not able to distinguish real and generated profiles, as none of the generated profiles were detected by a majority of the subjects, and the ratings of the components' suspiciousness were on average very close to the central "neither nor" value.

Although the focus of this paper is not on the process of developing the fake profiles, we want to point out the accessibility and availability of the tools described in the methodology section. The ease of producing both the fake posts as well as the profile pictures was staggeringly easy. While creating and maintaining social bots would, without doubt, require intermediate to advanced programming skills, producing the components for fake profiles and posts would not, and even individuals without much training could build multiple seemingly humanlike profiles. This is primarily due to several reasons: 1) the availability of GAN-images through websites that demonstrate StyleGAN, 2) the modern tools for text generation such as GPT-3 that have no-code user interfaces, and 3) apps hosted on webpages can be used to access them even without personal access to the API. Therefore, the emergence of a growing number of realistic fake profiles is possible unless companies such as Twitter and Meta begin more actively detecting and removing profiles that, for example, have been algorithmically detected to have GAN-generated profile pictures.

These findings raise an important research impact question. What can be done to address this issue? Suppose humans cannot detect fake profiles and posts and report them manually. In that case, the role of automated detection and development of other safeguards by the companies operating the social networking sites becomes more important than before. This is due to online communities no longer being able to support moderation on their own by flagging suspicious content. Ultimately, this increases the responsibilities of social networking sites.

Moreover, one could also question whether the companies producing tools that can be used to create computational propaganda are also accountable. It should be noted that the use of GANs and text generators for malicious purposes on social media is well beyond the intentions of their respective developers and that these technologies have many potential beneficial use cases meriting their development. Moreover, companies such as OpenAI has even specifically banned using GPT-3 for the generation of offensive texts and social media content, and access to the model is terminated when infringements are caught. As evidence of this policy being enforced, during the development of this paper one of the authors had their API keys and access to the system revoked.

As a conclusion, we believe that while the availability of tools that can be used to create malicious content at scale could in theory be limited, most of the technology or alternatives to them are already published as open-source software, and thus putting the genie back into the lamp is impossible so to speak.

5.1. Theoretical implications

While this paper is purely an empirical study, the results can have strong implications for several theories assuming the hypothesis holds. For instance, the severity of the spiral of silence (Noelle-Neumann, 1974) could be enhanced by an influx of humanlike malicious accounts. The role of bots and fake accounts and their impact on the formation of what is the general public opinion has been studied and it has been shown by simulation that a relatively low percentage of bots can tip the discussion and trigger a spiral of silence, where a small but loud group define the perceived prevailing opinion (Ross et al., 2019). In other words, in the context of this paper, it is possible that if bad actors could create realistic-looking profiles and posts at scale, they could use them to distort the perceived public opinion.

5.2. Limitations

The main limitations of this study are the small number of visible components per account and the homogeneity of the profiles as all were white adult males. Moreover, the participants were mainly from a narrow demographic, as over 80% were white. These reduce the realism and generalizability of the results but were nevertheless deemed acceptable given the scope of this paper. We address the limitations with the following arguments.

First, our goal was not to determine if humans can recognize the generated profiles when given full access to the profiles and historical posts, but rather to emulate a situation where a user of a social networking site scrolls through the feed and sees multiple posts made by different users. If the profiles are not suspicious, it is unlikely that an individual would go through each profile in detail. Thus, the realistic generated profiles could pass as genuine users, and for example affect the individual’s perception of what is the common opinion on a specific matter.

Second, while in a more realistic setting the experiment would have both generated and genuine profiles of various gender, ethnicity and origin as well as some with missing or less similar profile information, this would introduce too many variables that can influence the results. This could be addressed by conducting multiple experiments or introducing a significantly larger sample size. This will be addressed in our future work, which is described in the following section.

Finally, when recruiting participants, we opted not to attempt to reach a particular distribution regarding the demographics as we assumed we would in later experiments be able to make it more evenly distributed. Due to the low acceptance rate to the experiment, finding larger numbers of participants from less common demographic groups would have taken significantly more time.

5.3. Future work

After having conducted a pilot as well as the experiment presented in this paper, we have determined that the design of the generated profiles seems sufficient. However, the scope of the topics of the posts as well as the diversity of the profiles posting should be expanded. Previous work has suggested that GAN-generated images with non-white and female individuals are less realistic and easier to detect due to biases in the training data (Nightingale & Farid, 2022). Therefore, it would be interesting to see if this pattern remains in a richer setting where the profile pictures are accompanied by information such as a name and post.

Moreover, introducing a treatment where some participants would be given instructions on how to spot fakes could be used to determine if subjects can learn to detect fake profiles based on different components such as profile pictures or the text in the posts. This could provide valuable insights for scholars and practitioners on how to combat computational propaganda by providing users of social networking sites with appropriate instructions and training.

To reduce the possible bias of the respondents, we plan to recruit a more diverse set of participants in future experiments in both MTurk and by running the experiment using students in different regions. This will allow us to produce more robust findings as well as potentially reveal differences between groups of humans.

Lastly, to increase the realism of the simulated social media feed, the user interface will be upgraded in upcoming experiments to include more elements and the possibility to view the profile description of the accounts, as this can be done also on Twitter when hovering the cursor over a profile. This will require adding additional components for the participants to view, such as how many followers and how many accounts the profile is following, as well as the profile description, which is also known as the bio.

6. Conclusion

Previous experiments have shown how it is possible to create fully synthetic yet real looking pictures of faces with generative adversarial networks as well as machine generated texts, using pre-trained language models such as GPT-3 that are indistinguishable from those written by a human. In this paper, we attempted for the first time, as far as we know, to produce realistic social media profiles using these two methods to demonstrate that we have passed the point where fully generated posts and profiles can pass unnoticed by humans in a social media feed. The results of our experiment support this hypothesis as the classification accuracy was consistently low for the generated profiles. Since the generated profiles were mostly classified as genuine profiles during the experiment, we could not determine if individual components of the profiles could indicate to humans which profiles are real humans and which are generated.

However, we are careful of making strong claims or generalizing based on the results until further experiments are conducted and some of the limitations of this study are addressed. While we believe that detecting generated content and fake profiles in the feed is difficult, we hypothesize that if given access to full profiles it would be much easier for humans to spot suspicious accounts.

We believe though that most humans would not go through the effort of checking each profile they come across in a feed, and thus the results of this paper can be considered concerning. Ultimately, this study suggests that making believable fake profiles with minimal human involvement is possible. Considering that fake profiles can distort online discussions and efficiently spread misinformation (Bessi & Ferrara, 2016; Shao et al., 2018), automatic detection of removal of such accounts should be the top priority of social networking sites as the end user cannot be expected to distinguish fake from real.

7. References

Bessi, A., & Ferrara, E. (2016). Social Bots Distort The 2016 U.S. Presidental Election. *First Monday*, *21*(11).

Bond, S. (2022, March 27). That smiling LinkedIn profile face might be a computer-generated fake. *National Public Radio*. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?t=1654174474533

Brachten, F., Stieglitz, S., Hofeditz, L., Kloppenborg, K., & Reimann, A. (2017). Strategies and influence of social bots in a 2017 German state election—A case study on twitter. *ACIS 2017 Proceedings*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's' Human'Is Not Gold: Evaluating Human Evaluation of Generated Text. *ArXiv Preprint ArXiv:2107.00061*.

Cresci, S. (2020). A Decade of Social Bot Detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972. https://doi.org/10.1145/3041021.3055135

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., & Nakov, P. (2021). A Survey on Computational Propaganda Detection. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. *Research Ethics*, *12*(1), 14–28. https://doi.org/10.1177/1747016115599568

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 25–32. https://doi.org/10.1145/2808797.2809292

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410. https://doi.org/10.1109/tpami.2020.2970919

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly: Management Information Systems*, *43*(3), 1025–1039. https://doi.org/10.25300/MISQ/2019/15188

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Li, J., Tang, T., Zhao, W. X., & Wen, J.-R. (2021, August 19). Pretrained Language Models for Text Generation: A Survey. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*.

Marx, J., Brünker, F., Mirbabaie, M., & Hochstrate, E. (2020). Conspiracy Machines–The Role of Social Bots during the COVID-19 Infodemic. *ACIS 2020 Proceedings*.

Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at All. *MIS Quarterly: Management Information Systems*, *43*(4), 1343–1360. https://doi.org/10.25300/MISQ/2019/15505

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, *24*(2), 43–51.

Roozenbeek, J., & Linden, S. van der. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10. https://doi.org/10.1057/s41599-019-0279-9

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Rossi, S. (2022). The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation. *The 55th Hawaii International Conference on System Sciences: HISS 2022*.

Salge, C. A. de L., Karahanna, E., & Thatcher, J. B. (2022). Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter. *MIS Quarterly*, *46*(1). https://doi.org/DOI: 10.25300/MISQ/2021/15598

Shafahi, M., Kempers, L., & Afsarmanesh, H. (2016). Phishing through social bots on Twitter. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 3703–3712. https://doi.org/10.1109/BigData.2016.7841038

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06930-7

Stieglitz, S., Brachten, F., Ross, B., & Jung, A. K. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. *ACIS 2017 Proceedings*, 1–11.

Strick, B. (2021). *Analysis of the Pro-China Propaganda Network Targeting International Narratives*. Center for Information Resilience. https://www.info-res.org/post/revealed-coordinated-attempt-to-push-pro-china-anti-western-narratives-on-social-media

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2013). Social turing tests: Crowdsourcing sybil detection. *Proceedings of The 20th Annual Network & Distributed System Security Symposium (NDSS)*.

Wang, X., Guo, H., Hu, S., Chang, M.-C., & Lyu, S. (2022). GAN-generated Faces Detection: A Survey and New Perspectives (2022). *ArXiv Preprint ArXiv:2202.07145*.

Williamson, W., & Scrofani, J. (2019). Trends in Detection and Characterization of Propaganda Bots. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 7118–7123. https://doi.org/10.24251/hicss.2019.854

Yu, N., Davis, L., & Fritz, M. (2019). Attributing fake images to GANs: Learning and analyzing GAN fingerprints. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7556–7566.

Paper IV

The following paper has been submitted and is under review at the *Journal of Information Technology*.

AI-Generated Profiles Are Indistinguishable from Real Profiles in Social Media Feeds

Sippo Rossi
University of Copenhagen
siro@di.ku.dk

Youngjin Kwon
Washington State University
youngjin.kwon@wsu.edu

Odd Harald Auglend
Independent Researcher
oha.digi@cbs.dk

Raghava Rao Mukkamala
Copenhagen Business School
rrm.digi@cbs.dk

Matti Rossi
Aalto University
matti.rossi@aalto.fi

Peter Ractham
Thammasat University
peter@tbs.tu.ac.th

Jason Thatcher
University of Colorado
Boulder
jason.thatcher@temple.edu

Abstract

Generative artificial intelligence (GAI) and foundation models have made creating realistic images and humanlike text possible. These technologies have already been adopted by social media bot developers, who are using tools such as GPT-4 to create social bots that are harder to detect. However, there is little research on the average human's ability to distinguish such bot profiles from genuine human accounts on social media. To study this, we conducted two experiments where participants were asked to label accounts as bots or humans in a simulated social media feed populated by real Twitter users and bots created using generative AI. Our findings show that humans cannot accurately distinguish genuine AI-generated bot profiles from genuine human profiles.

Keywords: Social bot, social media, generative artificial intelligence, experiment, artificial intelligence, AI

1. Introduction

Fake profiles and bots on social media remain pervasive [1], despite the attempts by social networking platforms to remove them [2]. While crude spambots and simple bots are easy to distinguish from their human counterparts, modern social bots are increasingly humanlike and more capable of passing undetected by humans [3]. This ability to emulate humans has made social bots a favored tool for malicious and benevolent actors to spread content at scale, which in turn has made detecting the bots a vital issue for sites such as X (formerly known as Twitter) and Facebook as well as for researchers attempting to study bots and information dissemination.

As bot detection methods have evolved, so have bots. Modern deep learning tools are being used to produce realistic profile pictures [4,5], and content for posts is being generated using the latest Natural Language Processing (NLP) techniques [6,7]. Such innovations have resulted in an arms race between bot developers and bot detectors, with researchers having to respond to increasingly sophisticated bots with new detection techniques [3,8,9].

While bot detection has evolved, not as much attention has been paid to the underlying phenomenon, which is understanding why bots can fool and ultimately influence humans exposed to them. In doing so, we patterned our work after prior studies that focused on whether humans can distinguish between genuine and AI-generated profile pictures [10,11] or text written by AI [12,13]. These previous studies have only focused on one form of generated content at a time rather than on a combination of them. However, there is a need to go beyond individual components, because generative AI has gone beyond simply constructing authentic feeling pictures to creating customized content to populate social bots' profiles and posts [7]. Hence, this study builds upon the previous research on the ability of humans to detect AI-generated content, by combining image and text within one experiment. The research questions of this paper are:

RQ1: Can humans distinguish AI-generated bot profiles from genuine profiles that are commenting on a social media post?

RQ2: Which features on a bot profile can humans detect are AI-generated and not genuine?

To build an understanding of a human's ability to detect bots, we assess the interplay between images and content in shaping their ability to detect bots. We build on research that has shown on a limited scale that humans find AI-generated bot profiles indistinguishable from real profiles when each social media post is rated separately (e.g., in isolation) [14]. To extend this work, we used the latest generative AI technologies to create fake social media profiles and then asked human participants to detect these AI-generated bot profiles in a simulated social media feed. To further increase ecological validity, these AI-generated bot profiles were shown within discussion threads that included genuine posts by real accounts mixed in with the AI-generated fake bot accounts and posts.

Given recent advances in the ability to generate realistic pictures and text, we posit that most participants cannot distinguish between AI-generated and genuine profiles in a social media feed. If participants successfully detect the AI-generated profiles, then we evaluate whether features such as the profile picture or content of the post influenced their decision. In a sense, this study serves as a kind of revised version of the Turing Test [15], where instead of communicating by text with a bot and a human, the evaluator sees other social media profiles replying to a post and tries to determine which accounts are humans and which are AI-generated bots.

Understanding the conditions under which humans can detect bots is important, because some suggest that the labeled training data for supervised machine learning bot detection models can be created through crowdsourcing [16,17]. Early studies found crowds consistently detected bot profiles, especially when multiple individuals classified social media accounts using a majority vote system [18]. However, more recent works suggest that social bots are increasingly difficult to spot, even for people with experience classifying accounts [3,14]. The rapid development of applied deep learning, particularly in NLP and generative adversarial networks (GANs) for image and text generation, has likely further complicated bot detection because state-of-the-art social bots may emulate humans more effectively by using AI-generated content [7,9].

Our study departs from the computational literature on bot detection and instead directs attention to human evaluators. Even as platforms apply computational techniques to detect and remove many social bots, it is evident that many are still not caught by these automated detection techniques [7]. As a result, users will be exposed to bots and will need to assess whether an account that posts or comments on other posts on a social networking site (SNS) is, in fact, a bot or a human. When assessing an account's credibility, users will make decisions based on information found in the social media feed, such as profile pictures, names, and the content of posts. Our study aimed to emulate this process in the context of a user scrolling through a feed and seeing posts which have comments coming from both genuine human profiles as well as bots that have been created using generative AI.

Based on two experiments, we found that humans cannot distinguish bots with AI-generated posts and profile pictures from genuine Twitter profiles that belong to humans. Due to the poor performance in detecting and correctly labeling bot accounts as bots, we could not identify clear features that would lead to humans detecting bots. Thus, our study suggests that the average human can no longer detect modern bot profiles. This has implications for bot detection research, as many machine learning-based approaches have relied on human annotated training data. In other words, using crowdsourcing and majority voting to create a labeled dataset for training a bot classifier would no longer be effective. Moreover, for the average social media user, understanding the true opinion of other users will become more difficult to estimate if online discussions are infiltrated by bots pushing and promoting some agenda. The implications of our findings will be offered in the discussion section.

The remainder of our paper is organized as follows. First, we discuss the background of the study and provide an overview of recent advances in social bots and generative AI research in the literature review. Then, we describe the research design and provide an overview of the experiments that were conducted. This is followed by a presentation of the results, and the discussion of their implications to practice and theory. The paper concludes with remarks on potential future research.

2. Background and related research

In this section, we first provide a context for the study by discussing key studies on bots on social media. Then, we review recent findings on the human ability to detect deep learning-generated content such as images and texts. Lastly, to support our methodological choices, we summarize previous experiments focused on social bots and inauthentic social media content.

2.1. Bots on social media and their impact

As this research aims to make inferences on the human ability to detect advanced social bots in a social media feed, we will first briefly define what the somewhat ambiguous term, social bot, means. While diverse definitions and taxonomies for bots and social bots exist [19–21], we define social bots as accounts that are automated and that follow some model for controlling their behavior and or content that they produce. Social bot research has primarily focused on bots that are used for malicious purposes such as spamming, inflating follower counts [22] and influencing politics [23], but, arguably, social bots can be used for neutral or benevolent purposes as well such as information sharing or sales [5,24,25].

The impact social bots have on humans, for example, in the case of spreading fake news, has been a challenge to measure [26]. Some studies suggest that social bots' influence is minimal, and in the case of fake news and misinformation, humans pose a more significant threat than bots [27]. However, analysis suggests that bots represent a disproportionately larger share of accounts that share low-credibility content, such as fake news [26], as well as share content more rapidly and more successfully than regular accounts [29]. Furthermore, evidence suggests that bots amplify the initial spread of trending malicious content, helping it become viral [28]. Social bots have been used to manipulate online discussions and spread misinformation related to pressing topics such as the COVID-19 pandemic and vaccines [30,31]. Even more concerningly, a relatively small population of bots can tilt an opinion climate in social media and thus influence public opinion and perception [32].

Conducting an updated Turing Test with humans and social bots is important, as it helps researchers, platform owners, and policymakers understand whether humans know that it is a social bot authoring social media posts. The idea of a revised version of the Turing Test for social bots is not entirely new [18,33], but neither of the cited examples nor the methods employed by other studies identified in our literature review have become established. Due to the various existing papers which propose Twitter Turing Tests or Social Bot Turing Tests, we do not label our proposed experiments as Turing Tests, even though they share the same goal—assessing a human's ability to detect bots.

2.2. Human ability to detect generated content

Recent advances in large language models have made it possible to produce short but high-quality text, such as poems that are sometimes indistinguishable from those written by a human [34,35]. According to one study, some individuals can distinguish generated sentences from human written ones relatively consistently and at much higher rates than random chance [13]. At the same time, another suggests that text generated by the already outdated GPT-3 is no longer detected by humans [12]. Interestingly, one of the differences in these two studies is the participant pool, as one uses university students and the other individuals recruited via Amazon Mechanical Turk (MTurk), with the latter performing worse. As the study by Rossi et al. [14] used MTurk and suggested humans cannot identify fake accounts with GPT-3 generated posts, this paper tests the ability of both students and MTurkers to detect social bots, to see if there are differences in the results.

While powerful pre-trained foundation models for text generation have been accessible since late 2022, very few publications evaluate their capability to generate social media posts (besides [14]). Our search through Google Scholar and arXiv resulted in no papers that assessed humans' ability to identify AI-generated social media text content. Outside of controlled academic experiments, it is suspected that GPT-2 has already been used to produce propaganda texts in social media posts [6] and that ChatGPT has been used in a cryptocurrency scheme [7,36].

Because deep learning image generation methods matured earlier than text generation techniques, more studies examine the human ability to detect generated images than text. While images produced by generative adversarial network (GAN) models are detectable algorithmically for humans such images seem near photorealistic or even real [37,38]. Figure 1 demonstrates the quality of GAN models when applied to create realistic photos of humans. Similar to manual bot classification, few studies have evaluated the human ability to differentiate computer-generated images of humans from authentic photos, especially when compared to the extensive literature on computational detection of generated photos. However, recent studies have indicated that GAN-generated images of faces similar to those used as profile pictures are indistinguishable from real faces to humans [11] and even viewed as more trustworthy than authentic photos [10]. Moreover, academic and non-academic publications have examples of cases where numerous fake profiles using GAN-generated profile pictures have been discovered on various social networking sites [5,6,39].

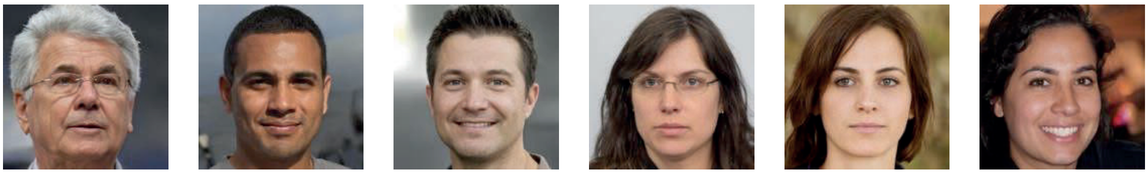


Fig 1. Examples of profile pictures generated with StyleGAN

Note: None of the people in the photos exist and any resemblance to actual individuals is coincidental.

2.3. Social media experiments

The definition of experiments as a research method varies by field, and social media experiments often take the more liberal definition as used by [10]. Experiments have been used to study bots as well as other forms of deception and misinformation on social media. They have been conducted within the "live" social media platforms [3,40,41], through test environments such as browser games [40], as well as with custom-built programs or websites that mimic real social networking sites [43].

For many social media studies, experiments in simulated environments or games are preferable and, in some cases, the only option due to the ethical issues tied to studying sensitive topics. For example, showing negative or harmful content to unsuspecting users to measure reactions or examine the spread of the content outside of controlled environments or even both, without the possibility of debriefing, is risky and ethically unacceptable. Studies conducted on social networking sites that have failed to consider such possible outcomes have resulted in criticism from researchers and those affected for the lack of informed consent [44]. Thus, experiments on controlled groups using different approaches like browser games or surveys are preferred due to the reduced ethical concerns as the participants are volunteers, are informed, and can be made aware of the nature of the experiment in advance or afterward.

For some studies, experiments can be conducted on the social networking site itself. Interesting examples have included social bots programmed by researchers to study infiltration methods and bot designs [40,41,45]. However, creating bot accounts goes against the policies of social networking sites and adds risks, as the experiment can end prematurely if the platform bans the researcher and the affiliated accounts. If successful, the main benefit of running experiments on actual social networking sites is the realism and, thus, ecological validity of the results. Due to the difficulty of measuring whether humans detect the bots in such a setting, this approach was deemed unsuitable for the goals of this paper.

Whether in a controlled environment or the field, few studies have used experiments to assess human ability to detect bots using a dataset or manipulation consisting of human and bot accounts. In two relevant examples, the participants were given access to the complete profiles rather than just a social media feed, and the content was not AI-generated [3,18]. The evolution of bots can be seen in the results as Wang et al.'s [18] experiment demonstrated the feasibility of crowdsourcing bot detection. In contrast, the more recent experiment by Cresci et al. [3] showed that even individuals specifically selected based on their ability to identify fake accounts could not accurately label social bots. As our work focused on profiles with purely AI-generated posts and profile pictures, in contrast to previous works which were conducted before these were available, thus rendering the bots much simpler, we believe our study provides novel information on the human ability to detect advanced social bots, rather than attempting to reproduce previous findings.

3. Method

In this section, we describe the procedure used in the experiments. Then, we elaborate on the methods used to produce the fake profiles and their posts. Lastly, we explain how the genuine social media profiles and posts were collected from Twitter. The institutional review board reviewed the procedure for collecting and generating fake profiles at one of the author's institutions. They complied with ethical standards and requirements for studies conducted with human subjects.

3.1. Procedure

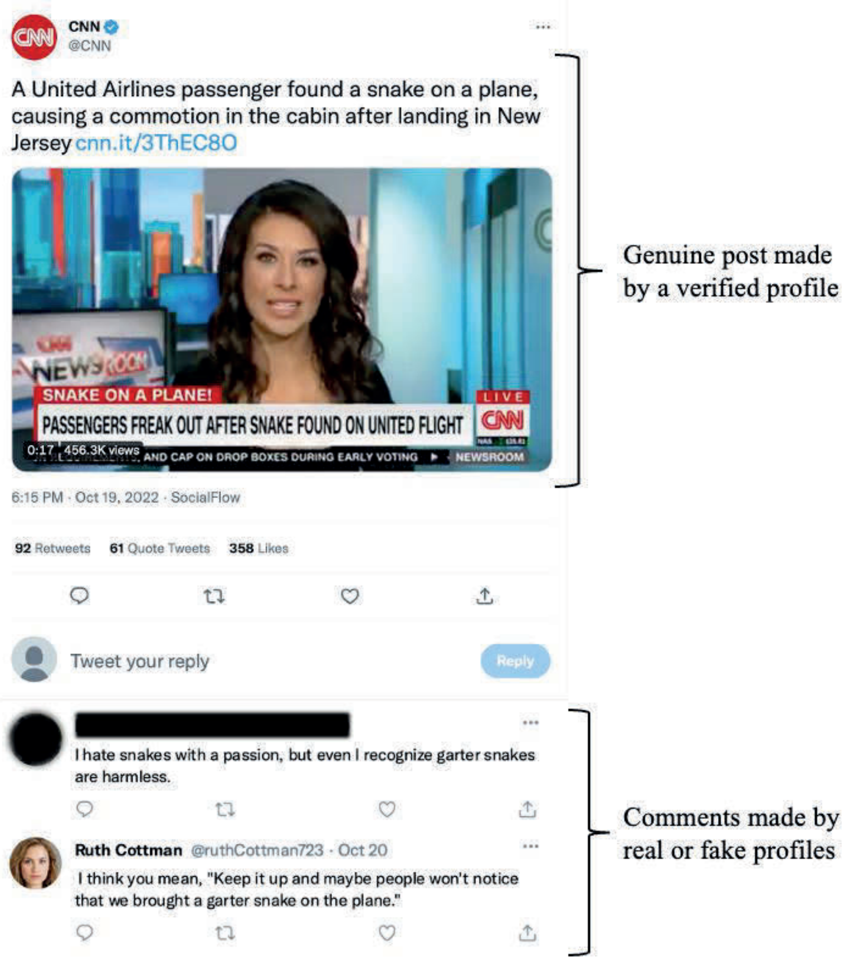
This study uses two experiments to probe whether humans can detect AI-generated social bots. We did so for two reasons. First, to rule out sample source as an explanation for variation. Previous studies on detecting generated content had varied results depending on whether the participants were students or recruited via Amazon Mechanical Turk (MTurk). Second, to evaluate our research question more rigorously because university students can be reasonably homogeneous and offer high internal validity while MTurkers can be more heterogeneous and offer high external validity. Thus, one experiment was administered to participants recruited through Amazon Mechanical Turk, and the other was administered to participants recruited from university courses.

The experiments were administered using a Qualtrics survey. The procedure was identical in both experiments, except that participants recruited through Amazon Mechanical Turk were prescreened with a pre-test to ensure they were not bots and understood English. The experiment starts with a warning about the content and informs the participants that they may opt out now or at any point during the experiment. This warning was followed by two screens containing explanations of the task and terminology within the task, which have a timer to encourage thorough reading before allowing to proceed to the next section. See Appendix A for full descriptions of the explanations. Lastly, before the actual experiments, the participants are asked to commit to thoughtful answers throughout the experiments as a soft attention check, which, if they refuse to commit, would lead to the end of the survey.

In the experiments, participants were shown a random sample of two artificial social media feeds, with each feed containing one main genuine post made by a verified user and two comments on the post. The comments belonged either to a genuine (made by real users of Twitter) or a bot account that the research team generated. Participants saw randomly assigned variations with some seeing only human or bot made comments, or a mix of both from a pool of 12 profiles in six different variations of the feed. See Figure 2 below, which contains one of the feeds used in the experiment.

The participants labeled two commenting accounts as a human or a bot. Labelling was followed by a second set of questions where they rate the likelihood of the account being a bot as well as each of the four visible features (profile picture, post, name, handle) for how much they contribute to the suspicion of the account being a bot. The task was kept short to reduce fatigue and learning effects (e.g., participants sorting out how to detect bots based on the questions).

After evaluating the feeds, the participants were asked to rate the clarity of the instructions, the difficulty of the task and to provide basic demographic information such as their gender, ethnicity, age, and highest earned degree. The participants also provided information on the frequency of their social media use, such as how often they use Twitter. The experiment concluded with a submit responses button, which needed to be pressed by the participants for the response to be considered valid.



CNN
@CNN

A United Airlines passenger found a snake on a plane, causing a commotion in the cabin after landing in New Jersey cnn.it/3ThEC8O

NEWSROOM
SNAKE ON A PLANE!
PASSENGERS FREAK OUT AFTER SNAKE FOUND ON UNITED FLIGHT
LIVE
CNN
0:17
456.3K views
AND CAP ON DROP BOXES DURING EARLY VOTING
NEWSROOM

6:15 PM · Oct 19, 2022 · SocialFlow

92 Retweets 61 Quote Tweets 358 Likes

Tweet your reply

[redacted] [redacted]

I hate snakes with a passion, but even I recognize garter snakes are harmless.

Ruth Cottman @ruthCottman723 · Oct 20

I think you mean, "Keep it up and maybe people won't notice that we brought a garter snake on the plane."

Genuine post made by a verified profile

Comments made by real or fake profiles

Fig 2. Examples of a simulated social media feed from the experiment

Notes: The genuine account's profile picture, name, and handle have been removed to maintain privacy.

3.2. Collecting the genuine profiles and posts

The setup of the experiment required collecting two sets of genuine posts and profiles. One set consists of the main posts made by verified users such as accounts owned by news media such as CNN or organizations such as NASA. The second set consists of accounts commenting on the posts made by the verified users. We collected the accounts and posts through a manual search of Twitter, looking for profiles with both a humanlike profile picture and account names containing both first and last names. Furthermore, to ensure we are not collecting bot profiles, each profile was checked to belong to an actual human based on their general behavior and visible profile information that links them to a real person (e.g., multiple online profiles with consistent names and photos), and by having a unique non-generated profile picture that cannot be found in other places via a reverse image search. To ensure the privacy of the commenting profiles, we cannot share the names, handles, or images of the accounts in this paper. Also, they were only visible during the experiment in images seen by participants. Moreover, each post was deemed neutral and thus incapable of negatively affecting a participant's perception of the person who wrote the post.

3.3. Generating the fake profiles and posts

Four features were generated for the social bot profiles so that they would have the same information visible as the genuine profiles. First, to create the profile pictures for the fake profiles, we used a script that visits the website "thispersondoesnotexist.com" multiple times and in each visit downloads the generated image that the website provides (it generates one unique image for each visit to the page). The website uses Nvidia's StyleGAN model to produce realistic images of human faces, similar to the typical online profile picture. Photos were rejected and not used in the experiment in the rare event that they contained a clear flaw in generation (such as two faces) or belonged to a child.

Two scripts were used to produce the textual features. First, a basic Python script was used to generate random names and handles, where the first and last names were drawn from a pool of US names [46]. Then random capitalization was added, and for some handles, a random number similar to those commonly present on Twitter (e.g. John12345678) were added, so the generated name and handle pairs were similar to those found on genuine users. The second script, that produced the posts, was connected to GPT-3's API [47] and would consider both the main post and the previous comments in it and attempt to "continue the discussion" by producing a tweet length text. While producing social media posts is against the terms of use of GPT-3, we received permission from OpenAI to use it for this research project (personal communication, February 9, 2022).

After all four features were produced, a final script combined them into a table, with each row containing the data used to populate the fake profiles. To produce the simulated social media feed, we first produced the individual posts using an online tool that enables the creation of fake Tweets by inputting the desired content of the post and the profile making the post. Thus, we inserted the collected real and fake profile information into the tool and produced images of the tweets, which were then combined into a thread of comments under the main post, as shown earlier in Figure 2.

4. Results

In this section, we present the results of the two experiments. Experiment 1, which had university students as participants, is presented first, followed by Experiment 2, which had participants recruited via MTurk.

4.1. Experiment 1

4.1.1. Participants

For the first experiment, participants were recruited at lecture or lab sessions of several university courses. 231 participants (126 female, 104 male, 1 other) completed the task. 35 responses were excluded because the participant quit before submitting the results. To incentivize participants to answer correctly, they were told that the best performing participants could win a restaurant gift card before participating in the experiment. The full table of demographic information is provided in Table 1. Most participants considered the instructions clear, as 81.2% agreed or strongly agreed with the statement “the study’s instructions and tasks were clear.” To the statement “the given tasks were easy to do” the participants responded with more variance as 48.9% agreed or strongly agreed, 26% neither agreed nor disagreed, and the remaining 25.1% disagreed or strongly disagreed, indicating that many students felt that distinguishing bots from humans was difficult.

Table 1

Demographics of participants who completed the task

Category	Subcategory	N
Gender	Male	104
	Female	126
	Other	1
	Total	231
Age	Mean	22.57
	Median	22
	Range	18 - 56
Ethnicity*	White	110
	Asian	110
	Middle Eastern or Northern African	6
	Hispanic	3
	Black	2
	Hawaiian or Pacific Islander	0
	Other	5
Education	High school	58
	Bachelor’s	156
	Master’s	16
	Doctoral	1

* Participants can pick multiple

4.1.2. Ability to distinguish and correctly classify profiles

In the experiment, we used six AI-generated bot accounts and six genuine human profiles. Overall, the participants did not perform well in bot detection, as the average accuracy in classifying the accounts as bots and humans was 56%, which is slightly above random chance in a binary classification problem. The best-performing bot account was correctly classified by only 6.3% of the participants who saw it. In comparison, the most misclassified genuine human account was labeled as a bot by 53.9% of the participants who saw it. In their estimate of the likelihood of the account being a bot, the participants were very uncertain, and the average response was 2.92 on a scale of 1-5, with 3 being neither unlikely nor likely. When rating the suspiciousness of each of the four components (profile picture, tweet, name, handle), the average response for all of these were 3.4, 4.16, 2.23 and 3.78, respectively, on a scale of 0 to 10, suggesting that none of the features stood out as an indicator of the account being a bot. Full results by profile are listed in Appendix B.

To statistically analyze the data collected from students, we employed a generalized linear mixed-effects model (GLMM), an analytical approach tailored to the structure of our data. The nature of our data informed the decision to use a mixed-effects model. First, we include the profile conditions (AI-generated or human profile) as a fixed effect in the model. Second, as the subjects encountered four out of 12 profiles, subject- and profile-level random effects are included to control for varying effects of subjects and profiles on the dependent variables.

Additionally, the response variable of our primary interest has a binary outcome, which is 1 (one) if a profile is correctly identified as AI-generated or human; otherwise, 0 (zero). This dichotomous nature violates the normality assumption of the linear mixed-effects model, thereby justifying the use of a GLMM. We fitted our GLMM using the glmer command from the lme4 package [48] with the binomial() option in R [49]. In addition, we used the ggplot2 package to plot interaction results [50].

The results of the GLMM with subject- and profile-level random effects show that the effect of profile type was statistically significant. The full results are visible in Table 2. The student subjects show a lower rate of correctly identifying AI-generated profiles than genuine human profiles, b = -.91, SE = .38, p = .02 (Model 2). This analysis suggests that when all other factors are held constant, the odds of correctly identifying profiles decrease by 59.7% (e^(-0.910)= .40) for the profiles generated by AI, compared to genuine human ones. In addition, the likelihood ratio test that compares pseudo-R2 indicates that the hypothesized model with the profile fixed effect (pseudo-R2 = .05) explains the correct identification of profiles better than the null model without it, χ^2(1) = 4.67; p = .03[17].

We replicated the main analysis to corroborate this finding, including only (a) subject-level random effects and (b) profiles-level random effects in the GLMM. We found results that affirm the main analysis. Specifically, compared to genuine human profiles, AI-generated profiles are associated with a lower rate of correct identification in the GLMM with subject-level random effects b = -.86, SE = .14, p < .001; and with profile-level random effects b = -.87, SE = .37, p = .02. We found a consistent pattern in the likelihood ratio test. The hypothesized model with the fixed effect and subject-level random effects explains (pseudo-R2 = .05) better than the null model, χ^2(1) = 38.02; p < .001; the model with the fixed effect and profile-level random effects (pseudo-R2 = .05) shows a statistical significance, compared to the null model, χ^2(1) = 4.67; p = .05.

We tested the moderating effects of suspicion of the profiles on the relationship between profiles and the outcome variable. For this moderation test, we included the suspicion measure and its interaction term with the profile fixed effect in our GLMM. As can be seen in Model 3, the interaction term is statistically significant, b = .52, SE = .06, p < .001. Figure 3 presents the plot

[17] pseudo-R^2 of the fixed effects is reported across this research.

of the interaction result. Interestingly, higher suspicion led to a higher correction rate when AI-generated profiles were presented. On the contrary, subjects with high suspicion of human profiles were worse at correctly identifying the profiles than those with low suspicion.

Table 2

Generalized Linear Mixed Model–Student Sample

	Model 1	Model 2	Model 3
(Intercept)	.43 (.23)	.89** (.27)	1.52*** (.27)
Profiles FE	-	-.91* (.38)	2.73*** (.39)
Suspicion	-	-	-.22*** (.04)
Profile FE X Suspicion	-	-	.52*** (.06)
Subject RE	*Incl.*	*Incl.*	*Incl.*
Image RE	*Incl.*	*Incl.*	*Incl.*
Pseudo-R^2	.00	.05	.19

Note. N = 924; 231 subject groups and 12 image groups; * p <.05, ** p < .01, *** p < .001; standard errors are in parentheses; Pseudo-R^2 of the fixed effect is reported; *Incl*.: included; RE: Random Effects; FE: Fixed Effects; Profiles FE: 1 if AI-generated profiles or 0 if human profiles.

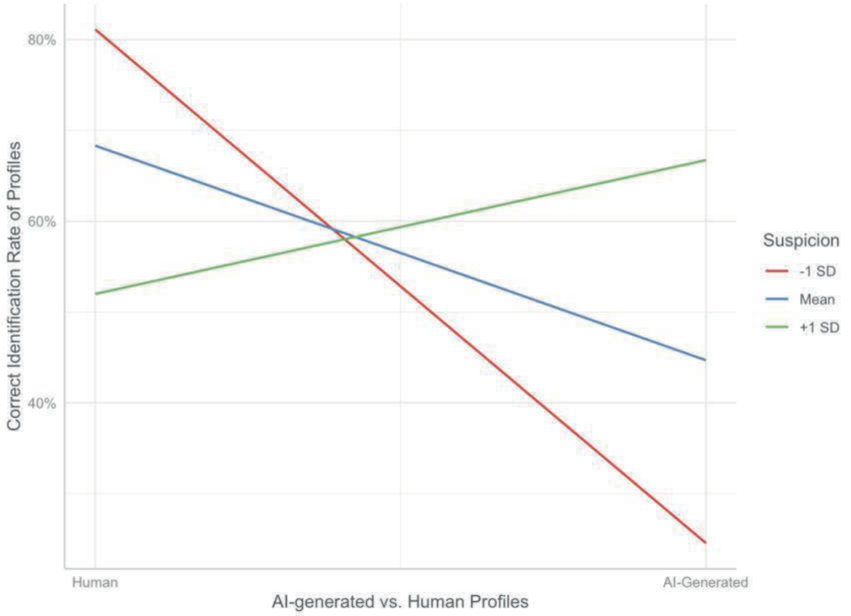


Fig 3. The interaction between profile type and suspicion on the outcome variable (students)

Notes. Mean suspicion means when the level of suspicion is average among subjects. +1 (-1) SD means when suspicion is 1 standard deviation higher (lower) than the mean level of suspicion.

4.2. Experiment 2

4.2.1. Participants

For the second experiment, participants were recruited via Amazon Mechanical Turk. A total of 252 participants (132 male, 118 female, 2 other) completed the task. To incentivize participants to answer correctly, a reward was promised based on performance, although ultimately, all who completed the experiment were given the same reward with a delay. The full table of demographic information is provided in Table 3. As with Experiment 1, in Experiment 2 the participants considered the instructions clear, as 92.9% of them agreed or strongly agreed with the statement "the study's instructions and tasks were clear." Furthermore, regarding the statement "the given tasks were easy to do," the majority of participants (82.9%) agreed or strongly agreed, which is in contrast to the responses given in Experiment 1. This finding could be explained by MTurkers viewing the tasks differently than students, and since MTurkers most likely have significantly more exposure to studies such as this one.

Table 3

Demographics of participants who completed the task

Category	Subcategory	N
Gender	Male Female Other Total	132 118 2 252
Age	Mean Median Range	36.45 34.5 22–70
Ethnicity*	White Asian Middle Eastern Hispanic Black Pacific Islander Other	236 5 0 4 6 1 0
Education	High school Bachelor's Master's Doctoral	13 184 51 4

* Participants can pick multiple

4.2.2. Ability to distinguish and correctly classify profiles

The participants in Experiment 2 were shown the profiles from the same pool of twelve accounts as in Experiment 1. The participants' overall performance in the task was poor, as the average accuracy in classifying the accounts as bots and humans was only 48%, which is very close to random chance, given that the task was binary classification. The least detected bot was labeled correctly as a bot by only 21.7% of the participants that were shown the profile. In comparison, the least correctly classified human profile was labeled as a bot by 59% of the participants. The participants' estimates of the accounts' likelihood of being a bot was, on average 3.6 on a scale of 1-5. When evaluating the suspiciousness of the four components (profile picture, tweet, name, handle), the average scores given were 6.08, 6.42, 6.41, and 6.61, respectively on a scale of 0-10, which indicates that the participants found most profiles moderately suspicious, and which is in line with their tendency to label even the genuine human profiles as bots. Full results by profile are listed in Appendix B.

For statistical analysis, the same analysis strategy was used with Experiment 2 as with the Experiment 1 as was presented in section 4.1.2, with the exception that we included only profiles-level random effects for this analysis, since the variation within the subject group is too low to estimate subject-level random effects. The results are presented in Table 4. The GLMM with profile-level random effects suggests that contrary to the student sample, MTurkers did not show a significant difference in correctly identifying AI-generated or Human profiles, b = -0.08, SE = .22, p = .72. The likelihood-ratio test shows a consistent result. The hypothesized model with the profile fixed effect (pseudo-R2 = .00) is not different than the null model with statistical insignificance, χ^2(1) = .13; p = .72.

Table 4

Generalized Linear Mixed Model–Student Sample

	Model 4	**Model 5**	**Model 6**
(Intercept)	-.00 (.11)	.04 (.16)	$.97^{***}$ (.25)
Profiles FE	-	-.08 (.22)	1.95^{***} (.36)
Suspicion	-	-	$-.16^{***}$ (.03)
Profile FE X Suspicion	-	-	$.30^{***}$ (.04)
Subject RE	*Incl.*	*Incl.*	*Incl.*
Image RE	.00	.00	.06
Pseudo-R^2	-.00 (.11)	.04 (.16)	$.97^{***}$ (.25)

Note. N = 1008; 12 image groups; * p <.05, ** p < .01, *** p < .001; standard errors are in parentheses; Pseudo-R2 of the fixed effect is reported; Incl.: included; RE: Random Effects; FE: Fixed Effects; Profiles FE: 1 if AI-generated profiles or 0 if human profiles.

Despite lacking the main effect, we found a significant interaction effect between profile fixed effects and suspicion on the outcome (Model 6). The pattern of the result is consistent with that in the subject sample. Subjects with high suspicion of AI-generated profiles were more likely to identify the profiles correctly than those with low suspicion. However, higher suspicion led to a lower correct identification rate when human profiles were presented. Figure 4 presents the plot of the interaction result.

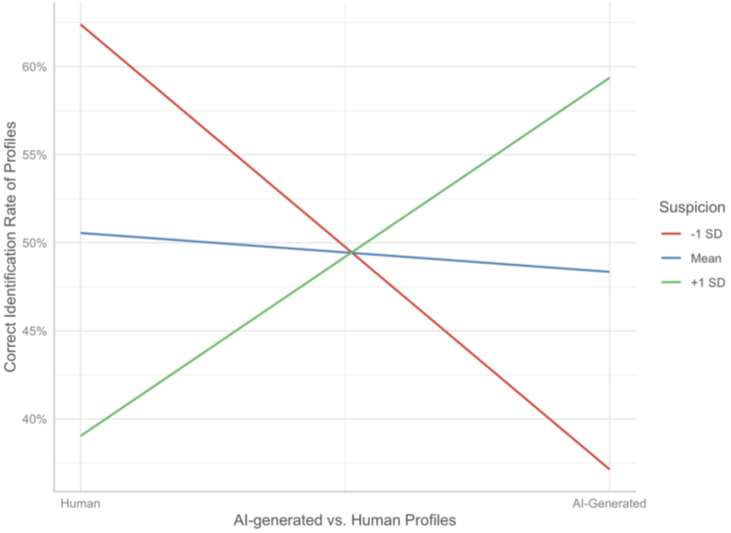


Fig 4. The interaction between profile type and suspicion on the outcome variable (MTurk)

Notes. Mean suspicion means when the level of suspicion is average among subjects. +1 (-1) SD means when suspicion is 1 standard deviation higher (lower) than the mean level of suspicion.

5. Discussion

In this section, we first compare the results of the two experiments and summarize the most important findings. We then review the practical as well as theoretical implications of the findings, and lastly discuss the limitations of the study.

Experiments 1 and 2 show that humans cannot distinguish AI-generated bot profiles from genuine human profiles in a social media feed. In both experiments, the participants frequently labeled the bot accounts as humans and, conversely, surprisingly, often misclassifying the genuine humans as bots. The students (Experiment 1) performed worse than the MTurkers (Experiment 2) in detecting bots specifically, while the MTurkers were more often mislabeling genuine humans as bots. When ordering the bot accounts by how often they were correctly identified, the order was the same in both experiments, indicating that participants in either experiment found the same profiles more challenging to detect. Statistical testing revealed that in Experiment 1 there was a statistically significant lower rate of correctly identifying AI-generated profiles, while this was not the case in Experiment 2, where the results seemed closer to random chance.

Research question 1 was, "Can humans distinguish AI-generated bot profiles from genuine profiles that are commenting on a social media post?" based on the findings, the answer is that they cannot. Research question 2, which was "Which features on a bot profile can humans detect are AI-generated and not genuine?" could not be answered, as the participants were unable to detect the bots consistently, and when assessing the suspiciousness of the different components of the bot profiles (profile picture, tweet, name, handle), the responses were mainly close to the middle values, indicating uncertainty.

5.1 Practical implications

The main practical implication of the study's findings is that creating labeled datasets of bots via qualitative labeling done by humans should not be relied on. As some supervised bot detection models relied on training data labeled via crowdsourcing, in the future, such an approach should not be taken. This implication affects bot research and applied settings, as any analysis that relies on large amounts of social media user data should consider the presence of bots. Especially for academic researchers and industry analysts who lack the competencies to develop or use open-source bot detection models, cleaning datasets of advanced social bots that use generative AI can be much more challenging. This is a significant concern now that bot detection is more complicated than before due to tools such as the Botometer and others that relied on Twitter's (X's) API having been disabled due to the social networking site's changes to the API.

Overall, these findings have two significant implications for the study of social bots. First, while there are still only a few known examples of bots on social media that are using generative AI, there will likely be more in the future, as bot developers adopt new tools, especially when they have the potential to improve the ability of bots to remain undetected [7,9]. Second, the detection of social bots can become more challenging not only due to the aforementioned effects on the human ability to qualitative detect them, but also due to the lack of accurate computational methods for the detection of generated text, even machine learning-based detection of generative AI-powered bots can prove difficult [7]. However, new ways to detect generative AI powered social bots will likely arise. For instance, early examples of bots using ChatGPT to produce Tweets were spotted as the accounts would post all outputs of the chatbot, including the phrase "as an AI language model" [7].

Lastly, as a more benign practical implication, social bots that use generative AI can also be used for non-malicious purposes, where their humanlike nature can be an advantage. As an example, previous studies have proposed using benevolent social bots to deradicalize online communities [24]. Moreover, researchers can develop humanlike social bots more easily given for research purposes thanks to the ease-of-use of tools such as large language models with chatbot interfaces and other generative AI models that can be accessed through a webpage interface.

5.2 Theoretical implications

While the goal of this paper was to provide empirical evidence showing that we have surpassed the point where humans can distinguish fully AI-generated bot profiles from genuine human profiles in social media, the findings can help guide future theoretical work. One of the theories discussed by previous work in relation to social bots and misinformation is the spiral of silence [14,32,51], which proposes that the perceived opinion of the majority influences an individual's willingness to express conflicting opinions on a matter [52]. In the context of social media discourse, this would imply that if most users discussing a topic are voicing that they have a specific opinion, those with opposing opinions are less likely to post contradicting opinions, which further helps increase the perceived popularity of the opinion of the majority. Moreover, by controlling a network of social bots, this perceived opinion of the majority could be manipulated by flooding discussions with posts taking specific stances that the bots support. Previous research has shown through simulation that a relatively small percentage of bot accounts can cause this tipping of the climate in online discourse [32].

As the findings of this study suggest that humans cannot detect AI-generated profiles in a social media feed, social bots manipulating the perceived public opinion on a topic has become a more realistic threat. However, further empirical work, such as experiments, is needed to validate whether the spiral of silence can be strengthened, or even caused by many posts made by generative AI-powered social bots.

5.3 Limitations

The study's primary limitation is the limited nature of the simulated social media feed, as the participants of the experiments could only see a limited amount of information on each profile. In a real social networking site, it would be possible to see more data points per profile by going to view each profile participating in a discussion individually, which would reveal information such as the bio description of the profile as well as all previous posts made by the account, assuming that the account is public. Given that the study's objective was explicitly related to evaluating the ability of an average human who uses social media to identify bots while browsing a social media feed, this limitation should not be reasonable. Furthermore, we justify this with the assumption that most social media users will not thoroughly investigate each profile that they come across on social media and that the profiles and posts seen momentarily in a feed can nevertheless influence the perception of the users of social media.

Due to this constraint in the design of the experiment, we limit the findings to suggesting that humans cannot distinguish bots from genuine accounts when seeing them within discussions on a social media feed. Thus, it is not yet determined whether the classification accuracy would improve when given access to complete profiles and posting history nor do we know if generative AI can be consistent enough over multiple posts to make it seem like there is one human author. This will be discussed more in the conclusion as potential future areas of research.

6. Conclusion

Previous research has thoroughly studied how bots on social media can be detected via computational methods [8] and recent studies have already begun investigating how to detect social bots that are using generative AI such as ChatGPT to produce their posts [7]. Despite the popularity of bot detection research, the human ability to qualitatively detect bots on social media has not been widely investigated. Building on previous studies on the human ability to spot AI-generated content, in this paper, we presented the results of two experiments, which both showed that humans cannot consistently distinguish AI-generated bot profiles from genuine human profiles in a social media feed. These findings concern researchers and society in general, as producing realistic computational propaganda and influencing opinions via social bots has become easier than before due to the availability of easy-to-use tools that can be used to generate realistic and humanlike social media profiles and posts.

However, further research is needed to validate and generalize the study's findings in broader settings and address the limitations of the two experiments. More specifically, first, we propose that future experiments look into the detectability of AI-generated social media profiles when the evaluator is given full access to the profiles of the humans and bots as well as their most recent posts. This way it could be determined if via more thorough examination of profiles and by having access to multiple posts it is still impossible to distinguish between genuine and generated profiles. Second, the theoretical implications regarding the spiral of silence could be further studied. This could be done by for example experimenting with larger feeds consisting of both genuine and bot profiles, and prompting participants to explain what they believe is the majority’s opinion on a topic, especially if told that some profiles might be bots.

Bots have been continuously evolving throughout their history as new technologies have become available and as they have had to adapt to avoid detection and deletion [8,19]. Social bots that use generative AI are the latest iteration in this arms race, and perhaps the most advanced variant of social bots to date. As large language models and generative AI still rapidly developing, future versions of these tools might help build social bots that are more humanlike than currently. While even their less sophisticated social bot predecessors have been claimed to be capable of avoiding detection and thus able to influence opinions in online communities, these concerns seem to have become more pressing due to the findings of this study.

7. References

[1] C.A. de L. Salge, E. Karahanna, J.B. Thatcher, Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter, MIS Quarterly 46 (2022). https://doi.org/DOI: 10.25300/MISQ/2021/15598.

[2] J. Nicas, Why can’t the social networks stop fake accounts?, The New York Times (2020).

[3] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017: pp. 963–972. https://doi.org/10.1145/3041021.3055135.

[4] S. Bond, That smiling LinkedIn profile face might be a computer-generated fake, National Public Radio (2022). https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?t=1654174474533.

[5] J.A. Goldstein, R. DiResta, Research Note: This Salesperson Does Not Exist: How Tactics from Political Influence Operations on Social Media are Deployed for Commercial Lead Generation, Harvard Kennedy School Misinformation Review 3 (2022) 1–15.

[6] G. Da San Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, P. Nakov, A Survey on Computational Propaganda Detection, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2021.

[7] K.-C. Yang, F. Menczer, Anatomy of an AI-powered malicious social botnet, arXiv Preprint arXiv:2307.16336 (2023).

[8] S. Cresci, A Decade of Social Bot Detection, Communications of the ACM 63 (2020) 72–83. https://doi.org/10.1145/3409116.

[9] E. Ferrara, Social bot detection in the age of ChatGPT: Challenges and opportunities, First Monday (2023).

[10] S.J. Nightingale, H. Farid, AI-synthesized faces are indistinguishable from real faces and more trustworthy, Proceedings of the National Academy of Sciences 119 (2022) e2120481119.

[11] S.D. Bray, S.D. Johnson, B. Kleinberg, Testing human ability to detect 'deepfake'images of human faces, Journal of Cybersecurity 9 (2023) tyad011.

[12] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N.A. Smith, All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text, CoRR abs/2107.00061 (2021). https://arxiv.org/abs/2107.00061.

[13] L. Dugan, D. Ippolito, A. Kirubarajan, S. Shi, C. Callison-Burch, Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023: pp. 12763–12771.

[14] S. Rossi, Y. Kwon, O.H. Auglend, R.R. Mukkamala, M. Rossi, J. Thatcher, Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?, in: Proceedings of the 56th Hawaii International Conference on System Sciences, 2023: pp. 134–143. https://hdl.handle.net/10125/102645 (accessed March 10, 2023).

[15] A.P. Saygin, I. Cicekli, V. Akman, Turing test: 50 years later, Minds and Machines 10 (2000) 463–518.

[16] V. Benjamin, T. Raghu, Augmenting social bot detection with crowd-generated labels, Information Systems Research 34 (2023) 487–507.

[17] S. Feng, H. Wan, N. Wang, J. Li, M. Luo, Twibot-20: A comprehensive twitter bot detection benchmark, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021: pp. 4485–4494.

[18] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, B.Y. Zhao, Social turing tests: Crowdsourcing sybil detection, in: Proceedings of The 20th Annual Network & Distributed System Security Symposium (NDSS), The Internet Society, 2013.

[19] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Communications of the ACM 59 (2016) 96–104. https://doi.org/10.1145/2818717.

[20] R. Gorwa, D. Guilbeault, Unpacking the Social Media Bot: A Typology to Guide Research and Policy, Policy and Internet 12 (2020) 225–248. https://doi.org/10.1002/poi3.184.

[21] R.J. Oentaryo, A. Murdopo, P.K. Prasetyo, E.-P. Lim, On Profiling Bots in Social Media, in: E. Spiro, Y.-Y. Ahn (Eds.), Social Informatics, Springer International Publishing, 2016: pp. 92–109.

[22] S. Cresci, R.D. Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Fame for sale: Efficient detection of fake Twitter followers, Decision Support Systems 80 (2015) 56–71. https://doi.org/10.1016/j.dss.2015.09.003.

[23] A. Bessi, E. Ferrara, Social Bots Distort The 2016 U.S. Presidental Election, First Monday 21 (2016).

[24] K.M. Blasiak, M. Risius, S. Matook, "Social Bots for Peace": A Dual-Process Perspective to Counter Online Extremist Messaging, in: ICIS 2021 Proceedings, 2021.

[25] C.A.D.L. Salge, N. Berente, Is that social bot behaving unethically?, Communications of the ACM 60 (2017) 29–31. https://doi.org/10.1145/3126492.

[26] D.M.J. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, M.J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S.A. Sloman, C.R. Sunstein, E.A. Thorson, D.J. Watts, J.L. Zittrain, The science of fake news, Science 359 (2018) 1094–1096. https://doi.org/10.1126/science.aao2998.

[27] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 1151 (2018) 1146–1151.

[28] C. Shao, G.L. Ciampaglia, O. Varol, K.C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nature Communications 9 (2018). https://doi.org/10.1038/s41467-018-06930-7.

[29] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, P. Zola, Coordinated inauthentic behavior and information spreading on Twitter, Decision Support Systems 160 (2022) 113819. https://doi.org/10.1016/j.dss.2022.113819.

[30] J. Marx, F. Brünker, M. Mirbabaie, E. Hochstrate, Conspiracy Machines–The Role of Social Bots during the COVID-19 Infodemic, ACIS 2020 Proceedings (2020).

[31] S. Rossi, The Scamdemic Conspiracy Theory and Twitter's Failure to Moderate COVID-19 Misinformation, in: The 55th Hawaii International Conference on System Sciences: HISS 2022, Hawaii International Conference on System Sciences (HICSS), 2022.

[32] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, S. Stieglitz, Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks, European Journal of Information Systems 28 (2019) 394–412. https://doi.org/10.1080/0960085X.2018.1560920.

[33] A. Alarifi, M. Alsaleh, A. Al-Salman, Twitter turing test: Identifying social machines, Information Sciences 372 (2016) 332–346. https://doi.org/10.1016/j.ins.2016.08.036.

[34] N. Köbis, L.D. Mossink, Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry, Computers in Human Behavior 114 (2021) 106553.

[35] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundagellya, I. Sutskever, Better Language Models and Their Implications, (2021). https://openai.com/blog/better-language-models/.

[36] W. Knight, Scammers Used ChatGPT to Unleash a Crypto Botnet on X, Wired (2023). https://www.wired.com/story/chat-gpt-crypto-botnet-scam/ (accessed December 7, 2023).

[37] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: pp. 4401–4410. https://doi.org/10.1109/tpami.2020.2970919.

[38] N. Yu, L. Davis, M. Fritz, Attributing fake images to GANs: Learning and analyzing GAN fingerprints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: pp. 7556–7566.

[39] B. Strick, Analysis of the Pro-China Propaganda Network Targeting International Narratives, Center for Information Resilience, 2021. https://www.info-res.org/post/revealed-coordinated-attempt-to-push-pro-china-anti-western-narratives-on-social-media (accessed September 5, 2021).

[40] C. Freitas, F. Benevenuto, S. Ghosh, A. Veloso, Reverse engineering socialbot infiltration strategies in twitter, Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015 (2015) 25–32. https://doi.org/10.1145/2808797.2809292.

[41] M. Shafahi, L. Kempers, H. Afsarmanesh, Phishing through social bots on Twitter, Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016 (2016) 3703–3712. https://doi.org/10.1109/BigData.2016.7841038.

[42] J. Roozenbeek, S. van der Linden, Fake news game confers psychological resistance against online misinformation, Palgrave Communications 5 (2019) 1–10. https://doi.org/10.1057/s41599-019-0279-9.

[43] P.L. Moravec, R.K. Minas, A.R. Dennis, Fake news on social media: People believe what they want to believe when it makes no sense at All, MIS Quarterly: Management Information Systems 43 (2019) 1343–1360. https://doi.org/10.25300/MISQ/2019/15505.

[44] C. Flick, Informed consent and the Facebook emotional manipulation study, Research Ethics 12 (2016) 14–28. https://doi.org/10.1177/1747016115599568.

[45] Y. Boshmaf, I. Muslukhov, K. Beznosov, M. Ripeanu, Design and analysis of a social botnet, Computer Networks 57 (2013) 556–578. https://doi.org/10.1016/j.comnet.2012.06.006.

[46] P. Remy, Name Dataset, GitHub Repository (2021). https://github.com/philipperemy/name-dataset.

[47] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, others, Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[48] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Using lme4, Journal of Statistical Software 67 (2015) 1–48.

[49] R Core Team, R: A language and environment for statistical computing, (2023).

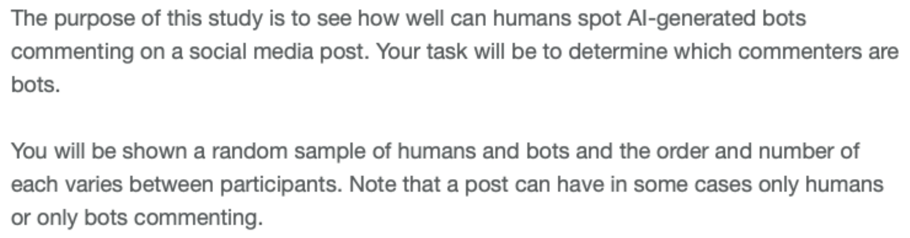
[50] H. Wickham, others, Elegant graphics for data analysis, Media 35 (2009) 10–1007.

[51] A. Shukla, N. Sahasrabudhe, S. Moharir, Opinion Dynamics: Bots and the Spiral of Silence, in: 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), IEEE, 2023: pp. 436–439.

[52] E. Noelle-Neumann, The spiral of silence a theory of public opinion, Journal of Communication 24 (1974) 43–51.

Appendices

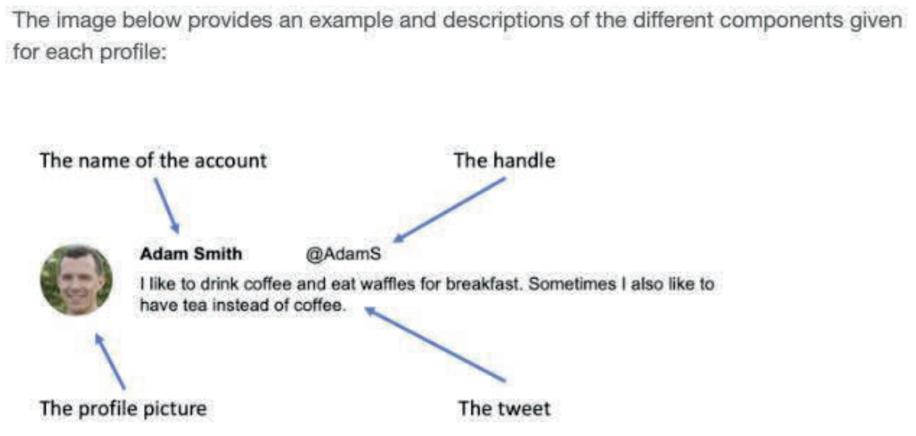
Appendix A: The instructions shown to participants



The purpose of this study is to see how well can humans spot AI-generated bots commenting on a social media post. Your task will be to determine which commenters are bots.

You will be shown a random sample of humans and bots and the order and number of each varies between participants. Note that a post can have in some cases only humans or only bots commenting.

Fig A1. The initial instructions shown to participants



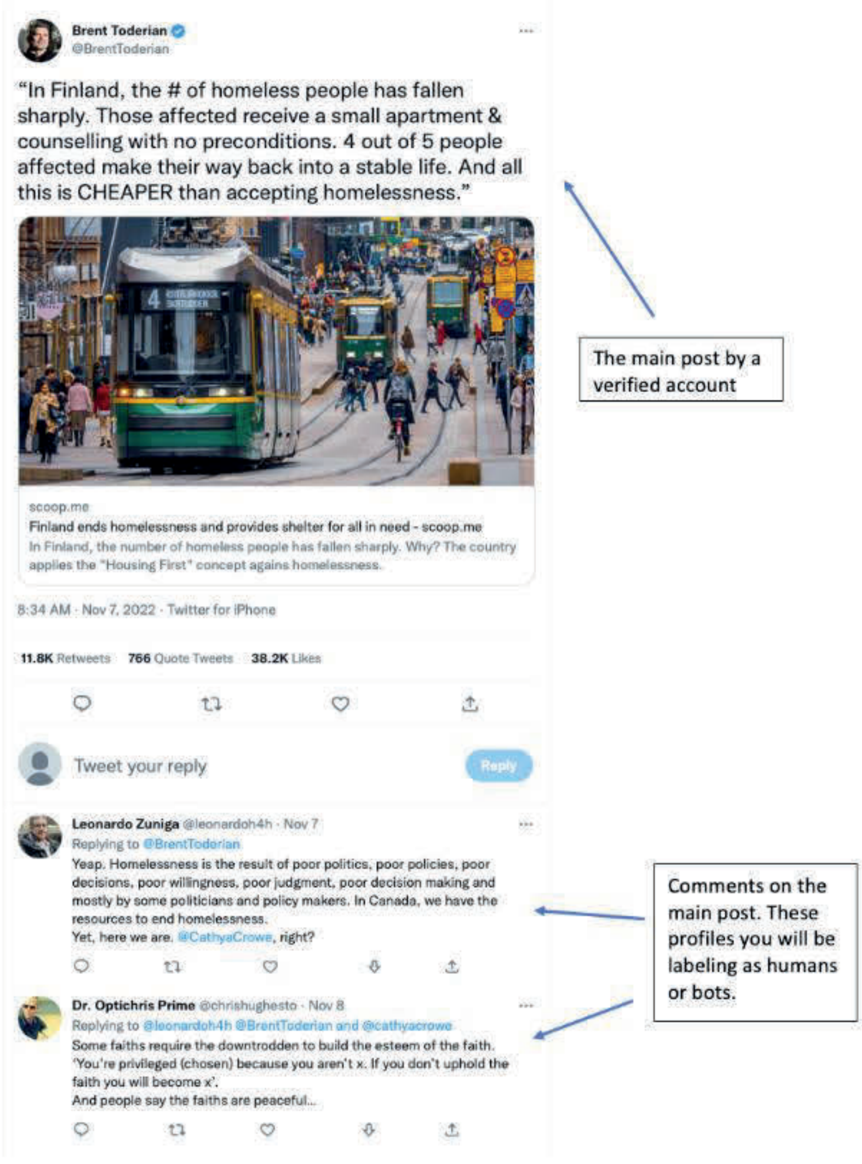
The image below provides an example and descriptions of the different components given for each profile:

I like to drink coffee and eat waffles for breakfast. Sometimes I also like to have tea instead of coffee.

Fig A2. The terminology explainer shown to participants

The post is always from a verified (real) profile, while there are two profiles commenting on the post. The accounts that are commenting on the post can be either humans or bots. Note that a post can have in some cases only humans or only bots commenting. Moreover, you should assess the profiles purely based on the name, handle, picture and post.

Below is an example of what you could see during the experiment:



Brent Toderian
@BrentToderian

"In Finland, the # of homeless people has fallen sharply. Those affected receive a small apartment & counselling with no preconditions. 4 out of 5 people affected make their way back into a stable life. And all this is CHEAPER than accepting homelessness."

scoop.me
Finland ends homelessness and provides shelter for all in need - scoop.me
In Finland, the number of homeless people has fallen sharply. Why? The country applies the "Housing First" concept agains homelessness.

8:34 AM · Nov 7, 2022 · Twitter for iPhone

11.8K Retweets 766 Quote Tweets 38.2K Likes

Tweet your reply

Reply

Leonardo Zuniga @leonardoh4h · Nov 7
Replying to @BrentToderian
Yeap. Homelessness is the result of poor politics, poor policies, poor decisions, poor willingness, poor judgment, poor decision making and mostly by some politicians and policy makers. In Canada, we have the resources to end homelessness.
Yet, here we are. @CathyaCrowe, right?

Dr. Optichris Prime @chrishughesto · Nov 8
Replying to @leonardoh4h @BrentToderian and @cathyacrowe
Some faiths require the downtrodden to build the esteem of the faith. 'You're privileged (chosen) because you aren't x. If you don't uphold the faith you will become x'.
And people say the faiths are peaceful...

The main post by a verified account

Comments on the main post. These profiles you will be labeling as humans or bots.

In the survey you will be asked to mark each account that you believe is a bot.

Fig A3. The feed demonstration shown to participants

Appendix B: Classification results by account

Results of Experiment 1 (Students)

Account	Average bot score	Accuracy	Likelihood of being a bot	Profile picture	Tweet	Name	Handle
Bot 1	0,063	0,063	3,076	4,278	4,190	3,291	3,646
Bot 2	0,303	0,303	2,605	3,053	3,421	2,395	2,684
Bot 3	0,338	0,338	2,701	3,182	3,662	3,013	3,078
Bot 4	0,434	0,434	2,921	2,947	4,421	3,105	3,289
Bot 5	0,628	0,628	3,474	5,577	5,269	3,962	4,615
Bot 6	0,763	0,763	3,789	5,579	6,132	4,079	4,974
Human 1	0,145	0,855	2,447	2,526	3,066	2,921	3,618
Human 2	0,250	0,750	2,408	1,553	3,632	1,974	2,329
Human 3	0,282	0,718	2,603	2,513	3,731	3,718	3,551
Human 4	0,291	0,709	2,797	3,506	4,000	3,114	3,949
Human 5	0,351	0,649	2,896	3,351	4,481	3,494	3,416
Human 6	0,539	0,461	3,276	2,737	3,895	3,645	6,250
Mean	0,370	0,560	2,920	3,400	4,160	3,230	3,780

Results of Experiment 2 (MTurk)

Account	Average bot score	Accuracy	Likelihood of being a bot	Profile picture	Tweet	Name	Handle
Bot 1	0,217	0,217	3,747	5,928	6,458	6,313	6,723
Bot 2	0,369	0,369	3,643	5,964	6,250	6,417	6,524
Bot 3	0,388	0,388	3,600	6,212	6,529	6,753	6,718
Bot 4	0,536	0,536	3,702	6,190	6,548	6,321	6,512
Bot 5	0,537	0,537	3,732	6,549	6,573	6,354	6,793
Bot 6	0,694	0,694	3,753	6,329	6,847	6,659	6,694
Human 3	0,390	0,610	3,585	5,780	6,378	6,256	6,573
Human 2	0,435	0,565	3,529	6,106	6,435	6,353	6,588
Human 1	0,459	0,541	3,671	6,247	6,647	6,776	7,000
Human 6	0,529	0,471	3,482	5,965	6,318	6,376	6,600
Human 4	0,541	0,459	3,565	5,706	6,012	6,165	6,412
Human 5	0,590	0,410	3,542	5,940	6,072	6,120	6,241
Mean	0,470	0,480	3,630	6,080	6,420	6,410	6,610

Paper V

This paper was published as a peer-reviewed editorial at the *International Journal of Information Management*.

Augmenting Research Methods with Foundation Models and Generative AI

Sippo Rossi
Copenhagen Business School

Matti Rossi
Aalto University

Raghava Rao Mukkamala
Copenhagen Business School

Jason Thatcher
Temple University

Yogesh K Dwivedi
Swansea University

Abstract

Deep learning (DL) research has made remarkable progress in recent years. Natural language processing and image generation have made the leap from computer science journals to open-source communities and commercial services. Pre-trained DL models built on massive datasets, also known as foundation models, such as the GPT-3 and BERT, have led the way in democratizing artificial intelligence (AI). However, their potential use as research tools has been overshadowed by fears of how this technology can be misused. Some have argued that AI threatens scholarship, suggesting they should not replace human collaborators. Others have argued that AI creates opportunities, suggesting that AI-human collaborations could speed up research. Taking a constructive stance, this editorial outlines ways to use foundation models to advance science. We argue that DL tools can be used to create realistic experiments and make specific types of quantitative studies feasible or safer with synthetic rather than real data. All in all, we posit that the use of generative AI and foundation models as a tool in information systems research is in very early stages. Still, if we proceed cautiously and develop clear guidelines for using foundation models and generative AI, their benefits for science and scholarship far outweigh their risks.

Keywords: Foundation model, Generative AI, Experiments, Synthetic data

1. Introduction

Generative AI has made its way into the public consciousness through applications and services based on foundation models. Foundation models refer to “any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks'' (Bommasani et al., 2022, p. 3). In other words, foundation models are pre-trained deep learning models that have been trained with extremely large datasets and which can be used directly or after further finetuning, for a broad variety of tasks such as text and image generation or classification. In late 2022, many scientists first encountered foundation models directly after universities began discussing how to respond to students using ChatGPT to complete coursework (Huang, 2023). Even before ChatGPT, some had stumbled upon generative AI through websites like thisxdoesnotexist.com, by hearing from acquaintances who received early access to OpenAI's GPT-3 (Brown et al., 2020) and DALL-E (Ramesh et al., 2021), or by reading about Meta's troubled and short-lived release of its large language model Galactica (Heaven, 2022), which would happily generate research papers about such non-existent things like Russian space bears.

While the focus on the negative impact of generative AI on foundation models in science and society is understandable, they also offer opportunities to enhance and improve existing research methods. For example, advances in natural language processing (NLP) have received much attention because of their practical applications (Brown et al., 2020), but it is worth noting that the advances in image generation have also been significant. It is now possible to generate realistic images of people that are indistinguishable from real photographs (Nightingale & Farid, 2022), and we are not far from being able to produce videos of limited length and quality (Esser et al., 2023). Similarly, large language models using the latest NLP methods can generate high-quality synthetic content for experiments, freeing researchers to focus on value-added activities such as interpreting the results. Clearly, generative AI has potential applications for researchers as well as implications for both authors and reviewers of the studies that use it.

Given the sudden surge in interest in generative AI and particularly large language models such as ChatGPT, their implications for research have already been discussed by several editorials (Dwivedi et al., 2023a; Susarla et al., 2023), but with an emphasis on the challenges and policy rather than explicit methodological commentary and guidance. In this editorial, we approach these new technologies as an opportunity for research and suggest use cases for foundation models and generative AI, as well as propose best practices for incorporating them into existing research methods. We do this by outlining how NLP and image generation models can be used to conduct research in novel ways. Our goal is to provide straightforward suggestions and guidelines on using foundation models, particularly in building realistic experiments and in different types of quantitative studies, such as those relying on supervised machine learning, where generated synthetic data can be used as a safe alternative to otherwise sensitive human data or where collecting sufficient enough data is prohibitively expensive or otherwise difficult. While many of our examples are based on specific well-known models, our proposed approaches are model agnostic and thus applicable to any appropriate foundation model.

The editorial is organized as follows. We begin with a brief overview of what has led to the current state of the art, discuss the limitations of existing technologies, and provide examples of early studies that have successfully applied foundation models. We then present how foundation models can be used in different scenarios and provide examples from our own experience on how they can be used. Finally, we conclude with a discussion of the implications of using foundation models and ethical considerations.

2. State of the art

The algorithms behind foundation models are not new, as they are based on deep neural networks, which have been around for decades (Bommasani et al., 2022). Deep learning has been practically implementable by individual researchers since around 2015-2016 when TensorFlow, Keras, and PyTorch were released, and the hardware needed to train the models became available even to the general public (Dean, 2022). Despite this, the competitive advantage of foundation models lies in the massive amount of model parameters and having been trained on colossal datasets using vast computational resources that are inaccessible to all but the wealthiest organizations (Bommasani et al., 2022). This pre-trained nature means that the end users can forgo expensive and time-consuming model training and focus on either applying the model or using transfer learning (Zhuang et al., 2020) to fine-tune and apply it to a specific task.

It should be noted that the exact definition of a foundation model is still slightly fuzzy. Besides Bommasani et al's., (2022) definition, which was given in the introduction, others have described foundation models, for example, as “flexible, reusable AI models that can be applied to just about any domain or industry” (Murphy, 2022) or as “AI neural networks trained on massive unlabeled datasets to handle a wide variety of jobs from translating text to analyzing medical images” (Merritt, 2023). Thus, the consensus is that they are complex deep learning models that have been trained using unsupervised or self-supervised learning, meaning that their training does not contain labels, allowing the use of much broader and larger datasets.

2.1. What can be done with foundation models?

A wide variety of speech and image generation models have become available through open or closed-source software. Large technology companies, specialized research labs, and startups have competed to develop and publish basic models that are, to varying degrees, publicly available as downloadable parameters to a model or accessible via chatbot interfaces or application programming interfaces (APIs). Some notable examples include OpenAI and its GPT-3 (Brown et al., 2020) and DALL-E (Ramesh et al., 2021), Stability AI’s Stable Diffusion (Rombach et al., 2022), Google’s BERT (Devlin et al., 2019) and LaMDA (Collins & Ghahramani, 2021) and Meta AI’s LLaMA (Touvron et al., 2023). While there are other areas where the same or completely different foundational models can be applied, such as image and text classification, they are omitted as we focus on generative methods.

Several studies have shown that large language models can produce text that is difficult or even impossible for a human to distinguish from human-written text (Clark et al., 2021; Köbis & Mossink, 2021). They can produce grammatically correct text, summarize articles, answer questions that are more sophisticated than decision tree-based back-end solutions for chatbots, or even rewrite existing text from one style to another (Bommasani et al., 2022; Brown et al., 2020; Floridi & Chiriatti, 2020).

Foundation models for image synthesis have also achieved impressive results. They can generate competition-winning pieces of art (Parshall, 2023) and strikingly realistic-looking photographs that are difficult for humans to recognize when shown alongside real photographs (Nightingale & Farid, 2022; Rossi et al., 2023). These models can respond to prompts given in text or image format, producing infinite variations of an image or generating entirely new ones (Bommasani et al., 2022; Karras et al., 2019; Rombach et al., 2022). The output can range from seemingly photorealistic to creative and computer-generated.

2.2. What can't (yet) be done with foundation models

The achievements of foundation models are remarkable, and particularly large language models have been able to create an illusion of an impressive performance (Cosmo, 2022); however, they are still examples of weak or narrow AI (Fei et al., 2022), containing historical and human-like biases from their training corpus (Dwivedi et al., 2023a; Stahl & Eke, 2024; Susarla et al., 2023). Therefore, the belief that large language models are sentient based on their ability to converse in a human-like manner is misguided and dangerous (Cosmo, 2022), and it is important to understand the limitations of foundation models, even in basic tasks such as text summarization.

Large Language Models (LLMs) are ultimately stochastic parrots, predicting which word fits in a sentence and repeating patterns (and biases) from the data they have been trained on without any judgment about what is right or wrong (Bender et al., 2021; Schramowski et al., 2022). In other words, based on linguistic and general knowledge learned from the training sets, they can produce grammatically correct but factually incorrect text with high confidence and even make up references or quote people who do not match the context or exist. This problem has been documented in all LLMs and the term hallucination refers to this phenomenon (Ji et al., 2023). Thus, LLMs cannot be relied upon to produce coherent or factually accurate text, and a human with adequate subject matter expertise should always review the generated text.

Image generation models are primarily based on generative adversarial networks (GAN), which can produce realistic images or videos, but occasionally produce flawed or even disruptive images due to problems with the underlying training data and the probabilistic nature of the generators. These errors or "artifacts" in images often occur in objects that should be symmetrical, such as teeth, glasses, or earrings (Karras et al., 2020; West and Bergstrom, 2019). Because these symmetrical objects tend to be rendered incorrectly, they can reveal that the photo is generated rather than authentic (West and Bergstrom, 2019). Furthermore, the generated images often repeat negative stereotypes due to the biases and lack of representation in the most common image datasets used to train models like DALL-E, StyleGAN, or Stable Diffusion. In practice, the models perform better in terms of image quality when used to generate images of white males rather than, for example, non-white females (Barr, 2022).

2.3. Foundation models as tools in research

Before the 2020s, there was much less emphasis on building APIs or services that give access to foundation models or generative AI to individuals outside of a narrow group of scientists with advanced skills in data science. Consequently, there is a dearth of published papers outside computer science journals where generative AI and foundation models have been integrated into the research design, although the number of publications is rapidly rising, and we are soon likely to see more examples across different fields (Kar et al., 2023). One example of an early paper that integrated foundation models into the research design is our study on how well humans can detect bots on social media, which used both NLP (GPT-3) and image generation methods (StyleGAN) to produce social media profiles with generated posts and profile pictures (Rossi et al., 2023). Other examples of foundation models being used include using StyleGAN to generate variations of faces in a study proposing a model for predicting matching in online dating (Kwon et al., 2022) and building a model for automatic procedure generation using BART (Geluykens et al., 2021). It should be noted that in each of these papers, the foundation models are not the focus of the research but are part of the method. Later in the editorial, we will use these studies and others to illustrate key points and opportunities.

3. Incorporating foundation models into existing research methods

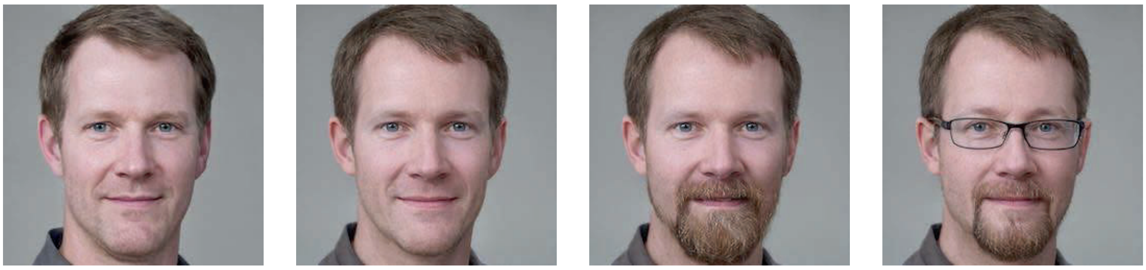
Existing articles and editorials on Generative AI primarily provide general discussions on its potential use or misuse (e.g., Dwivedi et al., 2023a; Dwivedi et al., 2023b; Susarla et al., 2023), as well as its feasibility as a tool for conducting literature reviews (Pan et al., 2023). Due to the nascent nature of foundation models, at the time of writing, there are few examples of their use as part of the method in information systems research. Still, examples can be found in related fields such as computer science. In this section, we outline ways in which generative AI can be applied to IS research, drawing inspiration from previous work as well as proposing entirely new approaches based on the capabilities of foundation models.

To find relevant literature, we searched for recent publications containing specific keywords within the AIS eLibrary, Scopus, Google Scholar and arXiv. The keywords used were "foundation model," "generative AI" and "generative artificial intelligence." Moreover, we used a snowballing approach to find further relevant publications by reviewing what the papers identified by the keyword search cited. Due to the novelty of the topic and methods, many identified papers were still arXiv pre-prints and thus not peer reviewed, although most of the pre-prints included in the references of this editorial have been heavily cited already at the time of writing.

3.1 Realistic experiments with generated content

Experiments are a popular method for studying various phenomena in human behavior and decision-making, but experimental research is time-consuming and somewhat difficult to conduct. Research methods can be evaluated by their generalizability, realism (for participants), and precision (in controlling and measuring variables) (Dennis & Valacich, 2001; McGrath, 1981). Addressing all three dimensions is difficult, but we argue that foundation models and generative AI can help achieve realism while improving the controllability of variables.

Depending on the experiment, it may contain elements such as text or images, as well as entire simulated environments (a social media feed, a web page, or posts in an online forum) that are shown to subjects. These elements must be collected from the real world or created by the researchers, but this can be challenging. Consider a situation in which a study wants to examine racism in hiring processes based on variations in the applicant's appearance in a resume photo. Rather than trying to collect a diverse set of photos with differences that cannot be controlled due to natural variations in appearance, image generation tools such as StyleGAN and Stable Diffusion make it possible to create entirely new images with multiple similar versions or slight variations of an existing image. For example, suppose the goal is to generate variations of a particular face. In that case, researchers can specify the image generator using different descriptions of facial features such as age, skin tone, or depth of smile. We illustrate this in Figures 1 and 2.

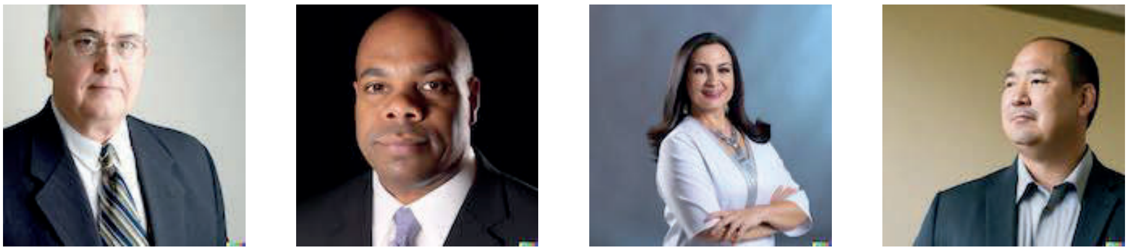


Models: StyleGAN for the leftmost image and StyleCLIP for other images
Original image description and target image instructions used with StyleCLIP:

1) A face; a young face
2) A face; a face with a beard
3) A face; a face with a beard and glasses

Images in Figure 1 generated by StyleGAN & StyleCLIP might inadvertently resemble real people. Any such similarity is purely coincidental and unintentional.

Figure 1: Generating variations with StyleGAN & StyleCLIP (Source: Created by Authors)



Model: DALL-E
Prompts used:

1) A photo of a middle-aged researcher from a business school
2) A photo of a middle-aged researcher from a business school from different backgrounds
3) A photo of a middle-aged Hispanic researcher from a business school
4) A photo of a middle-aged Asian researcher from a business school

Images included in Figure 2 generated by DALL-E might inadvertently resemble real people. Any such similarity is purely coincidental and unintentional.

Figure 2: Examples of images generated with prompts (Source: Created by Authors)

Like taking photographs, manually writing text to populate fields, and creating textual content in experimental environments is laborious. Finding and collecting text from a real-world environment and reusing it in an experiment is not always feasible. Foundation models such as GPT-3 or Llama 2 can be used to generate grammatically correct text for experiments with relative ease. For example, it can partially or fully populate a fake social media feed where the posts are indistinguishable from human-generated content, as we did in one of our recent studies (Rossi et al., 2023). Another potential application of LLMs in experiments is to generate variations of text, allowing for expansion from a small sample generated by humans. An example is in Table 1, where an initial sample of texts used in an experiment is enlarged.

Table 1: Producing new social media posts from a small sample produced by humans (Source: Created by Authors Using GPT-3)

Original text	Generated variation 1	Generated variation 2
This weekend take a moment to pause and reflect on the wonder of life. Spend time with your family and see your friends.	This weekend take a moment to unwind and relax. Recharge your batteries by catching up with your nearest and dearest. Enjoy a picnic in the park or a takeaway with friends.	Gather your loved ones and make some memories this weekend — take a moment in your hectic life to reflect on the beauty of life and all that it has to offer.
USA is a scientific hub for progressing quantum technology and computing. With this rapidly developing technology at our fingertips, potential beneficiaries must consider their quantum-computing strategies.	Germany is an innovation hub for advancing artificial intelligence. With AI technology gaining traction, those interested in exploiting the potential rewards must contemplate their AI plans.	Australia is a driving force in the research and development of autonomous vehicle technology. With this cutting-edge technology available, those wanting to capitalize on it must think through their self-driving car strategies.

Model: GPT-3
Parameters: mode: complete, model: text-davinci-003, temperature: 1.
Prompts used:

1) "Create two similar posts as the following: {*original text*}"
2) "Create two similar posts as the following changing the topic, country and structure: {*original text*}"

Note that the footnotes under Table 1, Figure 1 and Figure 2 contain the key information on the models and parameters used to generate the text and images. We will return to this approach to documenting the use of foundation models later in this editorial.

3.2 Using foundation models to produce synthetic data

Many research topics require the collection and/or storage of data on human subjects, and experiments may expose subjects to content such as images or text written by or associated with an individual. Collecting such data can be expensive, time-consuming, or otherwise impractical. Moreover, handling human data adds a layer of risk to the research and additional workload, as proper storage and sharing of the data must be considered to comply with regulations. This issue is particularly important when research involves sensitive topics or personally identifiable data.

One way to reduce these risks is to augment or replace human data with synthetic data generated by generative AI models. This approach can address ethical and regulatory concerns, making conducting research and storing data easier without risking individual privacy.

The goal of generating synthetic data is to produce data that closely resembles real data in terms of data points and distribution. Ideally, this allows researchers to build the same machine learning models, but without the risk of compromising individual privacy by having a dataset with real human data. Furthermore, even if the original data is not sensitive, in situations where there are not enough data points for the task, such as training a machine learning model, generative AI can be used to generate additional training data to augment the existing dataset (Chung et al., 2023; Frid-Adar et al., 2018), with an example being generated hate speech posts in social media that can be used to train detection models (Kirk et al., 2021; Wullach et al., 2021). Generative AI has already been used to generate synthetic research data in multiple domains, including healthcare, finance, and social media. For example, in healthcare, generative AI has been used to generate large datasets of synthetic medical images (Chambon et al., 2022; Zhou et al., 2021) and patient data (Guillaudeux et al., 2023) from the samples of smaller real-world datasets, which can be used to train machine learning models for disease diagnosis or treatment planning. In finance, generative AI has been used to generate synthetic financial data that can be used to test investment strategies or risk models (Eckerli & Osterrieder, 2021).

It should be noted that there are limitations to this approach. For instance, when using generative AI to increase the amount of training data for a machine learning model, these new data points will be similar to the data that they are generated from. Consequently, the machine learning model trained on the augmented dataset will become better at predicting patterns found in the training data but will not be able to generalize better on novel cases. As a concrete example in the context of bot detection, let's imagine the goal is to apply machine learning to detect bots based on a limited dataset. There are too few datapoints for the model to be accurate without augmenting, so generative AI is used to enlarge the training data and produce similar but still unique data points. The model can detect bots with sufficient training data, but only ones similar to those in its training data. Thus, we emphasize that synthetic data can be used when dealing with a phenomenon that is well understood and where performance is measurable, such as image classification or text classification. Furthermore, it cannot be used to predict novel patterns or scenarios not found in the training data.

In addition to being used as data in various models, synthetic data can also be used to replace or augment the content shown to human subjects in experiments. For example, in a study where subjects are shown posts from a social media feed, taking content from a real social networking site would be easy. Still, contacting all the authors of potentially anonymous or pseudonymous posts is impractical and asking for their permission is impractical. In one of the author's institutions, the ethics review board specifically had such concerns and requested whether a study could be conducted without real data that is traceable to individuals on the internet unless explicit permission is given. The author considered the alternatives and the feasibility of manually creating the content, but this would be extremely time-consuming. Moreover, it would be challenging to emulate social media content produced by multiple authors. Ultimately, it was realized that data could be anonymized by using foundation models to generate slight variations of genuine content that would be close to the original data.

Thus, synthetic data that is derived from genuine data can be used to provide realistic content for an experiment while preserving privacy. Using the example from Table 1, where multiple variations of posts were created, a similar approach could generate synthetic data by removing the original data from the dataset after using the generative AI model. It is important to recognize that this approach comes with caveats. First, careful assessment is needed to determine whether the fact that the content is synthetic can influence the results, and if yes, then using generative AI should be avoided. As an example, consider a study where participants are shown CVs belonging to individuals of different ethnicities and asked to rate them to see if ethnicity influences the results. If some of the CVs contain parts produced with generative AI, while others do not, this can result in a situation where participants' ratings are influenced by whether the content is generated or not, which is undesirable. However, if, instead, in all CVs shown to participants, the same parts are produced by generative AI (such as the photo of the applicant), this is no longer a problem. As a second consideration, the quality of the generated content must be up to the level of genuine content and preferably entirely or nearly indistinguishable to not influence the results. This could be validated, for example, with a limited pretest run through the MTurk or other crowdsourcing platforms.

As a final benefit of using synthetic data, we note that regulations, particularly in the EU, are becoming more stringent regarding the storage and use of data that can be used to identify individual people. The GDPR rules (General Data Protection Regulation, 2016), and the abrupt changes they have brought to journal publishers' policies, have made it more difficult to use data collected through previous studies or web scraping. The rules are intended to protect individuals and their data from misuse, but they also make many types of research projects very difficult or impossible to conduct. Synthetic data is compliant with GDPR, as it is not tied to a real person or identifiable. Therefore, using synthetic data may be preferable to avoid the risks associated with storing data associated with specific individuals.

All in all, generative AI models and synthetic data offer several advantages for research applications that need to avoid risks related to human subjects, privacy, and GDPR compliance. However, it is important to note that synthetic data is not a perfect substitute for real data, and researchers should carefully consider the limitations and assumptions of the generative AI models used to generate synthetic data. In particular, it is worth noting that generative models suffer from learning biases from the training sets (Schramowski et al., 2022), which should be considered before using them to generate synthetic data. Caution is required as research outcomes with synthetic data might differ from those obtained with the original data. Further evidence-based empirical research is needed to evaluate the reliability of synthetic data in different contexts, although initial results seem promising when synthetic data is used with consideration to the method and end goals. For example, we know synthetic data is useful in settings where there are clearly defined parameters and outcome variables, and where the goal is to augment data have a larger training data set (Trabucco et al., 2023). However, we will not know how to use synthetic data to emulate unstructured natural settings until further research completed.

3.3. Summary

We summarize the proposed approaches for augmenting existing research methods with foundation models in Table 2. The approaches are presented in a 2x2 matrix, where on the vertical axis is the task (generating content for experiments or generating synthetic data) and on the horizontal axis is the type of data to be generated (text or images).

Table 2: Ways to augment research with foundation models

Task/Content	Text	Images
Generating content for experiments	Generate realistic-looking text for an experiment, e.g., fake tweets. Example: Rossi et al., 2023	Generate variations of an image, e.g., fake profile pictures. Example: Boyd et al., 2023
Generating synthetic data	Replace genuine text with generated text to mask the original data, e.g., visual storytelling, table-to-text generation, knowledge bases-to-text generation, and so on. Examples: Chung et al. 2023; Chen et al. 2021; Kirk et al.; 2021; Wullach et al., 2021	Use generated images instead of real pictures, e.g., domain-specific images for training a model, text-to-image generation. Examples: Chambon et al., 2022; Trabucco et al., 2023(Chambon et al., 2022; Trabucco et al., 2023)

In this editorial we propose two use cases for foundation models and generative AI as research tools. These use cases are generating content for experiments and generating synthetic data, which are seen in the leftmost column of Table 2. For both tasks, we propose how generative AI and foundation models can be used for generating text and images, and for each of these four possible combinations of task and content, we provide an example of a specific use case for the models.

For more detailed examples and further reading, we recommend looking at the papers provided as examples for each task and content pair. For instance, in the Rossi et al. (2023) paper, there is a description of how GPT-3 was used to produce realistic tweets for a social media study. As the generation of synthetic data is a broader topic, and the usage is very rapidly evolving in several research fields, there are more examples of papers available. We assume that there will be a literal explosion of works employing these tools in the near future.

Lastly, when using foundation models and generative AI, it is important to consider biases that the models contain and biases that using synthetic data can introduce. This applies to all of the use cases discussed in this editorial. More specifically, as foundation models are trained with such large datasets that controlling for biases is practically impossible, the models also will output text or images that represents the biases found across texts written by humans as well as images created by humans (Bender et al., 2021; Schramowski et al., 2022). Furthermore, even when ignoring bias in terms of equality and diversity, using synthetic data can introduce another form of bias, which is that the data may not contain similar variance as genuine data. If this data is then used to for example train a machine learning model, it can result in a model performing extremely well, such as overfitting, but only usable in a very narrow setting. Thus, consideration must be used in the generation process, and afterward, it should be validated that the characteristics of the generated synthetic data are not too distinct from genuine data, using various validation techniques, such as evaluation metrics and human-in-the-loop tests (Chen et al. 2021).

4. Discussion

Based on the examples of how foundation models have been used in recent research, we believe that generative AI and foundation models will become more prominent in future research. As we have shown with text and images, generative AI has the potential to free up the time and mental bandwidth of researchers working with experiments to do value-added work. Specifically, the creation of realistic, immersive environments for experiments will be much faster with the help of foundation models, allowing the researcher to spend more time designing, conducting, and analyzing the results of the experiment rather than painstakingly building an ecologically valid experiment itself.

In addition, using synthetic data produced by generative AI can reduce the amount of sensitive data that needs to be collected and stored, making it easier to comply with the tightening regulatory landscape and share data. Most importantly, it reduces the risk of personal data being exposed through data leaks and eliminates the privacy risks associated with, for example, training models on large datasets consisting of data points that are directly related or connectable to individual people. Another exciting opportunity with foundational models is synthetic data generation using multimodal data. In the future, foundational models will be very useful for combining different multimodal datasets such as images, video, text, and speech to generate synthetic datasets. A concrete example could be the generation of synthetic datasets for hate speech identification on social media platforms based on tweet texts combined with images and videos.

Next, we discuss in more practical terms how to use foundation models, providing a list of some known models at the time of writing and how they can be used. Finally, this section concludes with an overview of the ethical considerations involved in using foundation models and generative AI in research.

4.1 Using foundation models

The first step in using foundation models is to determine what they are intended to do, as different models have different capabilities and limitations. For example, in terms of image generation, StyleGAN can generate extremely realistic photographs, while prompt-based models like DALL-E and Stable Diffusion can generate images based on a textual description with few restrictions on what can be included in the images. Second, there can be significant differences in ease of use as well as the computational resources required. While there have been steady improvements in making foundation models accessible to a wider range of users, many still require some basic knowledge of a high-level programming language such as Python to be used effectively (e.g., to use the APIs). Even models like GPT-4, DALL-E, and Stable Diffusion, which are accessible via websites, are ultimately more efficient and scalable when used via a Python script to consume their APIs. In addition, models that are intended to be used on local computer hardware rather than through a cloud service require at least a high-end computer.

Table 3: Examples of foundation models and how to use them

Task	Model	What it can do	How it is used
Text generation	GPT-3 and GPT-4	Can, among other tasks, generate texts or modify given text, both based on a given prompt and parameters	Used through a web interface or via an API
	LLAMA 2	Generates texts based on a given prompt	Installed locally and used through a high-level programming language such as Python
	Bard	Generates texts based on a given prompt	Used through a web interface at the time of writing with an API being released in the future
	BERT	An open-source language model that can be fine-tuned for various domain-specific supervised/unsupervised downstream tasks	Installed locally and used through a high-level programming language such as Python
	PaLM	Generates and modifies texts. Capable of generating explanations that require multi-step logical inference, world knowledge, and language understanding (Chowdhery et al., 2022)	Used through an API
Image generation	DALL-E	Generates images from textual descriptions or produces variations of a given image	Used through a web interface or via an API
	StyleGAN	Can generate very realistic new images or variations of a given image based on parameters	Installed locally and used through a high-level programming language such as Python
	Stable Diffusion	Generates images from text prompts or produces variations of a given image	Used through a web interface or via an API

Table 3 provides a high-level overview of some of the most popular foundation models for text and image generation at the time of this writing, with notes on their capabilities and whether the model can be accessed through a web page or requires writing code to use. It should be noted that the table is far from exhaustive, and due to the rapid developments in the discipline and productization of foundation models, especially the "how to use" column may change as more user-friendly interfaces are developed. The foundation models that are accessible through a web page or API typically have a pay-by-use scheme, where the cost of use is based on the length of text processed or the number of images produced. Some notable exceptions are StyleGAN, BERT, and Llama 2, which are free and have a large open-source community that develops and shares versions of the models that have already been fine-tuned for more narrowly defined tasks.

We would like to caution that while there is excitement around the capabilities of new models and it can be tempting to use the latest possible alternatives, it is always not preferable, and an older model can be equally good or at least a safer option for the given task. This is due to older and more established models having had time to have been vetted more thoroughly by the userbase, and their limitations especially regarding bias can be more well understood, and consequently the bias in the outputs might also be easier to detect and mitigate thanks to already available literature. Thus, opting to use a newer model should be based on the needs of the task at hand rather than on novelty.

As a concrete example to illustrate the point made in the previous paragraph, we have previously demonstrated that generative AI can be used to enlarge the training datasets used by image classification machine learning model by generating new variations. But if for the given task sufficient improvements in accuracy can be achieved with simpler and more conservative approaches to data augmentation such as random mirroring and cropping (Shorten & Khoshgoftaar, 2019), then there is no need to apply generative AI. Thus, generative AI and foundation models should be employed only when simpler tools fail and not merely for the sake of employing the latest available tools.

4.2 Ethical considerations

Foundation models, like other AI tools, have several important ethical considerations that need to be considered and addressed. First and foremost, we would like to emphasize that due to the possibility of generative AI suddenly producing artifacts or unsafe content due to its training data containing unsafe content, humans should always be involved in evaluating the outcomes of generative AI to assess the validity of the results. While the human-in-the-loop approach is easy to maintain when the number of generated items is limited, it becomes less practical when the number of generated data points or items increases. When manual inspection by humans is not possible, the importance of evaluating the generated content for biases such as class imbalance (lack of representation), stereotyping, or otherwise harmful content needs to be addressed with statistical or computational methods. Fortunately there are tools available for evaluating content that is created with generative AI, such as Stable Bias for images (Luccioni et al., 2023) and Perspective API (Lees et al., 2022) that can be used to evaluate the toxicity of short texts, although it should be noted that even these systems have shortcoming in their ability to detect toxic or otherwise problematic content (Gehman et al., 2020).

In addition to evaluating the results of the models before using them, we argue that research using foundation models should adhere to the following principles to ensure transparency and help make the results reproducible. First, the use of foundation models should be disclosed like any other significant methodological choice, even if it is used only to a limited extent in the paper. Second, sufficient documentation of the model, version, prompt, and parameters should be provided either in the description of the methodology or in a separate appendix. This type of documentation was demonstrated in the footnotes of Table 1. While in most cases, it is impossible to perfectly reproduce the results since the outputs will vary each time the model is run due to the stochastic nature of generative AI, knowing the model version, the exact parameters, and the prompt will allow for a more informed evaluation of the output. Third, whenever possible, the generated data should be made available, if not publicly, then to reviewers or researchers upon request. These principles are listed in Table 4, with additional information on how to address and evaluate them both as an author and as a reviewer. At the time of writing many publishers have already added policies and guidelines to authors regarding what type of use of generative AI is permitted, and generally documenting it is required in the method or acknowledgment sections (Dwivedi et al., 2023a).

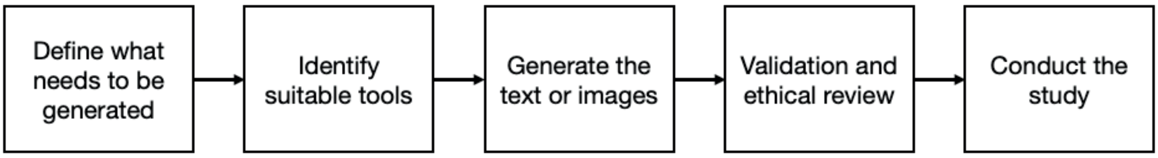

Define what needs to be generated
Identify suitable tools
Generate the text or images
Validation and ethical review
Conduct the study

Figure 3: A process for using foundation models and generative AI

Table 4: Ethical guidelines for using foundation models and generative AI

Principle	**What to evaluate**	**How to address**
Humans should be kept in the loop	Is there any step where subjects of an experiment are shown generated data? Or is an ML model trained with generated content?	All materials shown to humans should be evaluated by humans. Generated data should be tested rigorously for biases using statistical methods.
Disclose the use of foundation models and generative AI	Is it clearly stated where and how foundation models and generative AI were used?	Describe the use of foundation models in a manner that the setup could be replicated. For an example, see Rossi *et al.*, 2023.
Provide sufficient documentation	Is the name of the model, version of the model, as well as the used prompts and parameters listed?	See the footnotes of Table 1, Figure 1 and Figure 2 for an example.
Provide access to the generated data	Can the raw generated data be made available as a supplement or provided upon request?	Consider all generated data as any other data used in research.

To summarize the entire process of using foundation models and generative AI, we illustrate a five-step process that we used in one of our recent papers (Rossi et al., 2023) in Figure 3. The process starts with identifying what needs to be generated, which is followed by reviewing the available tools and selecting the most suitable one given the task and available resources. In our project, we needed to produce realistic social media profiles and posts, and thus, we needed a model capable of generating text as well as a model for generating profile pictures and ended up selecting GPT-3 and StyleGAN. In the third step, the content, such as text, images, or both, is generated using the foundation model. At this point, attention should be paid to documenting the process so that none of the settings used are lost. As an example, in our project, we were using GPT-3 through the API via a Jupyter notebook, so it was easy to maintain information on what the parameters were as they needed to be explicitly stated and were stored then within the file. This is followed by the fourth step, which contains rigorous validation that the outputs of the generative AI are not biased, harmful, or otherwise of undesirable quality. As noted previously, this step is easy if the volume of generated content is limited but becomes quickly challenging as the number of generated data increases. In our case, we generated approximately 30 profile pictures and 30 social media posts, and thus, a manual inspection of each generated item was conducted. After these four steps are completed, the study can be conducted, and the final phase completed.

5. Conclusion

Throughout its history, artificial intelligence has gone through several cycles of optimistic hype followed by disillusionment. Recent developments in generative AI and the introduction of foundation models have already had an impact on academia, although it is unclear how profound this impact will be once the initial excitement wears off. Based on the emerging work that has already adopted generative AI and foundation models, we believe that these new technologies present many opportunities for scholars and that some of these technologies and the methods built on or extended by them will survive the initial excitement and become part of the established methodology of information systems research.

To demonstrate the possibilities of foundation models for IS research, we evaluated the previous work using these models in related fields and pointed to a few early examples within information systems research. To help future researchers in their endeavors, we outlined usage scenarios for text and image generation and gave a list of tools that can be used for each type of task. In terms of method, we stress that the users of these tools should record and publish the prompts they have used to generate the data for reproduction and evaluation purposes. Finally, we outlined potential ethical issues and provided a list of principles for how to avoid them. Thus, we have outlined a straightforward approach for utilizing foundation models in IS research.

While we have focused in this editorial on how generative AI can augment experiments and produce synthetic data for other quantitative research methods, there are vast possibilities beyond these. In the current state of generative AI, it is not sentient, and neither can we see it producing new theories on its own, so we have omitted these considerations from this paper. Ultimately, we hope to see more information systems research that uses foundational models like those described in this paper, as well as entirely new ways while adhering to ethical guidelines and best practices.

References

Barr, K. (2022). AI Image Generators Routinely Display Gender and Cultural Bias. Gizmodo. https://gizmodo.com/ai-dall-e-stability-ai-stable-diffusion-1849728302 (accessed 14 March 2023)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2021, pp. 610–623. FAccT '21. Association for Computing Machinery. Available at: https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E, et al. (2022). On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv. Available at: https://doi.org/10.48550/arXiv.2108.07258 (accessed 7 March 2023)

Boyd, A., Tinsley, P., Bowyer, K., & Czajka, A. (2023). The value of AI guidance in human examination of synthetically-generated faces. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(5), pp. 5930–5938. doi: https://doi.org/10.1609/aaai.v37i5.25734

Brown, T., Mann, B., Ryder, et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chambon, P., Bluethgen, C., Langlotz, C. P., & Chaudhari, A. (2022). Adapting pretrained vision-language foundational models to medical imaging domains. arXiv Preprint arXiv:2210.04133. Epub ahead of print 2022

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6), 493-497. doi: https://doi.org/10.1038/s41551-021-00751-8

Chowdhery, A., Narang, S., Devlin, J., et al. (2022). Palm: Scaling language modeling with pathways. arXiv Preprint arXiv:2204.02311. Epub ahead of print 2022

Chung, J. J. Y., Kamar, E., & Amershi, S. (2023). Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. arXiv Preprint arXiv:2306.04140.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's' Human'Is Not Gold: Evaluating Human Evaluation of Generated Text. arXiv Preprint arXiv:2107.00061. Epub ahead of print 2021

Collins, E., & Ghahramani, Z. (2021). LaMDA: Our breakthrough conversation technology. Google. https://blog.google/technology/ai/lamda/ (accessed 10 March 2023)

Cosmo, L. D. (2022). Google Engineer Claims AI Chatbot Is Sentient: Why That Matters. Scientific American. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/ (accessed 10 March 2023)

Dean, J. (2022). A Golden Decade of Deep Learning: Computing Systems & Applications. Daedalus, 151(2), 58–74. doi: https://doi.org/10.1162/daed_a_01900

Dennis, A. R., & Valacich, J. S. (2001). Conducting experimental research in information systems. Communications of the Association for Information Systems, 7(1), 5. doi: https://doi.org/10.17705/1CAIS.00705

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186. doi: https://doi.org/10.48550/arXiv.1810.04805

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., & others. (2023a). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71. doi: https://doi.org/10.1108/IJCHM-05-2023-0686

Dwivedi, Y. K., Pandey, N., Currie, W., & Micu, A. (2023b). Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: Practices, challenges and research agenda. International Journal of Contemporary Hospitality Management. doi: https://doi.org/10.1108/IJCHM-05-2023-0686

Eckerli, F., & Osterrieder, J. (2021). Generative adversarial networks in finance: An overview. arXiv Preprint arXiv:2106.06364. doi: https://doi.org/10.48550/arXiv.2106.06364

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and Content-Guided Video Synthesis with Diffusion Models (arXiv:2302.03011). arXiv. http://arxiv.org/abs/2302.03011

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 119 (2016).

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., & Wen, J.-R. (2022). Towards artificial general intelligence via a multimodal foundation model. Nature Communications, 13(1), Article 1. doi: https://doi.org/10.1038/s41467-022-30761-2

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30(4), 681–694. doi: https://doi.org/10.1007/s11023-020-09548-1

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321–331. doi: https://doi.org/10.1016/j.neucom.2018.09.013

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369. doi: https://doi.org/10.18653/v1/2020.findings-emnlp.301

Geluykens, J., Mitrović, S., Ortega Vázquez, C. E., Laino, T., Vaucher, A., & De Weerdt, J. (2021). Neural Machine Translation for Conditional Generation of Novel Procedures. Hawaii International Conference on System Sciences. doi: https://doi.org/10.24251/HICSS.2021.132

Guillaudeux, M., Rousseau, O., Petot, J., Bennis, Z., Dein, C.-A., Goronflot, T., Vince, N., Limou, S., Karakachoff, M., Wargny, M., & others. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. Npj Digital Medicine, 6(1), 37. doi: https://doi.org/10.1038/s41746-023-00771-5

Heaven, W. D. (2022, December 18). Why Meta's latest large language model survived only three days online. MIT Technology Review. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (accessed 10 March 2023)

Huang, K. (2023, January 16). Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach. The New York Times. https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html (accessed 10 March 2023)

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 248:1-248:38. doi: https://doi.org/10.1145/3571730

Kar, A. K., Varsha, P., & Rajan, S. (2023). Unravelling the Impact of Generative Artificial Intelligence (GAI) in Industrial Applications: A Review of Scientific and Grey Literature. Global Journal of Flexible Systems Management, 1–31. doi: https://doi.org/10.1007/s40171-023-00356-x

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401–4410. doi: https://doi.org/10.1109/tpami.2020.2970919

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8107–8116. doi: https://doi.org/10.1109/CVPR42600.2020.00813

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2021). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv Preprint arXiv:2108.05921.

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? Communications of the ACM, 59(10), 16–17. doi: https://doi.org/10.1145/2983270

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114, 106553. doi: https://doi.org/10.1016/j.chb.2020.106553

Kwon, S., Park, S. H., Lee, G., & Lee, D. (2022). Learning Faces to Predict Matching Probability in an Online Matching Platform. International Conference on Information Systems 2022. https://aisel.aisnet.org/icis2022/digital_commerce/digital_commerce/9

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3197–3207.

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable bias: Analyzing societal representations in diffusion models. arXiv Preprint arXiv:2303.11408.

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. American Behavioral Scientist, 25(2), 179–210. doi: https://doi.org/10.1177/000276428102500205

Merritt, R. (2023, March 13). What Are Foundation Models? NVIDIA Blog. https://blogs.nvidia.com/blog/what-are-foundation-models/ (accessed 10 March 2023)

Murphy, M. (2022, May). What are foundation models? IBM Research Blog. https://research.ibm.com/blog/what-are-foundation-models (accessed 10 March 2023)

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences, 119(8), e2120481119.

Parshall, A. (2023). How This AI Image Won a Major Photography Competition'. Scientific American, 21. https://www.scientificamerican.com/article/how-my-ai-image-won-a-major-photography-competition/ (accessed 10 March 2023)

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation (arXiv:2102.12092). arXiv. http://arxiv.org/abs/2102.12092

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10674–10685. doi: https://doi.org/10.1109/CVPR52688.2022.01042

Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2023). Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles? Proceedings of the 56th Hawaii International Conference on System Sciences, 134–143. https://hdl.handle.net/10125/102645

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, 4(3), 258–268. doi: https://doi.org/10.1038/s42256-022-00458-8

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1–48. doi: https://doi.org/10.1186/s40537-019-0197-0

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. International Journal of Information Management, 74, article number 102700. doi: https://doi.org/10.1016/j.ijinfomgt.2023.102700

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. Information Systems Research. doi: https://doi.org/10.1287/isre.2023.ed.v34.n2

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. arXiv Preprint arXiv:2302.07944.

West, J., & Bergstrom, C. (n.d.). Which Face Is Real? Retrieved March 14, 2023, from https://www.whichfaceisreal.com/learn.html

Wullach, T., Adler, A., & Minkov, E. (2021). Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech. Findings of the Association for Computational Linguistics: EMNLP 2021, 4699–4705. doi: https://doi.org/10.18653/v1/2021.findings-emnlp.402

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5), 820–838. doi: https://doi.org/10.1109/JPROC.2021.3054390

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43–76. doi: https://doi.org/10.1109/JPROC.2020.3004555

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or,
How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning
– A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus
– Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain
Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter.
En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India’s National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism’

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin
– en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy
Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness
The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management
A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship
Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness
An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur
Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing
The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence
A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog
Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management
- A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat
En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd
Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing
A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions
in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være?
Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sense-making in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a
Partly-Digital World
Comparing Digital Piracy and Legal
Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in
organizational decision-making

11. Mia Kaspersen
The construction of social and
environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading
names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte
til velgørenhed og understøtte
relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt
perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig
kontekst, strategisk tilgang, samt
anvendte redskaber og teknologier for
udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization
of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural
Production
The case of travel guidebook
production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie
i det 19. århundrede
Omtydningen af melankolien og
manien som bipolære stemningslidelser
i dansk sammenhæng under hensyn til
dannelsen af det moderne følelseslivs
relative autonomi.
En problematiserings- og erfarings-
analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters
of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of
Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize
social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL
RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om
selvledelsens paradoksale dynamik og
eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis
– barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life
- Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelsesstudie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity:
Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering: Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig – er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm – A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was
The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation
Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone
With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope
Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics
Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions
– on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A
– Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats?
A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management
A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management
Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management
– Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me
- The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration
A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce:
Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca
An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability
in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'. A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World
Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber
Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts
Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity
– How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation
The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition.
Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsification and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints
A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS
Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics:
-An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
   ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
   Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
   Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
   Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
   Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
   Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
   Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
   The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
   Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
    Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
    Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
    Essays on Crude Oil Tanker Markets

13. Cancan Wang
    Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
    Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
    Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
    Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
    The Comparative Dynamics of Private Governance
    The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
    STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
    Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
    Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability
Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

23. Henriette Sophia Groskopff Tvede Schleimann
Creating innovation through collaboration – Partnering in the maritime sector

24. Kristian Steensen Nielsen
The Role of Self-Regulation in Environmental Behavior Change

25. Lydia L. Jørgensen
Moving Organizational Atmospheres

26. Theodor Lucian Vladasel
Embracing Heterogeneity: Essays in Entrepreneurship and Human Capital

27. Seidi Suurmets
Contextual Effects in Consumer Research: An Investigation of Consumer Information Processing and Behavior via the Applicati on of Eye-tracking Methodology

28. Marie Sundby Palle Nickelsen
Reformer mellem integritet og innovation: Reform af reformens form i den danske centraladministration fra 1920 til 2019

29. Vibeke Kristine Scheller
The temporal organizing of same-day discharge: A tempography of a Cardiac Day Unit

30. Qian Sun
Adopting Artificial Intelligence in Healthcare in the Digital Age: Perceived Challenges, Frame Incongruence, and Social Power

31. Dorthe Thorning Mejlhede
Artful change agency and organizing for innovation – the case of a Nordic fintech cooperative

32. Benjamin Christoffersen
Corporate Default Models:
Empirical Evidence and Methodical Contributions

33. Filipe Antonio Bonito Vieira
Essays on Pensions and Fiscal Sustainability

34. Morten Nicklas Bigler Jensen
Earnings Management in Private Firms: An Empirical Analysis of Determinants and Consequences of Earnings Management in Private Firms

2020

1. Christian Hendriksen
Inside the Blue Box: Explaining industry influence in the International Maritime Organization

2. Vasileios Kosmas
Environmental and social issues in global supply chains:
Emission reduction in the maritime transport industry and maritime search and rescue operational response to migration

3. Thorben Peter Simonsen
The spatial organization of psychiatric practice: A situated inquiry into 'healing architecture'

4. Signe Bruskin
The infinite storm: An ethnographic study of organizational change in a bank

5. Rasmus Corlin Christensen
Politics and Professionals: Transnational Struggles to Change International Taxation

6. Robert Lorenz Törmer
The Architectural Enablement of a Digital Platform Strategy

7. Anna Kirkebæk Johansson Gosovic
Ethics as Practice: An ethnographic study of business ethics in a multinational biopharmaceutical company

8. Frank Meier
Making up leaders in leadership development

9. Kai Basner
*Servitization at work:
On proliferation and containment*

10. Anestis Keremis
Anti-corruption in action: How is anti-corruption practiced in multinational companies?

11. Marie Larsen Ryberg
Governing Interdisciolinarity: Stakes and translations of interdisciplinarity in Danish high school education.

12. Jannick Friis Christensen
Queering organisation(s): Norm-critical orientations to organising and researching diversity

13. Thorsteinn Sigurdur Sveinsson
Essays on Macroeconomic Implications of Demographic Change

14. Catherine Casler
Reconstruction in strategy and organization: For a pragmatic stance

15. Luisa Murphy
*Revisiting the standard organization of multi-stakeholder initiatives (MSIs):
The case of a meta-MSI in Southeast Asia*

16. Friedrich Bergmann
Essays on International Trade

17. Nicholas Haagensen
European Legal Networks in Crisis: The Legal Construction of Economic Policy

18. Charlotte Biil
Samskabelse med en sommerfugle-model: Hybrid ret i forbindelse med et partnerskabsprojekt mellem 100 selvejende daginstitutioner, deres paraplyorganisation, tre kommuner og CBS

19. Andreas Dimmelmeier
The Role of Economic Ideas in Sustainable Finance: From Paradigms to Policy

20. Maibrith Kempka Jensen
*Ledelse og autoritet i interaktion
- En interaktionsbaseret undersøgelse af autoritet i ledelse i praksis*

21. Thomas Burø
LAND OF LIGHT: Assembling the Ecology of Culture in Odsherred 2000-2018

22. Prins Marcus Valiant Lantz
Timely Emotion: The Rhetorical Framing of Strategic Decision Making

23. Thorbjørn Vittenhof Fejerskov
Fra værdi til invitationer - offentlig værdiskabelse gennem affekt, potentialitet og begivenhed

24. Lea Acre Foverskov
Demographic Change and Employment: Path dependencies and institutional logics in the European Commission

25. Anirudh Agrawal
A Doctoral Dissertation

26. Julie Marx
Households in the housing market

27. Hadar Gafni
Alternative Digital Methods of Providing Entrepreneurial Finance

28. Mathilde Hjerrild Carlsen
Ledelse af engagementer: En undersøgelse af samarbejde mellem folkeskoler og virksomheder i Danmark

29. Suen Wang
Essays on the Gendered Origins and Implications of Social Policies in the Developing World

30. Stine Hald Larsen
The Story of the Relative: A Systems-Theoretical Analysis of the Role of the Relative in Danish Eldercare Policy from 1930 to 2020

31. Christian Casper Hofma
Immersive technologies and organizational routines: When head-mounted displays meet organizational routines

32. Jonathan Feddersen
The temporal emergence of social relations: An event-based perspective of organising

33. Nageswaran Vaidyanathan
ENRICHING RETAIL CUSTOMER EXPERIENCE USING AUGMENTED REALITY

2021

1. Vanya Rusinova
The Determinants of Firms' Engagement in Corporate Social Responsibility: Evidence from Natural Experiments

2. Lívia Lopes Barakat
Knowledge management mechanisms at MNCs:
The enhancing effect of absorptive capacity and its effects on performance and innovation

3. Søren Bundgaard Brøgger
Essays on Modern Derivatives Markets

4. Martin Friis Nielsen
Consuming Memory: Towards a conceptualization of social media platforms as organizational technologies of consumption

05. Fei Liu
Emergent Technology Use in Consumer Decision Journeys: A Process-as-Propensity Approach

06. Jakob Rømer Barfod
Ledelse i militære højrisikoteams

07. Elham Shafiei Gol
Creative Crowdwork Arrangements

08. Árni Jóhan Petersen
Collective Imaginary as (Residual) Fantasy: A Case Study of the Faroese Oil Bonanza

09. Søren Bering
"Manufacturing, Forward Integration and Governance Strategy"

10. Lars Oehler
Technological Change and the Decomposition of Innovation:
Choices and Consequences for Latecomer Firm Upgrading: The Case of China's Wind Energy Sector

11. Lise Dahl Arvedsen
Leadership in interaction in a virtual context:
A study of the role of leadership processes in a complex context, and how such processes are accomplished in practice

12. Jacob Emil Jeppesen
Essays on Knowledge networks, scientific impact and new knowledge adoption

13. Kasper Ingeman Beck
Essays on Chinese State-Owned Enterprises:
Reform, Corporate Governance and Subnational Diversity

14. Sönnich Dahl Sönnichsen
Exploring the interface between public demand and private supply for implementation of circular economy principles

15. Benjamin Knox
Essays on Financial Markets and Monetary Policy

16. Anita Eskesen
Essays on Utility Regulation:
Evaluating Negotiation-Based Approaches inthe Context of Danish Utility Regulation

17. Agnes Guenther
Essays on Firm Strategy and Human Capital

18. Sophie Marie Cappelen
Walking on Eggshells: The balancing act of temporal work in a setting of culinary change

19. Manar Saleh Alnamlah
About Gender Gaps in Entrepreneurial Finance

20. Kirsten Tangaa Nielsen
Essays on the Value of CEOs and Directors

21. Renée Ridgway
Re:search - the Personalised Subject vs. the Anonymous User

22. Codrina Ana Maria Lauth
IMPACT Industrial Hackathons:
Findings from a longitudinal case study on short-term vs long-term IMPACT implementations from industrial hackathons within Grundfos

23. Wolf-Hendrik Uhlbach
Scientist Mobility: Essays on knowledge production and innovation

24. Tomaz Sedej
Blockchain technology and inter-organizational relationships

25. Lasse Bundgaard
Public Private Innovation Partnerships:
Creating Public Value & Scaling Up Sustainable City Solutions

26. Dimitra Makri Andersen
Walking through Temporal Walls:
Rethinking NGO Organizing for Sustainability through a Temporal Lens on NGO-Business Partnerships

27. Louise Fjord Kjærsgaard
Allocation of the Right to Tax Income from Digital Products and Services:
A legal analysis of international tax treaty law

28. Sara Dahlman
Marginal alternativity: Organizing for sustainable investing

29. Henrik Gundelach
Performance determinants:
An Investigation of the Relationship between Resources, Experience and Performance in Challenging Business Environments

30. Tom Wraight
Confronting the Developmental State: American Trade Policy in the Neoliberal Era

31. Mathias Fjællegaard Jensen
Essays on Gender and Skills in the Labour Market

32. Daniel Lundgaard
Using Social Media to Discuss Global Challenges: Case Studies of the Climate Change Debate on Twitter

33. Jonas Sveistrup Søgaard
Designs for Accounting Information Systems using Distributed Ledger Technology

34. Sarosh Asad
CEO narcissism and board composition: Implications for firm strategy and performance

35. Johann Ole Willers
Experts and Markets in Cybersecurity On Definitional Power and the Organization of Cyber Risks

36. Alexander Kronies
Opportunities and Risks in Alternative Investments

37. Niels Fuglsang
The Politics of Economic Models: An inquiry into the possibilities and limits concerning the rise of macroeconomic forecasting models and what this means for policymaking

38. David Howoldt
Policy Instruments and Policy Mixes for Innovation: Analysing Their Relation to Grand Challenges, Entrepreneurship and Innovation Capability with Natural Language Processing and Latent Variable Methods

2022

01. Ditte Thøgersen
Managing Public Innovation on the Frontline

02. Rasmus Jørgensen
Essays on Empirical Asset Pricing and Private Equity

03. Nicola Giommetti
Essays on Private Equity

04. Laila Starr
When Is Health Innovation Worth It? Essays On New Approaches To Value Creation In Health

05. Maria Krysfeldt Rasmussen
Den transformative ledelsesbyrde – etnografisk studie af en religionsinspireret ledelsesfilosofi i en dansk modevirksomhed

06. Rikke Sejer Nielsen
Mortgage Decisions of Households: Consequences for Consumption and Savings

07. Myriam Noémy Marending
Essays on development challenges of low income countries: Evidence from conflict, pest and credit

08. Selorm Agbleze
A BEHAVIORAL THEORY OF FIRM FORMALIZATION

09. Rasmus Arler Bogetoft
Rettighedshavers faktisk lidte tab i immaterialretssager: Studier af dansk ret med støtte i økonomisk teori og metode

10. Franz Maximilian Buchmann
Driving the Green Transition of the Maritime Industry through Clean Technology Adoption and Environmental Policies

11. Ivan Olav Vulchanov
The role of English as an organisational language in international workplaces

12. Anne Agerbak Bilde
TRANSFORMATIONER AF SKOLELEDELSE - en systemteoretisk analyse af hvordan betingelser for skoleledelse forandres med læring som genstand i perioden 1958-2020

13. JUAN JOSE PRICE ELTON
EFFICIENCY AND PRODUCTIVITY ANALYSIS: TWO EMPIRICAL APPLICATIONS AND A METHODOLOGICAL CONTRIBUTION

14. Catarina Pessanha Gomes
The Art of Occupying: Romanticism as Political Culture in French Prefigurative politics

15. Mark Ørberg
Fondsretten og den levende vedtægt

16. Majbritt Greve
Maersk's Role in Economic Development: A Study of Shipping and Logistics Foreign Direct Investment in Global Trade

17. Sille Julie J. Abildgaard
Doing-Being Creative: Empirical Studies of Interaction in Design Work

18. Jette Sandager
Glitter, Glamour, and the Future of (More) Girls in STEM: Gendered Formations of STEM Aspirations

19. Casper Hein Winther
Inside the innovation lab - How paradoxical tensions persist in ambidextrous organizations over time

20. Nikola Kostić
Collaborative governance of inter-organizational relationships: The effects of management controls, blockchain technology, and industry standards

21. Saila Naomi Stausholm
Maximum capital, minimum tax: Enablers and facilitators of corporate tax minimization

22. Robin Porsfelt
Seeing through Signs: On Economic Imagination and Semiotic Speculation

23. Michael Herburger
Supply chain resilience – a concept for coping with cyber risks

24. Katharina Christiane Nielsen Jeschke
Balancing safety in everyday work - A case study of construction managers' dynamic safety practices

25. Jakob Ahm Sørensen
Financial Markets with Frictions and Belief Distortions

26. Jakob Laage-Thomsen
Nudging Leviathan, Protecting Demos - A Comparative Sociology of Public Administration and Expertise in the Nordics

27. Kathrine Søs Jacobsen Cesko
Collaboration between Economic Operators in the Competition for Public Contracts: A Legal and Economic Analysis of Grey Zones between EU Public Procurement Law and EU Competition Law

28. Mette Nelund
Den nye jord – Et feltstudie af et bæredygtigt virke på Farendløse Mosteri

29. Benjamin Cedric Larsen
Governing Artificial Intelligence – Lessons from the United States and China

30. Anders Brøndum Klein
Kollektiv meningsdannelse iblandt heterogene aktører i eksperimentelle samskabelsesprocesser

31. Stefano Tripodi
Essays on Development Economicis

32. Katrine Maria Lumbye
Internationalization of European Electricity Multinationals in Times of Transition

33. Xiaochun Guo
Dynamic Roles of Digital Currency – An Exploration from Interactive Processes: Difference, Time, and Perspective

34. Louise Lindbjerg
Three Essays on Firm Innovation

35. Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

36. Hanna Nyborg Storm
Cultural institutions and attractiveness – How cultural institutions contribute to the development of regions and local communities

37. Anna-Bertha Heeris Christensen
Conflicts and Challenges in Practices of Commercializing Humans – An Ethnographic Study of Influencer Marketing Work

38. Casper Berg Lavmand Larsen
A Worker-Centered Inquiry into the Contingencies and Consequences of Worker Representation

39. Niels le Duc
The Resource Commitment of Multinational Enterprise R&D Activities

40. Esben Langager Olsen
Change management tools and change managers – Examining the simulacra of change

41. Anne Sophie Lassen
Gender in the Labor Market

42. Alison E. Holm
Corrective corporate responses to accusations of misconduct on societal issues

43. Chenyan Lyu
Carbon Pricing, Renewable Energy, and Clean Growth – A Market Perspective

44. Alina Grecu
UNPACKING MULTI-LEVEL OFFSHORING CONSEQUENCES: Hiring Wages, Onshore Performance, and Public Sentiment

45. Alexandra Lüth
Offshore Energy Hubs as an Emerging Concept – Sector Integration at Sea

2023

01. Cheryl Basil Sequeira
Port Business Development – Digitalisation of Port Authroity and Hybrid Governance Model

02. Mette Suder Franck
Empirical Essays on Technology Supported Learning – Studies of Danish Higher Education

03. Søren Lund Frandsen
States and Experts – Assembling Expertise for Climate Change and Pandemics

04. Guowei Dong
Innovation and Internationalization – Evidence from Chinese Manufacturing Enterprises

05. Eileen Murphy
In Service to Security – Constructing the Authority to Manage European Border Data Infrastructures

06. Bontu Lucie Guschke
THE PERSISTENCE OF SEXISM AND RACISM AT UNIVERSITIES – Exploring the imperceptibility and unspeakability of workplace harassment and discrimination in academia

07. Christoph Viebig
Learning Entrepreneurship – How capabilities shape learning from experience, reflection, and action

08. Kasper Regenburg
Financial Risks of Private Firms

09. Kathrine Møller Solgaard
Who to hire? – A situated study of employee selection as routine, practice, and process

10. Jack Kværnø-Jones
Intersections between FinTech Imaginaries and Traditional Banking – A study of disciplinary, implementary, and parasitic work in the Danish financial sector

11. Stine Quorning
Managing Climate Change Like a Central Banker – The Political Economy of Greening the Monetary Technocracy

12. Amanda Bille
No business without politics – Investigating the political nature of supply chain management

13. Theis Ingerslev Jensen
Essays on Empirical Asset Pricing

14. Ann Fugl-Meyer
The Agile Imperative – A Qualitative Study of a Translation Process in the Danish Tax Administration

15. Nicolai Søgaard Laursen
Longevity risk in reinsurance and equity markets

16. Shelter Selorm Kwesi Teyi
STRATEGIC ENTREPRENEURSHIP IN THE INFORMAL ECONOMY

17. Luisa Hedler
Time, Law and Tech – The introduction of algorithms to courts of law

18. Tróndur Møller Sandoy
Essays on the Economics of Education

19. Nathan Rietzler
Crowdsourcing Processes and Performance Outcomes

20. Sigrid Alexandra Koob
Essays on Democracy, Redistribution, and Inequality

21. David Pinkus
Pension Fund Investment: Implications for the Real Economy

22. Sina Smid
Inequality and Redistribution – Essays on Local Elections, Gender and Corruption in Developing Countries

23. Andreas Brøgger
Financial Economics with Preferences and Frictions

24. Timothy Charlton-Czaplicki
Arendt in the platformised world – Labour, work and action on digital platforms

25. Letícia Vedolin Sebastião
Mindfulness and Consumption: Routes Toward Consumer Self-Control

26. Lotte List
Crisis Sovereignty – The Philosophy of History of the Exception

27. Jeanette Walldorf
Essays on the Economics of Education and Labour Market

28. Juan Camilo Giraldo-Mora
It is Along Ways – Global Payment Infrastructure in Movement

29. Niels Buus Lassen
THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

30. Frederik Bjørn Christensen
Essays on the Intergenerational Welfare State

31. Shama Patel
The Summer of 2020: Situating Digital Media in Scaling Affective Contagion: A Case of the George Floyd Video

32. Federico Jensen
Who rules the waves in the 21st Century?
The international political economy of global shipping

33. Tobias Berggren Jensen
Selvledende organisationer i den offentlige sektor – modsætninger og konflikter i radikal decentralisering

34. Jonathan Harmat
The Affects By Which We Are Torn
Four Essays on Government and Affect

35. Jørgen Valther Hansen
The Big 4 Audit Firms and the Public Interest
Public oversight & Audit Firm Governance

36. Stig Strandbæk Nyman
The Birth of Algorithmic Aspirational Control

37. Morten Tinning
Steaming Ahead
Experiences and the Transition from Sail to Steam

38. Oguzhan Cepni
Essays in Applied Financial Economics

39. Tim Dominik Maurer
Essays on Pension Policy

40. Aixa Y. Alemán-Díaz
Exploring Global Ideas in National Policy for Science, Technology and Innovation an Isomorphic Difference Approach

41. Michael Güldenpfennig
Managing the interrelationships between manufacturing system elements for productivity improvement in the factory

42. Jun Yuan (Julian) Seng
Essays on the political economy of innovative startups

43. Jacek Piosik
Essays on Entrepreneurial Finance

44. Elizabeth Cooper
Tourists on the Edge
Understanding and Encouraging Sustainable Tourist Behaviour in Greenland

2024

01. Marija Sarafinovska
Patients as Innovators: An Empirical Study of Patients' Role in Innovation in the Healthcare Industry

02. Niina Hakala
Corporate Reporting in the Governance of Climate Transition – Framing agency in a financialized world

03. Kasper Merling Arendt
Unleashing Entrepreneurial Education
Developing Entrepreneurial Mindsets, Competencies, and Long-Term Behavior

04. Kerstin Martel
Creating and dissolving 'identity' in global mobility studies - a multi-scalar inquiry of belongingness and becoming on-the-move

05. Sofie Elbæk Henriksen
Big Tech to the Rescue?
An Ethnographic Study of Corporate Humanitarianism in the Refugee Crisis

06. Christina Kjær
Corporate scandals
- in the age of 'responsible business'

07. Anna Stöber
Embedded Self-Managing Modes of Organizing
Empirical Inquiries into Boundaries, Momentum, and Collectivity

08. Lucas Sören Göbeler
Shifting and Shaping
Physicality in Digital Innovation

09. Felix Schilling
Department of International Economics, Government and Business

10. Mathias Lund Larsen
China and the Political Economy of the Green State

11. Michael Bennedsen Hansen
At få sjælen med
En narrativ analyse af danske containersøfolks erindringer, fortidsbrug og identitetskonstruktioner

12. Justyna Agata Bekier
More than a numbers game
Accounting for circular economy performance in collaborative initiatives in cities

13. Frederik Schade
The Question of Digital Responsibility
An Ethnography of Emergent Institutional Formations in the Contemporary Governance of Technology

14. Alexandrina Schmidt
The Mundane in the Digital: A qualitative study of social work and vulnerable clients in Danish job centres

15. Julian Fernandez Mejia
Essays on International Finance

16. Leonie Decrinis
Nudging in the Workplace: Exploring a Micro-level Approach Towards Corporate Sustainability

17. Nina Frausing Pedersen
A Framing Contest between Institutional Actors on Crypto-Asset Policymaking in the EU

18. Amalie Toft Bentsen
The Internal Market & the EU
Climate Regime
Interactions and frictions in the
legal norm systems

19. Sippo Rossi
Bots on Social Media
The Past, Present and Future

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens