# AGENCY, TRUST AND BLAME IN DECISION-MAKING ALGORITHMS: AN ANALYSIS OF TWITTER DISCOURSES

DANIEL HEATON


Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy.

Supervisors: Professor Joel Fischer and Dr Jérémie Clos

October 2024

# DECLARATION

I hereby declare that, except where specific reference is made to the work of others, the contents of this thesis, titled "Agency, Trust And Blame in Decision-Making Algorithms: An Analysis of Twitter Discourses", is my own work from the time of my PhD at the University of Nottingham under the supervision of Professor Joel Fischer and Dr Jérémie Clos, and that it has never been submitted for the award of any degree, diploma, or any similar type of recognition.

*Nottingham, October 2024*

Daniel Heaton

## ABSTRACT

This PhD project focuses on exploring the concepts of agency, blame and trust concerning public-facing decision-making algorithms through an interdisciplinary linguistic approach. Public-facing decision-making algorithms, such as the algorithms underpinning A Level grade calculations in 2020, the NHS Covid-19 contact tracing app and ChatGPT, have increasingly impacted global conversations, yet concerns have emerged regarding their perceived social agency, particularly when negative outcomes arise. Determining responsibility and accountability for these outcomes is challenging due to the complex and opaque nature of how these algorithms are designed and deployed. Trust in such systems is vital for the future development of artificial intelligence technologies, emphasising the need for algorithms to be perceived as trustworthy by design and in practice. Despite this, little exploration exists regarding how trust and blame are influenced by the perceived agency, responsibility and accountability of these systems. By analysing Twitter discourses, where views have been expressed about the three aforementioned public-facing decision-making algorithms, this research aims to provide nuanced insights into how these systems are perceived in society.

The approach used in this PhD thesis involves analysing the relationship between social agency and grammatical agency through computational and discursive linguistics. While popular Natural Language Processing (NLP)-based tools – like sentiment analysis, topic modelling and emotion detection – are commonly used for social media research, they struggle to capture the nuanced discursive and conversational aspects of opinions on decision-making algorithms. To address this gap, this research adopts a combined approach, incorporating Corpus Linguistics (CL) and Discourse Analysis (DA),

with an aim to provide deeper insights into the nuances of discourses surrounding agency, blame and trust in decision-making algorithms, which traditional NLP-based methods may overlook. Methodologically, the research involved three key steps: initial analysis using NLP-based tools, followed by deeper examination using CL tools to explore grammatical constructions and, finally, employing DA with Social Actor Representation (SAR). Here, active and passive presentations were examined and social actors identified, providing insights into trust or blame attributed to decision-making algorithms on Twitter.

For the first of the three case studies, the 2020 A Level algorithm, the initial NLP analysis highlighted discussions around government involvement, flaws in the algorithm and impact on schools, teachers and students, with sentiment remaining negative throughout the discourse. Through the CL and DA investigation, it was found that Twitter users attributed blame to various social actors, including the algorithm itself, the UK government and Ofqual for the A Level results. Users employed active agency and personalisation in their tweets, with blame shifting more prominently towards these entities as the discourse progressed, which was seen through techniques like assimilation and individualism.

Secondly, the study into the NHS Covid-19 app showed three primary topics, with an increase in discussion related to the government's role and a dip in sentiment occurring at the time of the second national lockdown. CL and DA exploration revealed that Twitter users predominantly portrayed the app as a social actor, particularly in informing, instructing, and disrupting, while also assigning responsibility for users' welfare and safety. Despite occasional passive presentations and instances of ridicule, the discourse consistently emphasised the app's perceived responsibility for processing information, especially during significant events like its launch, subsequent lockdowns, and the 'pingdemic' phase, highlighting the app's significant social impact and the public's expectations of its role.

Finally, the ChatGPT discourse saw topics spanning text generation, chatbot development and cryptocurrency, alongside a more positive sentiment trajectory. ChatGPT was, again, predominantly depicted as an active social actor, influencing content creation and information dissemination, while trust in its outputs evolved over time, influenced by perceptions of its agency and occasional blame for errors. There were times where, even though ChatGPT was portrayed actively, its status as a social actor was diluted due to users presenting it solely as an information source.

When looking holistically at the three discourses, while all three algorithms were portrayed actively as social actors, variations in blame attribution and trust were observed. ChatGPT was presented as more trustworthy and less blameworthy compared to the A Level algorithm and the Covid-19 app. The two pandemic-based case studies showcased how the agency that was ascribed to the systems from Twitter users unveiled a more overt degree of accountability and responsibility, resulting in decreased trust and clearer blame.

In terms of the main contributions of this thesis, this work provides insights into the dynamics of discourse surrounding decision-making algorithms through an analysis of Twitter discourses. This research offers a nuanced examination of how these algorithms are portrayed as social actors, highlighting strategic manipulations of blame attribution and foregrounding trust and blame perceptions in response to societal events. Additionally, this thesis demonstrates the effectiveness of integrating computational and discursive linguistic analytics, laying the foundations for future research using this approach. Overall, by understanding the roles, contexts and perceptions associated with decision-making algorithms, researchers can contribute to the responsible development and deployment of decision-making algorithms. This understanding can help foster public trust and effectively address societal concerns, particularly those related to perceptions of social algorithmic agency and its implications for trust and blame attribution.

# ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisors, Professor Joel Fischer and Dr Jérémie Clos, for their guidance and support throughout the PhD. I am very grateful to Joel for his belief in my research potential from the very beginning of my studies, and to Jérémie for his insight and fresh perspective on all things NLP. Encouraging me to publish my work has been such an enjoyable challenge and, for that, I am incredibly appreciative. The PhD is a pretty unique combination of all of our interests and would certainly not be the same without their involvement.

Of course, I was also incredibly lucky to have a third supervisor, Dr Elena Nichele, for the first three years of the PhD. This journey has felt like one tremendous, four-year identity crisis at times, and Elena helped me realise that I was not the only researcher occupying the vague, liminal space between Computer Science and Linguistics. Her extraordinary support and unwavering commitment to the PhD, even after she departed to pastures new, leaves me very thankful indeed.

I would also like to thank Professor Svenja Adolphs and Professor Kate Devlin for being wonderful and detailed examiners. It was a true pleasure to listen and respond to their questions and insights and I only wish that all PhD vivas ran as smoothly and diligently as mine. In particular, a huge thanks to Svenja for listening to my ideas before the PhD journey even began, reviewing my work every year, and, most importantly, encouraging me to remove every instance of 'hybrid' from the thesis!

The PhD would not have been possible, in the first instance, without the Horizon Centre for Doctoral Training's 'Creating Our Lives in Data' programme. A huge thanks goes to all the staff who made Horizon a warm and enriching place to learn and thrive. This thanks also

extends to the Horizon 2020 cohort: from being faces on a screen during lockdown in 2020, to being fully realised colleagues and friends, I doubt that the PhD would have been as much fun without having each other to bounce off.

I would like to explicitly thank my PhD partner, the Trustworthy Autonomous Systems Hub, for the development opportunities and project support. TAS has been such an invaluable resource and the community of researchers within it have been fantastic sounding boards for ideas. I'd like to give a special thanks to Xin Liew, who manually reviewed so many NLP outputs with me that I'm not sure how her eyes coped. I would also like to thank my other industry partner, Ipsos UK, in particular Colin Strong, for offering me a placement in summer 2022. I learnt such a great deal during our time working together and have thoroughly enjoyed being part of the consultancy team ever since; I'm certain that there will be many more exciting projects on which we will collaborate.

Professionally, there are too many people to express my gratitude to for getting me to this point. However, a special thanks goes to all my former colleagues working in schools who listened to my 'aspiration rants' about doing a PhD. They not only nodded politely and refrained from humouring me, but for also genuinely believed that I could do this. Personally, while there are still too many to thank individually, there are a few specific mentions I would like to give. To my family – my Mum, my Dad and brother Tom – a huge thanks for the encouragement and unwavering support in all my endeavours, but particularly this one. And, to Jon: no words could ever do it justice.

# CONTENTS


1 INTRODUCTION 1

1.1 Background and Motivation 1

1.2 Research Question 6

1.3 Research Objectives 7

1.4 Publication of the Thesis 8

1.5 Research Areas 9

1.6 Contributions 11

1.7 Structure of the Thesis 12

I BACKGROUND AND APPROACH 15

2 LITERATURE REVIEW 17

2.1 Grammatical and Social Agency 17

2.1.1 Social Actor Representation 19

2.2 Reviewing Agency, Responsibility and Accountability 22

2.2.1 Agency and Decision-Making Algorithms 22

2.2.2 Responsibility and Decision-Making Algorithms 27

2.2.3 Accountability and Decision-Making Algorithms 29

2.2.4 Section Summary 31

2.3 Influencing Trust and Blame 32

2.3.1 Trust and Decision-Making Algorithms 33

2.3.2 Blame and Decision-Making Algorithms 36

2.3.3 Section Summary 40

2.4 Decision-Making Algorithm Case Studies 41

2.4.1 The 2020 A Level Algorithm 41

2.4.2 The NHS Covid-19 App 44

2.4.3 ChatGPT 47
2.4.4 Section Summary 54
2.5 Chapter Summary 56
3 APPROACH 59
3.1 Data Collection 60
3.2 Overarching Analysis Process 61
3.3 NLP-Based Computational Linguistic Approaches 64
3.3.1 Background 64
3.3.2 Application 72
3.3.3 Section Summary 77
3.4 Corpus Linguistics 79
3.4.1 Background 79
3.4.2 Application 81
3.4.3 Section Summary 82
3.5 Discourse Analysis 83
3.5.1 Background 83
3.5.2 Application 88
3.5.3 Section Summary 90
3.6 Chapter Summary 91
II EMPIRICAL WORK 93
4 2020 A LEVEL ALGORITHM 95
4.1 Study Background 95
4.1.1 Study Research Question and Objectives 97
4.2 Study Approach 98
4.2.1 Data Collection and Processing 98
4.2.2 NLP-Based Techniques 99
4.2.3 Corpus Linguistics and Discourse Analysis 100
4.3 NLP-Based Techniques Analysis 101
4.3.1 Topics 101
4.3.2 Sentiment 107

4.3.3 Emotions 111
4.3.4 Section Summary 115
4.4 Corpus Linguistics and Discourse Analysis 116
4.4.1 Keyword Analysis of Potential Social Actors 116
4.4.2 The Algorithm 117
4.4.3 Ofqual 122
4.4.4 The UK Government 127
4.4.5 Students 132
4.4.6 Section Summary 133
4.5 Discussion 134
4.5.1 Topics, Sentiment and Emotions 134
4.5.2 Blame for The A Level Results 136
4.5.3 CL and DA to Complement NLP-Based Tools 138
4.5.4 Limitations and Future Work 140
4.6 Chapter Summary 141
5 COVID-19 CONTACT-TRACING APP 145
5.1 Study Background 145
5.1.1 Study Research Question and Objectives 147
5.2 Study Approach 148
5.2.1 Data 148
5.2.2 NLP-Based Techniques 149
5.2.3 Corpus Linguistics 150
5.2.4 Discourse Analysis 152
5.3 NLP-Based Techniques Analysis 153
5.3.1 Topics 153
5.3.2 Sentiment 158
5.3.3 Emotions 163
5.3.4 Section Summary 166
5.4 Corpus Linguistics and Discourse Analysis 167
5.4.1 Keyword Analysis 167

5.4.2 Presentations of The App: Timeline Overview 168
5.4.3 App Launch: September 2020 169
5.4.4 Early Months: October-December 2020 175
5.4.5 Second National Lockdown: January-February 2021 179
5.4.6 Later Months: March-June 2021 182
5.4.7 'Pingdemic': July 2021 184
5.4.8 Section Summary 188
5.5 Discussion 189
5.5.1 Topics, Sentiment and Emotions 190
5.5.2 Trends of Active Agency 191
5.5.3 Trends of Passivisation 196
5.5.4 Limitations and Future Work 197
5.6 Chapter Summary 198
6 CHATGPT 201
6.1 Study Background 201
6.1.1 Study Research Question and Objectives 205
6.2 Study Approach 206
6.2.1 Data Collection and Processing 206
6.2.2 Natural Language Processing Approaches 207
6.2.3 Corpus Linguistics 208
6.2.4 Discourse Analysis 209
6.3 NLP-Based Techniques Analysis 210
6.3.1 Topics 210
6.3.2 Sentiment 218
6.3.3 Emotions 222
6.4 Corpus Linguistics and Discourse Analysis 227
6.4.1 Timeline Overview of Results 227
6.4.2 Time Period 1: Launch (November to December 2022) 228
6.4.3 Time Period 2: Popularity (January 2023) 236

6.4.4 Time Period 3: Developments (February to March 2023) 242
6.4.5 Section Summary 249
6.5 Discussion 249
6.5.1 NLP-Based Analysis 250
6.5.2 Trends of Active Agency 254
6.5.3 Passive Presentations of ChatGPT 260
6.5.4 Ethical Implications 261
6.5.5 Limitations and Future Work 262
6.6 Chapter Summary 264

III SYNOPSIS 269

7 DISCUSSION 271
7.1 Introduction 271
7.1.1 Research Question and Sub-Questions 271
7.1.2 Research Objectives 272
7.2 Sub-research questions 272
7.2.1 SRQ1: A Level Algorithm 272
7.2.2 SRQ2: Covid-19 App 275
7.2.3 SRQ3: ChatGPT 280
7.3 Overarching Research Question 284
7.3.1 Depictions of Active Agency 284
7.3.2 Passivisation 287
7.3.3 Trust 288
7.3.4 Blame 290
7.3.5 Section Summary 292
7.4 Implications for Research 295
7.5 Limitations 297
7.6 Chapter Summary 299
8 CONCLUSION 303
8.1 Thesis Summary 303
8.2 Main Contributions 307
8.3 Future Work 309

8.4 Concluding Remarks 311

IV APPENDIX 313

A APPENDIX A: PRIVACY NOTICE 315

B APPENDIX B: PROJECT INFORMATION 319

BIBLIOGRAPHY 321

# LIST OF FIGURES


Figure 1 An illustration of the relationship between the overarching thesis research question (RQ) and the sub-research questions (SRQs). 7

Figure 2 Figure to illustrate the proposed analytical approach. 59

Figure 3 A diagram to illustrate the borrowed best practices analysis process, first set out by Heaton et al. (2023b). 76

Figure 4 Trajectories of topics detected in tweets relating to the A Level algorithm. 102

Figure 5 TextBlob and VADER sentiment analysis of tweets relating to Ofqual A Level algorithm in August and September 2020. 108

Figure 6 Emotions detected in tweets relating to the Ofqual A Level algorithm. 112

Figure 7 Temporal trajectory of LogDice scores of collocates of *algorithm* - part A. 119

Figure 8 Temporal trajectory of LogDice scores of collocates of *algorithm* - part B. 119

Figure 9 Temporal trajectory of LogDice scores of collocates of *Ofqual* - part A. 124

Figure 10 Temporal trajectory of LogDice scores of collocates of *Ofqual* - part B. 124

Figure 11 Temporal trajectory of LogDice scores of collocates of *government* - part A. 129

Figure 12 Temporal trajectory of LogDice scores of collocates of *government* - part B. 129

Figure 13 Trajectories of topics detected in tweets containing 'NHSCovid19App'. 155
Figure 14 Evolution of the sentiment of tweets containing 'NHSCovid19App' using TextBlob and VADER from September 2020 to July 2021. 159
Figure 15 Emotions detected in tweets relating to 'NHSCovid19App'. 164
Figure 16 Comparison of the percentage of each theme found when the app is presented as a social actor. 189
Figure 17 Trajectories of ChatGPT compared with other generative AI systems. Data source: Google Trends (https://www.google.com/trends) 203
Figure 18 Trajectories of topics detected in tweets relating to ChatGPT. 212
Figure 19 Evolution of the sentiment of tweets relating to ChatGPT using VADER from November 2022 to March 2023. 219
Figure 20 Emotions detected in tweets relating to 'ChatGPT'. 223

# LIST OF TABLES


Table 1 The different approaches used in this thesis, including the data and analysis procedures and the intended insight. 63

Table 2 Ranking of the top 10 lexical items associated with each latent topic 102

Table 3 The top ten words with the highest keyness score. 117

Table 4 Collocational strength of *algorithm*. 118

Table 5 Collocational strength of *Ofqual*. 123

Table 6 Collocational strength of *government*. 128

Table 7 Collocational strength of *students*. 133

Table 8 Ranking of the top 10 lexical items associated with each latent topic 154

Table 9 The top ten words with the highest keyness score. 168

Table 10 Frequency of active and passive presentation of *app*. 169

Table 11 Top 20 words ranked by collocational strength of 'app' + R1 in September 2020. 170

Table 12 Top 6 words ranked by collocational strength of 'by the app' + L1 in September 2020. 172

Table 13 Top 20 word ranked by collocational strength of 'app' + R1 in October, November and December 2020. 176

Table 14 Top 3 words ranked by collocational strength of 'by the app' + L1 in October, November and December 2020. 178

Table 15 Top 18 words ranked by collocational strength of 'app' + R1 in January and February 2021. 180
Table 16 Top 4 words ranked by collocational strength of 'by the app' + L1 in January and February 2021. 181
Table 17 Top 6 words ranked by collocational strength of 'app' + R1 in March, April, May and June 2021. 183
Table 18 Top 20 words ranked by collocational strength of 'app' + R1 in July 2021. 185
Table 19 Top 6 words ranked by collocational strength of 'by the app' + L1 in July 2021. 187
Table 20 Ranking of the top 10 lexical items associated with each latent topic 211
Table 21 Frequency of active and passive constructions of *ChatGPT*. 228
Table 22 Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in November and December 2022. 229
Table 23 Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in November and December 2022. 234
Table 24 Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in January 2023. 236
Table 25 Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in January 2023. 241
Table 26 Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in February and March 2023. 242
Table 27 Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in February and March 2023. 247

# ACRONYMS

AI Artificial Intelligence

API Application Programming Interface

ChatGPT Chat Generative Pre-Trained Transformer

CDA Critical Discourse Analysis

CL Corpus Linguistics

CNN Convolutional Neural Networks

CSV Comma-Separated Values

DA Discourse Analysis

DL Deep Learning

EUACM Association for Computing Machinery Europe Council Policy Committee

FAT Fairness, Accountability and Transparency

GCSE General Certificate of Secondary Education

HCI Human-Computer Interaction

HRI Human-Robot Interaction

LDA Latent Dirichlet Allocation

LLM Large Language Model

LSS Latent Semantic Scaling

NHS National Health Service

NLP Natural Language Processing

RLHF Reinforcement Learning from Human Feedback

SAR Social Actor Representation

SAT Social Action Theory

STS Sociotechnical Systems

TF-IDF Term Frequency Inverse Document Frequency

UK United Kingdom

USACM Association for Computing Machinery United States Public Policy Council

VADER Valence Aware Dictionary and sEntiment Reasoner

XAI eXplainable Artificial Intelligence

# 1

# INTRODUCTION

## 1.1 BACKGROUND AND MOTIVATION

This PhD thesis will examine the scope for investigating agency, blame and trust in public-facing Autonomous Systems – in this case, defined as decision-making algorithms (Kochenderfer, Wheeler, and Wray, 2022) – using an interdisciplinary lens, which combines language analysis approaches from Computer Science and Linguistics to deliver nuanced insights into views expressed on social media. Public-facing Autonomous Systems aim to increase productivity and enable more efficient and informed decision-making (Royal Academy of Engineering, 2017). Examples, which have been used as case studies in this thesis, include the National Health Service (NHS) Covid-19 contact-tracing application, used for mitigating the spread of coronavirus in the United Kingdom (UK) (Heaton et al., 2024b), the Ofqual algorithm, used for automating UK Advanced Level results in 2020 (Heaton et al., 2023c), and Chat Generative Pre-Trained Transformer (ChatGPT), a text-generative large language model (Heaton et al., 2024a,c). These systems are of interest as they have had an impact on both a national and global scale and have generated conversation on social media (Kelly, 2021; Kretzschmar et al., 2020), despite working with little human supervision.

This thesis intends to address the concerns regarding the agency of these decision-making algorithms that have arisen recently, particularly when these have negative consequences (Bryson, 2020; Burrell, 2016). The perceived social agency of a decision-making algorithm can impact whether it is trusted, mistrusted, celebrated or blamed (Nowotny, 2021; Schoenherr and Thomson, 2024). In this context, de-

termining responsibility is challenging due to complexity and opacity (Holford, 2022; Selbst et al., 2019; Tsoukias, 2021), yet investigating trust in decision-making algorithms is critical for the development of future artificial intelligence technologies (Shahrdar, Menezes, and Nojoumian, 2019) as they become more relevant to our daily lives.

The Trustworthy Autonomous Systems Hub, a research organisation that enables the development of socially beneficial technologies, adopts the view by Devitt (2018) that decision-making algorithms must be trustworthy by design and perception. This means that developers and promoters should not only ensure that the algorithms are robust, reliable and transparent in their functioning but also foster a perception of trust among users, stakeholders and society at large (Devitt, 2018; Lee, 2018). Despite this, little exploration exists on how trust and blame are impacted by perceived social agency, responsibility and accountability of systems. Therefore, the agency remains unclear, necessitating further research to better understand its relevance in the context of decision-making algorithms.

A way of understanding the agency of an entity is to examine the relationship between social agency and grammatical agency. While social agency can be investigated in multiple ways, such as through interview or observation (Ahearn, 1999; Grillitsch, Rekers, and Sotarauta, 2021), grammatical agency -– or transitivity — can show whether an entity is presented actively, i.e. performing an action, or passively having an action performed onto them (Leslie, 1993). This is useful for analysing written social media texts as deconstructing the agency of these systems can shed light on the perceived power relations between entities and how these can ultimately indicate whether an algorithm is perceived as a social actor (Clark, 1998; Van Leeuwen, 2008). This may reveal if the system has a role in determining the outcome of a situation, and whether they are blamed or trusted, thus signalling what their perceived level of responsibility for such an outcome may be.

As previously mentioned, many have offered views about public-facing decision-making algorithms on social media sites such as Twitter (now X)[1], providing a large data set that an Application Programming Interface (API) can analyse in real-time (Agarwal et al., 2011; Gupta and Hewett, 2020; Kumar and Suresh, 2012). To analyse the views expressed on Twitter, popular Natural Language Processing (NLP) based computational tools, like sentiment analysis, topic modelling and emotion detection tools, are usually deployed. These are commonly used to undertake social media research due to the vast amounts of data that can easily be collected, using scraping techniques to analyse (Blei, Ng, and Jordan, 2003; Liu, 2010; Nikolenko, Koltcov, and Koltsova, 2017a; Vyas and Uma, 2018). These can be less intrusive, time efficient and cost-effective, as opposed to interviews or experiments (Rout et al., 2018). However, these popular NLP-based computational linguistic tools struggle to account for the discursive and conversational exchanges in which opinions on decision-making algorithms develop on social media. If aiming to understand how the presentation of grammatical and social agency impacts trust and blame (Kapidzic et al., 2019), these technical challenges should be taken into consideration. Other shortcomings include, but are not limited to, difficulty in the detection of negation, sarcasm and irony, and difficulty in interpretation (Jiang, Brubaker, and Fiesler, 2017; Maier et al., 2018; Stine, 2019).

Examples of how this has impacted research include instances where sentiment analysis tools fail to accurately capture the nuances of sentiment expressed in sarcastic or negated statements on social media platforms like Twitter (Maier et al., 2018). Similarly, topic modelling approaches may struggle to identify subtle shifts in discourse themes or the context-dependent nature of language usage, particularly in dynamic and rapidly evolving online conversations (Jiang, Brubaker, and Fiesler, 2017). Moreover, emotion detection algorithms

1 It is important to note that this thesis will use the term Twitter, rather than X, as data was collected and analysed at a time when this was the website's name.

may misinterpret the emotional tone of text, or offer categories that are leading to inaccuracies (Stine, 2019).

A new approach may overcome these shortcomings by combining popular NLP-based computational linguistic and sociolinguistic analyses. Here, Corpus Linguistics (CL) and Discourse Analysis (DA) may help culminate a corpus-driven approach to language exploration, underpinned by Social Actor Representation (SAR).

Firstly, CL involves analysing language data on a large scale, allowing for the comparison of multiple corpora to identify trends and patterns in texts (McEnery and Hardie, 2011). CL employs various analytical tasks, including collocation analysis, which examines the co-occurrence of words within a defined word span (Jaworska, 2017). Instead of relying solely on frequency, statistical significance measures are used to indicate lexical and grammatical associations, helping to identify meaningful collocations and themes (Baker, 2006; Mautner, 2007). Concordance lines in CL assist in showcasing the context surrounding a word of interest (Hoey, 2007). CL offers efficient analysis of social media datasets due to its ability to automatically scan large volumes of data for frequency patterns and keywords (Jaworska, 2017). Furthermore, similarly to NLP-based tools, it grants access to authentic texts and provides high processing speed (Tognini-Bonelli, 2001), as well as facilitating diachronic comparisons across corpora, aiding in the examination of lexical usage over time (Baker, 2010). Consequently, CL could be commonly used as a complementary approach to analysing language patterns in Twitter datasets.

DA complements CL by focusing on nuanced meanings and pragmatic interpretations in text analysis (Reeves, Kuper, and Hodges, 2008). While CL analyses language patterns, DA delves into implied meanings and contextual nuances, making it suitable for exploring trust and blame in decision-making algorithm discourse on Twitter. DA provides the opportunity to further scrutinise transitivity patterns, which can be explored through linguistic emphasis, manipulation or concealment (Leslie, 1993; Richardson, Mueller, and Pihlaja,

2021). Transitivity analysis, which examines agency in text, looks at the use of active and passive voice, revealing the language user's attitude and ideology (Leslie, 1993; Richardson, Mueller, and Pihlaja, 2021). Ultimately, this would aid the identification of agents, actions and responsibilities in discourse (Amoussou and Allagbe, 2018), enabling the exploration of trust and blame. This analysis of vocabulary and implicit information unveils ideological nuances, making transitivity and grammatical agency focal points of this PhD thesis (Clark, 1998; Goatly, 2007).

There are shortcomings to these approaches – such as CL results providing great evidence but limited explanation (Rose, 2017), the effort and time required to perform a successful discourse analysis, especially on a large data set (Wetherell and Potter, 1988), and the subjective nature of DA (Gill, 2000). However, there are ways in which they can complement the popular NLP-based computational linguistic tools to deliver insights into agency, blame and trust in this public discourse, specifically the analysis of the nuances of discourses that popular NLP-based methods may not account for.

Methodologically, the examination of each case study presented in this thesis will involve three parts. Firstly, popular NLP-based computational linguistic tools – topic modelling, sentiment analysis and emotion detection – will be used to gain an overview of the discourse in question, presenting these as trajectories, where areas of interest (particularly fluctuations that seem unexpected) can be highlighted for further exploration. Secondly, CL tools will be used to examine active and passive grammatical constructions in the discourse, with greater attention to be paid to the moments in the discourse highlighted from the initial analysis. Thirdly, DA, underpinned by SAR, will be used to further examine these active and passive presentations, identifying social actors and providing insights into how Twitter users trust or blame decision-making algorithms, and what the implications of these findings are.

To summarise, this thesis aims to investigate the views expressed about three different public-facing decision-making algorithms on Twitter. This is done through the use of NLP-based tools (namely sentiment analysis, topic modelling and emotion detection), CL and DA. This is underpinned by SAR, with a particular focus on how the relationship between grammatical and social agency can elucidate insights into whether these systems are portrayed as social actors, plus the implications for responsibility, accountability, trust and blame. Overall, this PhD thesis sets out to provide a comprehensive understanding of the dynamics surrounding social media discourse on decision-making algorithms. The findings presented here have the potential to be used by developers and promoters of these decision-making algorithms to uncover barriers to adoption and continued usage. This chapter outlines the thesis research question (RQ), sub-research questions (SRQs) and research objectives, whilst also commenting on the specialist areas that this PhD project is concerned with and the contributions it intends to make. Alongside this, this chapter also presents the peer-reviewed articles that have been published from this thesis, as well as the structure of the thesis.

## 1.2 RESEARCH QUESTION

Therefore, the main research question for this PhD project is:

> What insights into agency, trust and blame in the Twitter discourse surrounding decision-making algorithms can be achieved through combining language analysis approaches?

This research question will be explored in each of the three case studies, forming the sub-research questions. These are:

1. The 2020 A Level Calculation Algorithm
2. The NHS Covid-19 Contact-Tracing App

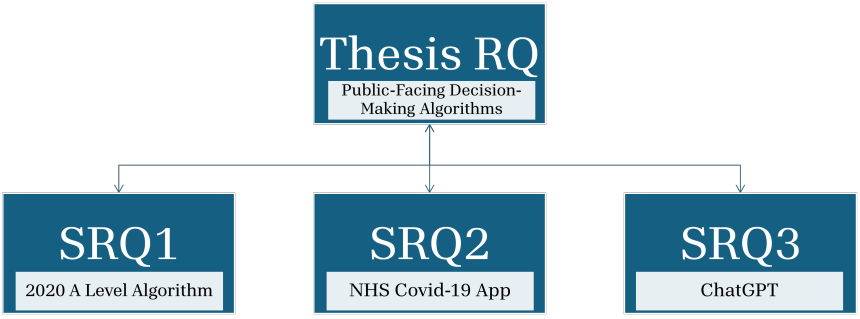


Figure 1: An illustration of the relationship between the overarching thesis research question (RQ) and the sub-research questions (SRQs).

3. ChatGPT

These three sub-research questions will feed into answering the overarching thesis research question. This is exemplified in Figure 1. The framing for each of these sub-research questions has been intentionally maintained throughout the thesis. This was to ensure the focal points of each study feed into answering the overarching PhD thesis research question.

## 1.3 RESEARCH OBJECTIVES

Four objectives will apply to each case study, which are:

a Demonstrate how Natural Language Processing tools (sentiment analysis, topic modelling and emotion detection) provide insight into public discourses surrounding decision-making algorithms.

b Demonstrate how Corpus Linguistics, particularly collocation, provides insight into public discourses surrounding the agency of decision-making algorithms.

c Demonstrate how Discourse Analysis provides insight into public discourses surrounding the agency, trust and blame of decision-making algorithms.

d Identify the strengths and limitations of using the three approaches to investigate public discourses surrounding decision-making algorithms.

## 1.4 PUBLICATION OF THE THESIS

Parts of this thesis have appeared in peer-reviewed publications. The publications derived from this thesis are listed below, indicating the specific chapters in which they appear.

Literature reviewed in Chapter 2 has appeared in:

- Heaton, Dan, Jérémie Clos, Elena Nichele, and Joel E Fischer (2023). "The Social Impact of Decision-Making Algorithms: Reviewing the Influence of Agency, Responsibility and Accountability on Trust and Blame." In: *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pp. 1–11.

Empirical work using NLP-based analyses in Chapters 4 and 5 has appeared in:

- Heaton, Dan, Jérémie Clos, Elena Nichele, and Joel Fischer (2023). "Critical reflections on three popular computational linguistic approaches to examine Twitter discourses." In: *PeerJ Computer Science* 9, e1211.

Empirical work using CL and DA in Chapter 4 has appeared in:

- Heaton, Dan, Elena Nichele, Jérémie Clos, and Joel E Fischer (2023). ""The algorithm will screw you": Blame, social actors and the 2020 A Level results algorithm on Twitter." In: *Plos One* 18.7, e0288662

Empirical work using CL and DA in Chapter 5 has appeared in:

- Heaton, Dan, Elena Nichele, Jérémie Clos, and Joel E. Fischer (2024). "Perceptions of the agency and responsibility of the NHS COVID-19 app on Twitter: critical discourse analysis." In: *Journal of Medical Internet Research* 26.1, e50388.

Empirical work using NLP-based analyses in Chapter 6 has appeared in:

- Heaton, Dan, Elena Nichele, Jérémie Clos, and Joel E. Fischer (2024). ""The ChatGPT bot is causing panic now — but it'll soon be as mundane a tool as Excel": analysing topics, sentiment and emotions relating to ChatGPT on Twitter." In: *Personal and Ubiquitous Computing*, pp. 1–20.

Empirical work using CL and DA in Chapter 6 has appeared in:

- Heaton, Dan, Elena Nichele, Jeremie Clos, and Joel E. Fischer (2024). ""ChatGPT says no": agency, trust, and blame in Twitter discourses after the launch of ChatGPT." In: *AI and Ethics*, pp. 1-23.

## 1.5 RESEARCH AREAS

This thesis is related to the following research areas:

- **Human-Computer Interaction:** This thesis makes significant contributions to the field of Human-Computer Interaction (HCI) by examining the intricate dynamics between humans and decision-making algorithms in online environments. Specifically, it enhances understanding of user interactions with decision-making algorithms by examining how these interactions shape public sentiment, agency attributions, and perceptions of trust and blame. Through empirical analyses of Twitter data, this thesis uncovers the nuanced ways in which users discuss and react to algorithmic decisions, highlighting the complexities inherent in human-algorithm interactions.

- **Computational Linguistics:** By integrating NLP tools and CL techniques, this thesis explores the complexities of language usage in online discourse. As computational linguistic techniques are usually employed to process and analyse vast language datasets and facilitate the recognition of linguistic patterns, this interdisciplinary approach relates to computational linguistics by enhancing the understanding of how decision-making algorithms are perceived in online contexts. Additionally, by addressing the challenges and constraints of current computational linguistic tools, this research lays the groundwork for future investigations to develop more sophisticated analytical frameworks for examining intricate language phenomena in digital settings.

- **Sociolinguistics:** By employing an interdisciplinary approach that integrates NLP, CL and DA, this thesis introduces a combined approach for studying linguistic phenomena in online environments. This combination of existing approaches enables researchers to analyse large-scale datasets from social media platforms like Twitter, offering new avenues for empirical linguistic research. Specifically, this research draws on sociolinguistic principles to examine how language is used to construct and negotiate social meaning in the context of human-algorithm interactions. Moreover, through the analyses of Twitter data, this thesis investigates how individuals attribute agency to decision-making algorithms, as potential social actors, in Twitter discourse. By examining linguistic features such as agency metaphors and personalisation of algorithms, the thesis sheds light on the complex ways in which language shapes perceptions of agency and responsibility and the impact this has on trust and blame. This analysis reveals how sociolinguistic variables like power dynamics are reflected in and shaped by online discourse surrounding algorithmic decision-making.

## 1.6 CONTRIBUTIONS

This thesis makes one main and subsidiary contribution, which are as follows:

1. **Insights into views expressed about decision-making algorithms online:** Using three case studies, this thesis examines the views expressed about public-facing decision-making algorithms. By scrutinising Twitter discussions surrounding these algorithms, this research presents a case that investigating agency, responsibility and accountability in these systems can ascertain implications for trust and blame. Thus, this contributes to the broader discourse on the societal implications of decision-making technologies by highlighting the role of social media views and attitudes.

2. **Complementary language analysis approach from existing methodologies:** The subsidiary contribution of this thesis lies in its combined approach to language analysis, which integrates and complements existing techniques. By combining NLP, CL and DA, this research offers a multifaceted examination of online discourse, allowing for a nuanced interpretation of linguistic patterns and communicative dynamics. This interdisciplinary lens not only enriches the findings contained in this PhD but also has the potential to be expanded for other researchers to study complex phenomena on social media.

Additionally, although not a core contribution of the thesis, the findings detailed in this research may be beneficial for policymakers, developers and promoters of decision-making algorithms, which may assist in overcoming barriers to adoption.

## 1.7 STRUCTURE OF THE THESIS

This thesis is structured into three main parts, with eight chapters in total. A summary of each part, and the chapters contained within them, is outlined below.

Part I, the background work to the PhD, begins with Chapter 2, which reviews key literature relevant to the research questions. Section 2.1 discusses the link between grammatical and social agency, while section 2.2 examines agency, responsibility, and accountability in prior research. Section 2.3 explores their impact on perceptions of trust and blame in decision-making algorithms. Additionally, three decision-making algorithm case studies – the 2020 A Level grade calculation algorithm, the NHS Covid-19 contact tracing app and ChatGPT – are introduced in section 2.4. Following this, Chapter 3 begins by reporting information about data collecting and handling in section 3.1. It also includes details on the overarching approach used in this PhD thesis in section 3.2. Finally, this chapter reports the specifics of the individual approaches used: NLP-based computational linguistic approaches (topic modelling, sentiment analysis and emotion detection) in section 3.3, CL in section 3.4 and DA in section 3.5.

Part II encompasses the empirical work of the thesis. Chapter 4 examines the first of the three Twitter case studies on the 2020 A Level results calculation algorithm and addresses SRQ1, along with objectives 1a, 1b, 1c and 1d. This is done in two parts: the NLP-based analysis in section 4.3 (which provides an overview of the computational findings in the discourse) and the CL and DA investigation in section 4.4 (which specifically examines the system's portrayal in Twitter discourses regarding its social agency). This same structure is replicated in Chapter 5, where the case study is the Covid-19 contact tracing app (addressing SRQ2 and objectives 2a, 2b, 2c and 2d) and Chapter 6, which is concerned with the ChatGPT case study (investigating SRQ3 and objectives 3a, 3b, 3c and 3d).

Finally, Part III synthesises the findings of the PhD. In Chapter 7, the main findings from the three empirical results chapters are discussed in sections 4.5, 5.5 and 6.5 respectively. This is followed by the answering of the overarching research question in section 7.3, where implications for agency, trust and blame and discussed. Additionally, section 7.4 outlines the implications for research and limitations of the thesis are explored in section 7.5. At the end of the thesis, Chapter 8 summarises the main findings and contributions of the thesis in sections 8.1 and 8.2 respectively. The thesis ends by outlining avenues for future work in this area in section 8.3.

# Part I

# BACKGROUND AND APPROACH

# 2

# LITERATURE REVIEW

This chapter of the thesis introduces key literature motivating the research questions. Specifically, section 2.1 details the link between grammatical and social agency and how this can be used to establish social actors. Also, in order to make the case for the importance of examining the social agency of decision-making algorithms in this thesis, it was important to examine how the concepts of agency, responsibility and accountability had been viewed in prior research in section 2.2. Further to this, the impact that agency, responsibility and accountability has on perceptions of trust and blame in decision-making algorithms is explored in section 2.3.

Additionally, this chapter introduces the three decision-making algorithm case studies: the NHS Covid-19 contact tracing application, the 2020 A Level grade calculation algorithm and ChatGPT. These reviews begin in section 2.4. A detailed literature survey for the 2020 A Level algorithm can be found in subsection 2.4.1, the NHS Covid-19 app can be found in subsection 2.4.2, and ChatGPT can be found in subsection 2.4.3. These subsections provide background to each of the systems and the social impact they have had, in precursor to their individual analysis chapters in Chapters 4, 5 and 6 respectively.

## 2.1 GRAMMATICAL AND SOCIAL AGENCY

This section focuses on the relationship between grammatical and social agency, which is used to underpin a substantial amount of the analysis in this thesis. This concept is related to the domain of sociolinguistics, which examines how social factors influence language use and variation in different contexts (Holmes and Wilson, 2022). It

investigates the relationship between language and society, studying how variables like social class, gender, ethnicity and age affect linguistic choices and patterns. Hudson (1996) emphasises that sociolinguistics explores both how language affects society and how society affects language, which is pertinent for this type of analytical exploration.

Agency is conceptualised as the capacity of individuals to act independently and make their own free choices within social structures (Giddens, 1984). In this sense, it represents the power of social actors to impact their circumstances, though this ability is influenced by social, cultural and material constraints. Emirbayer and Mische (1998) frame agency as a temporally embedded process informed by past experiences while oriented toward the future and engaged with the present. . For the purpose of this thesis, when defining agency, Gallagher (2000) stated that it is a clear feeling of control and suggests it impacts human consciousness. Silver et al. (2021) said that a sense of agency also encompassed the responsibility felt due to actions undertaken and the effects they have. Therefore, social agency could be uncovered by examining grammatical agency (Clark, 1998; Leslie, 1993; Richardson, Mueller, and Pihlaja, 2021).

From a sociolinguistic perspective, agency is not simply an individualistic characteristic but is also shaped by the social contexts in which individuals interact (Ahearn, 1999; Duranti, 2008). This means that factors like social norms, power dynamics, and access to resources can influence an individual's capacity to act (Norton and Toohey, 2011). Therefore, to understand how agency is constructed and negotiated, it is essential to examine how individuals use language within these social contexts (Duranti, 2008).

Grammatically, Leslie (1993) defined an agent as an entity with an internal source of energy through which it exerted force supposedly to carry out the activities referred to in a text. Expanding on this, Richardson, Mueller, and Pihlaja (2021) stated that agency in linguistics is often explored by examining how it was emphasised,

manipulated, or concealed. As such, transitivity analysis – the examination of agency in the text – looked at the use of active and passive voice or nominalisation, where verbs were the word class converted to nouns. Accordingly, choices revealed the attitude and ideology of the language user or the perceived agent (Oktar, 2001). Additionally, research showed that passive constructions tended to remove agency from the subject or dilute its impact (Comrie, 1977). Especially when the subject was absent from the clause, implied responsibility shifted (Clark, 1998). Arguably, this referred to the decision-making power, which is investigated in this thesis through the exploration of the decision-making algorithms and the implications for trust and blame.

Alternatively, the agency can be conveyed through lexical choices. For instance, Morris et al. (2007) suggest that an acceding trajectory evokes the impression of high animacy, which would be caused by enduring internal property, i.e., the volitional action (e.g., "the NASDAQ fought its way upward"). On the other hand, a descending trajectory suggests inanimacy, as a result of lack of external forces (e.g., "stocks drifted higher"). This phenomenon, called agency metaphor constituted the focus of this analysis, as well as transitivity, as they both communicate the capacity or potential to finalise decisions. In this sense, examining the portrayal of decision-making algorithms online can unveil attitudes towards these systems. Through the use of terms like 'decided', 'chose' or 'determined' algorithms may be imbued with a sense of agency, suggesting an active role in decision-making processes.

### 2.1.1 *Social Actor Representation*

In order to successfully analyse the relationship between grammatical and social agency, the theoretical framework of Social Actor Representation (SAR), which is drawn from Social Action Theory (SAT) can be used to underpin analysis. According to SAT, people create society,

institutions and structures (Weber, 1978), hence examining social actions can provide an explanation for human behaviour and societal change (Engeström et al., 1999), including the perceptions of system users that this thesis focuses on.

More specifically, SAR looks at how grammatical structures convey social agency, for example, active or passive constructions and transitivity structures can be employed to communicate who social actors are in a discourse of interest (Van Leeuwen, 2008). Similarly, syntactical features, sentence structures and verbs within tweets will give an indication of how users perceive each system.

Multiple concepts can be considered as key when it comes to socio-semantic categories for analysing social actors (Van Leeuwen, 2008). Among them, removing grammatical agents is called *excluding*. Conversely, when clues are left in as to who the agent is, this is called *backgrounding*. Also, actors can be *personalised* through word choices pertaining to the semantic nature of being 'human' or *impersonalised*. Moreover, examining agency metaphor, previously outlined by Morris et al. (2007), can signal further personification of non-human entities. All these aspects are especially relevant to this thesis because they imply human-like perception, possibly indicating whether responsibility for consequences is attributed as a result of social agency.

At the same time, social actors could be a group of people (*genericised*), or represented as single individuals (*specified*). In this sense, *indetermination* is when social actors are not specified (like 'someone'), whereas *determination* is when their identity is (made) known. All of these representation structures play a role in indicating the social and power dynamics within discourse, as shown in Twitter case studies that used SAR (Bernard, 2018; Fadanelli, Dal Pozzo, and Fin, 2020; McGlashan, 2020). For example, Bernard (2018) used elements of DA and Van Leeuwen's socio-semantic framework based on SAR to study the representation of social actors in the business discourses of two South African mining companies. She found that companies draw on a fixed set of linguistic devices and strategies when representing

higher- and lower-wage employees respectively. Therefore, these linguistic representations have an important role in maintaining relationships of power, dominance and social inequality in the South African mining industry. Other work includes Fadanelli, Dal Pozzo, and Fin (2020), who used DA and SAR to investigate Brazilian president Jair Bolsonaro's tweets, finding that they fulfill an important function of ideological rapprochement between supporters and Bolsonaro, which has influenced his popularity. Additionally, McGlashan (2020) used a corpus-based DA approach, including the analysis of collocates underpinned by SAR, to investigate the language used by followers of the Football Lads Alliance — a protest group who say they are 'against all extremism'. This exemplifies that this theoretical underpinning is used successfully in many similar studies.

By analysing all these characteristics in the Twitter discourse collected, it is the intention to identify common presentations of decision-making algorithms, ultimately displaying how power relations are communicated in real-life data dealing with algorithmic-operated decisions, even when mechanisms are not fully clear. After establishing these, similar semantically-related thematic groups could be identified. Examples of this being done successfully include Razis, Anagnostopoulos, and Saloun (2016), who applied thematic categories automatically depending on the appearance in their dataset, and Kitishat, Al Kayed, and Al-Ajalein (2020), who organised collocates into two thematic categories based on their analysis of the Syrian refugee crisis in Jordanian newspapers. Overall, using thematic categories has the potential to aid the analysis of the presentation and perceptions of various systems in this thesis.

## 2.2 REVIEWING AGENCY, RESPONSIBILITY AND ACCOUNTABILITY

This section of the literature review will focus on agency, accountability and responsibility. For the purposes of this exploration, agency refers to the ability of an entity to act on its own, without being influenced by external factors (Bandura, 2001; Giddens, 1986; Zimmerman, 2000). Responsibility and accountability are related concepts, but they have different meanings. Responsibility refers to the obligation to take action or make decisions based on one's role or position (Baumeister and Leary, 2017; Pettit, 2001), while accountability refers to the responsibility of an individual or entity for the consequences of their actions (Bovens, 2007; Mulgan, 2000). Each of these concepts will be explored in relation to decision-making algorithms.

### 2.2.1 *Agency and Decision-Making Algorithms*

Agency, for the purpose of this exploration, refers to the capacity of individuals or groups to act intentionally, make choices and exert influence over their environment (Bandura, 2001; Giddens, 1986; Zimmerman, 2000). According to Bandura (2001), agency involves a range of cognitive, behavioural and motivational processes that enable individuals to set goals, develop plans and execute actions to achieve desired outcomes. Agency is also influenced by social and cultural contexts, as individuals' beliefs, values and norms shape their understanding of the available options and the extent of their freedom to act (Arnett, 2015; Marková, 2003). As Zimmerman (2000) argues, agency is a dynamic and interactive process that requires an ongoing negotiation between individuals and their environment, as individuals adjust their strategies and goals in response to changing circumstances. Therefore, agency is not a fixed or innate characteristic of individuals, but rather a complex and dynamic process that is shaped by multiple

factors, including personal attributes, social and cultural contexts and environmental constraints. One issue related to agency is the degree of autonomy decision-making algorithms possess.

Algorithms have a degree of autonomy and agency, especially when they can learn and adapt from data inputs (Bryson, 2020). In particular, Bryson argues that algorithmic autonomy is not an all-or-nothing concept, but rather exists along a spectrum. At one end of the spectrum are algorithms that are highly deterministic and programmed to follow specific rules and decision-making processes (such as sorting algorithms or expert systems in medical diagnosis), while at the other end are algorithms that are able to learn and adapt from data inputs, making decisions that are not explicitly programmed or predetermined by humans (like decision trees, neural networks or self-learning systems like the computer game AlphaGo). Additionally, she highlights the potential benefits and risks associated with algorithmic autonomy. On one hand, autonomous algorithms can help improve efficiency and accuracy in a wide range of fields, from medical diagnosis to self-driving cars. On the other hand, if not properly designed and regulated, autonomous algorithms can pose significant risks, such as perpetuating bias or making decisions that harm people or society.

However, Floridi et al. (2018) argue that algorithms are not autonomous and that their outputs are the result of human design choices. In their work, they contend that algorithms cannot be considered autonomous as they are created by humans and, therefore, influenced by human biases, values and intentions, such as image recognition and recruitment algorithms. Although algorithms can be programmed to learn and adapt from data inputs, human designers are responsible for selecting the data to be used, analysing it and deciding what to do with the outputs. The authors propose that algorithms should be used to enhance human capabilities rather than replace them and urge us to consider the ethical implications of using algorithms in decision-making. In particular, they highlight the

potential impact of algorithmic decisions on areas like healthcare, criminal justice and finance. Ultimately, Floridi et al. (2018) advocate for a human-centered approach to algorithm development and usage, which prioritises the common good and human values. Therefore, this suggests that agency should be attributed to human designers rather than algorithms.

The increasing use of decision-making algorithms in various domains has raised concerns about their impact on social agency, particularly in terms of how they affect the decision-making processes of individuals, the accountability of the systems and those affected by them. Some scholars have argued that decision-making algorithms may impede social agency by replacing human judgement and decision-making with automated processes. For example, Burrell (2016) highlights the problem of opacity in machine learning algorithms. She argues that opaque algorithms, which are those that are difficult or impossible to interpret, can lead to blurred agency and transparency, and may perpetuate bias and discrimination. Additionally, the analysis by Pasquale (2015) on the rise of algorithms and their impact on society, specifically in the realms of finance and information, is relevant to the concept of agency. He argues that the opacity of these algorithms removes agency from the public, as decisions are being made without their input or understanding. This lack of transparency also limits the agency of those who are impacted by algorithmic decisions, as they have little recourse to challenge or contest those decisions.

However, others have emphasised the need to consider how algorithms can be designed to support rather than undermine social agency and how regulation can be ensured in algorithmic decision-making. Diakopoulos (2016) examined accountability in algorithmic decision making and argues that transparency and explainability are necessary for accountability to co-exist. He suggests that a combination of technical and social solutions are needed to address this challenge to regulate decision-making algorithms. Moreover, he stresses

the importance of involving affected communities and stakeholders in the design and implementation of algorithmic decision-making systems. This is similar to the findings of Selbst et al. (2019), who suggest that incorporating diverse perspectives and values into the design and implementation of sociotechnical systems can help promote fairness, which increases the agency of individuals and groups impacted by these systems. Additionally, the level of detail in which data is collected and analysed impacts fairness, which can affect the agency of individuals and groups if they are not fairly represented or impacted by the system. For instance, if a facial recognition algorithm disproportionately misidentifies certain demographic groups due to biases in the training data, individuals from those groups may experience reduced agency in situations where the technology is used for security or access control (Schuetz, 2021).

Above all, one of the key issues in the literature on social agency and decision-making algorithms is the role of transparency and explainability in fostering agency. Some scholars have argued that transparency is necessary to enable individuals to understand the decision-making process and to challenge algorithmic decisions if necessary (Gillespie, 2014; Zarsky, 2016). However, others have noted that the complexity of algorithms and the lack of transparency in their development and implementation may hinder accountability and undermine agency (Burrell, 2016; Pasquale, 2015).

Another important aspect of the literature on social agency and decision-making algorithms is the need for interdisciplinary approaches to address the ethical and social implications of algorithmic decision-making. Scholars from computer science, philosophy, social sciences and humanities have emphasised the importance of considering the broader societal and political implications of algorithms, beyond their immediate technical functionality (Floridi and Cowls, 2022; Mittelstadt et al., 2016). Such interdisciplinary approaches can help to develop more nuanced understandings of the relationship between social agency, decision-making algorithms and accountability,

and to identify strategies for ensuring that these systems are developed and implemented in ways that support rather than undermine social agency.

To address some of the challenges, eXplainable Artificial Intelligence (XAI) has been proposed as a solution for enhancing transparency and accountability in decision-making algorithms. XAI involves designing algorithms that can provide explanations for their outputs in a human-understandable format (Gunning et al., 2019). By doing so, developers, operators and regulators can gain deeper insights into how these algorithms work and why they produce certain outputs, enhancing accountability and responsibility (Selbst et al., 2019).

The exploration of agency in the context of decision-making algorithms reveals complexities involving individual and contextual factors shaping decision-making and influence. Within this context, debates on algorithmic autonomy and social agency intersect. While some argue algorithms possess autonomy, Bryson (2020) notes the spectrum of algorithmic decision-making from rule-bound to adaptive systems, emphasising benefits – like predictability and reliability for rule-bound and flexibility for adaptive – and risks – such as inflexibility for rule bound and bias and difficulty in interpretation for adaptive – of both. However, Floridi et al. (2018) contend algorithms lack autonomy as they are crafted by humans, reflecting human biases and values in data representation. They advocate for human-centered approaches in algorithm usage, attributing agency to human designers. Concerns arise regarding algorithmic impact on social agency, with opacity hindering transparency and accountability (Burrell, 2016; Pasquale, 2015). Conversely, others stress the need for transparency and diverse stakeholder involvement to enhance accountability and fairness in algorithmic decision-making, thereby supporting social agency (Diakopoulos, 2016; Selbst et al., 2019). The discourse emphasises the role of transparency, explainability and interdisciplinary collaboration, where XAI emerges as a potential solu-

tion by offering human-understandable explanations for algorithmic outputs, fostering accountability and responsibility.

### 2.2.2 *Responsibility and Decision-Making Algorithms*

Responsibility is defined as the moral and social obligations of individuals or groups to act in accordance with certain standards or norms (Baumeister and Leary, 2017; Pettit, 2001). Responsibility can be understood as a combination of two key elements: attribution and accountability. Attribution refers to the recognition of one's role in a particular situation or outcome, while accountability refers to the expectation that one will take action to address or repair any harm caused (Bovens, Bovens, et al., 1998; Pettit, 2001). Responsibility is also closely linked to agency, as individuals' capacity to act intentionally and make choices is a precondition for holding them responsible for their actions (Wallace, 1998). However, the extent to which individuals are held responsible for their actions is also influenced by various social and cultural factors, such as norms, values and power dynamics (Archer, 2000; Miller, 2001). Therefore, responsibility is not an absolute or fixed concept, but rather a dynamic and context-dependent process that is shaped by a range of individual and social factors. Because of this, the responsibility that decision-making algorithms possess has to be viewed from several perspectives.

Contributions offering frameworks for understanding include Tsoukias (2021), who examined the social responsibility of algorithms in society. They highlight the long-standing use of autonomous artefacts and categorise the impact of their use on data collection, manipulation, recommendation and decision-making. The framework identifies challenges for decision analysts, researchers and practitioners and emphasises the need for a community effort in addressing the ethical implications of algorithmic decision-making.

Additionally, others argue that the drive towards responsible adoption of automated decision-making systems fails to take into account the complexities of human judgement and the relevance of the human ability to discern ethical cues and actions. For example, through examining the representational limitations of Artificial Intelligence (AI) systems in discerning relevant cues and actions critical to ethical deliberations, Holford (2022) contrasts them to the twin-perspectives of pragmatism and phenomenology that provide lenses through which to unpack the human process of ethical deliberation. He concluded that a socio-technical system can only meet its moral responsibilities by attributing it directly onto the human decision maker's shoulders with full human meaningful control.

There have also been studies that have paid specific attention to social responsibility. Social responsibility refers to the ethical and moral obligations of organisations to act in the best interests of society (Carroll, 1979) and decision-making algorithms must also uphold these principles, particularly given the potential biases and discrimination that may result from their use (Diakopoulos, 2016). As a result, there has been a growing interest in developing frameworks for ethical decision-making and the responsible use of algorithms.

One such framework is the Fairness, Accountability and Transparency (FAT) framework proposed by Mittelstadt et al. (2016). This framework emphasises the importance of incorporating ethical principles into the design and implementation of algorithms, with a focus on ensuring fairness, accountability and transparency. Specifically, they suggest that algorithms must be designed to avoid perpetuating or amplifying biases and discrimination, and that users must be able to understand how the algorithms work and how they arrived at their decisions. This is essential to promote trust, mitigate potential harm and uphold ethical standards in algorithmic decision-making.

Similarly, Selbst et al. (2019) proposed the Sociotechnical Systems (STS) framework, which considers the interplay between technology and social systems in promoting ethical decision-making. This frame-

work emphasises the importance of incorporating diverse perspectives and values into the design and implementation of sociotechnical systems to promote fairness and accountability. Specifically, the authors suggest that systems must be designed to reflect the values and needs of all stakeholders, ensuring system design processes are transparent and inclusive, much like the work by Mittelstadt et al. (2016). Therefore, the development of frameworks for ethical decision-making and the responsible use of algorithms reflects a growing recognition of the need for decision-making algorithms to be responsible to society.

Overall, these frameworks highlight the importance of incorporating ethical principles and values into the design and implementation of algorithms. They promote fairness and transparency and ensure that these technologies are used in a responsible and socially beneficial manner.

### 2.2.3 *Accountability and Decision-Making Algorithms*

At its core, accountability refers to the extent to which individuals or organisations are held responsible for their actions or decisions, and the consequences that result from those actions or decisions (Bovens, 2007). While accountability is often associated with concepts such as transparency and control, it also has broader implications related to trust, legitimacy and democratic governance (Koppell, 2005; Mulgan, 2000). Accountability can be viewed as a mechanism for ensuring individuals or organisations are answerable to those who are affected by their actions or decisions and that they are held responsible for the outcomes they produce (Bovens, 2007; Mulgan, 2000). This can include various forms of accountability: legal, political and social (Schedler, Diamond, and Plattner, 1999). In practice, accountability is often implemented through mechanisms such as performance monitoring, evaluation and auditing; it is seen as a key factor in promoting

effective and responsible governance (Koppell, 2005; Mulgan, 2000). Nevertheless, this is not clearly applicable to algorithms.

Debate exists regarding who is accountable when algorithms do not achieve the expected outcomes. One of the Association for Computing Machinery United States Public Policy Council (USACM) and Association for Computing Machinery Europe Council Policy Committee (EUACM) principles for algorithmic fairness is accountability, which ensures those who deploy an algorithm cannot eschew responsibility for its actions, therefore not deflecting responsibility to an automated system (Garfinkel et al., 2017). Despite this, research suggests many individuals and groups do shift responsibility from humans if an algorithm is involved with a decision-making process. For example, Turton (2017) stated that Google and Meta deflect responsibility onto their social media algorithms despite being in control of their own code. This reiterates that the accountability is still debatable when it comes to algorithms.

Feier, Gogoll, and Uhl (2021) looked at whether an agent is systematically judged differently when the agent is artificial rather than human. They found decision-makers can actually rid themselves of guilt more easily by delegating to machines than by delegating to other people, thus showing the availability of artificial agents could provide stronger incentives for decision makers to delegate morally sensitive decisions. Therefore, it could be interpreted that decision-making algorithms are used to deflect accountability from human decision-makers.

Similarly, Bucher (2017) coined the term 'algorithmic imaginary' - how one imagines, perceives and experiences algorithms and what these imaginations make possible. This has been applied in many contexts – most suitably for this strand of research by Benjamin (2022), who recently studied the response to the 2020 A Level algorithm on Twitter through the examination of the "fuck the algorithm" chant as an imaginary of resistance to confront power in sociotechnical systems. Their analysis argued that this chant made algorithms more

visible to the public and prompted questions about social algorithms that shape the lives of many, every day.

The study by Burrell (2016), discussed earlier, is relevant to accountability also. To address issues with agency, she suggests designers and developers of machine learning algorithms need to take steps to increase transparency, including developing tools for auditing algorithms and making their workings more transparent to users.

Overall, these studies demonstrate the existing work on how humans attribute responsibility to decision-making algorithms socially and could pave the way for further investigation into how these algorithms could influence everyday life when accountability is placed solely on them. The seeming removal of autonomy and accountability from the human(s) that devise these algorithms could be replicated in online discourses or perceived in another way. In order for creators and promoters of these systems to be successful, having accurate insights into the current perceptions of these algorithms is important so potential misleading information can be combated.

### 2.2.4 *Section Summary*

This section has examined the intricate dynamics of agency, responsibility and accountability within decision-making algorithms. The exploration of agency underscores the debate on algorithmic autonomy, where some advocate for autonomous algorithms, while others contend that their outputs reflect human biases, raising concerns about opacity hindering transparency and accountability (Bandura, 2001; Bryson, 2020; Burrell, 2016; Diakopoulos, 2016; Floridi et al., 2018). Conversely, efforts to enhance fairness and accountability in algorithmic design through interdisciplinary collaboration XAI aim to support societal agency (Gunning et al., 2019; Selbst et al., 2019).

Regarding responsibility, research has investigated the moral obligations within decision-making, debating the role of human judg-

ment versus AI limitations (Baumeister and Leary, 2017; Holford, 2022). Frameworks like FAT and STS have highlighted the necessity of ethical principles and diverse perspectives in algorithm design for societal benefit (Mittelstadt et al., 2016; Selbst et al., 2019). Accountability intertwines with algorithmic decision-making, discussing the attribution of responsibility and the challenges in transparency for accountability, including the concept of the 'algorithmic imaginary' and its impact on public perceptions (Bucher, 2017; Feier, Gogoll, and Uhl, 2021; Garfinkel et al., 2017).

To summarise, this exploration has showcased the perceptions of balance between human agency and algorithmic autonomy, stressing the crucial roles of transparency, stakeholder involvement and interdisciplinary collaboration in ethically designing algorithms that support societal agency. Nevertheless, challenges persist in attributing responsibility and ensuring accountability in algorithmic decision-making, necessitating further examination of societal perceptions and the ethical implications surrounding these systems, especially when it comes to discourses online.

## 2.3 INFLUENCING TRUST AND BLAME

The concepts of trust and blame in the context of decision-making algorithms are becoming staple topics of autonomous system literature. Trust is a crucial factor in ensuring ethical and responsible use of algorithms as it is essential for users to trust that algorithms produce accurate and reliable outcomes (Mayer, Davis, and Schoorman, 1995). Developers, operators or other actors involved in the development and use of these algorithms are often held accountable when blame is assigned (Floridi et al., 2018). While trust and blame may seem contradictory, they can coexist (Baumeister et al., 2001). For example, when users have confidence in the overall integrity of the algorithms while still holding developers and operators accountable for negative

outcomes. The following sub-sections explore how agency, responsibility and accountability can impact the trust in – and the blaming of – decision-making algorithms in existing literature.

### 2.3.1 *Trust and Decision-Making Algorithms*

Trust, as a multi-dimensional construct, has been extensively studied in multiple fields, including sociology (Tan and Sutherland, 2004; Yousafzai, Pallister, and Foxall, 2009), psychology (Rousseau et al., 1998; Tan and Sutherland, 2004), economics (Lee et al., 2021; Zhang, Cui, and Wang, 2013) and management (Cho and Park, 2011; Radomska et al., 2019). Scholars have examined trust in diverse contexts, such as interpersonal relationships (Chang et al., 2016; Klein et al., 2019), organisational settings (Dietz, Gillespie, and Chao, 2010; Fang et al., 2008) and cross-cultural interactions (Dirks and Ferrin, 2002; Mayer, Davis, and Schoorman, 1995), as well as decision-making algorithms (Alaieri and Vellino, 2016; Bonnefon, Shariff, and Rahwan, 2016; Lyons et al., 2017; Shahrdad and Amirani, 2018). As a dynamic process, trust is influenced by individual differences, context and various factors like power dynamics and external events (Meyerson, Weick, Kramer, et al., 1996), which is particularly crucial for the analysis of agency in this thesis. The concept of trust has many different definitions and interpretations and there is currently no uniformed or universally agreed definition (Adams et al., 2003). For this thesis, the epistemological stance undertaken will be that trust is a socially constructed concept created within an individual internally (Weber, Weber, and Carter, 2003) as a result of interaction and experience (Green, 2007). The process of building and maintaining trust involves communication, mutual exchange and negotiation (Cook and Hegtvedt, 1983). Furthermore, trust is shaped by an individual's experiences, cultural background and context and is considered a socially constructed concept (Giddens, 2007; Luhmann, 1979). This person-

centered outlook is particularly relevant as this thesis specifically examines the human views expressed about decision-making algorithms.

The successful adoption and deployment of decision-making algorithms depend on the level of trust users have in them, which is influenced by the concepts of agency, responsibility and blame. For example, studies have shown that users are more likely to trust algorithms that operate autonomously and produce reliable outcomes (Bonnefon, Shariff, and Rahwan, 2016; Floridi et al., 2018). If algorithms are perceived as being influenced by external factors, such as human biases, their trustworthiness may be questioned (Bonnefon, Shariff, and Rahwan, 2016), instead.

The fulfilment of responsibilities and accountability of actors involved in the development and use of decision-making algorithms also affects trust. When developers and operators fulfill their obligations and are held accountable for their actions, users may have greater trust in the overall integrity and reliability of the algorithms. In contrast, failure to fulfill these obligations and responsibilities may lead to blame being assigned and may reduce users' trust in the algorithms (Bonnefon, Shariff, and Rahwan, 2016).

Specifically to decision-making algorithms, Shahrdad and Amirani (2018) examined trust in light of the exponential growth of the use of these systems in daily life and reviewed existing literature on the topic. They found prior studies indicated trust towards fully autonomous and semi-autonomous systems — such as home service robots and flight management systems — is low (Madhavan and Wiegmann, 2007; Muir, 1987). They concluded that managing trust affects the development of future acceptance and adoption of these systems.

Moreover, Lyons et al. (2017) studied the verification and validation of similar decision-making algorithms and created an approach to certify trust in them. They argue 'transparency facets' — an established communication channel between the designer, tester and user

–– enable the user to understand the goals of the system to verify its trustworthiness. Similarly, Kwiatkowska and Lahijanian (2016) called for the channels of communication to be re-examined to improve the perception of trustworthiness of these decision-making algorithms. A necessity to advance the role of social trust within HCI and Human-Robot Interaction (HRI) underpinned this theory. This accounts for competence, disposition, dependence and fulfilment. Ultimately, although both studies were inconclusive and called for more investigation, they highlight the importance of user feedback in the design and evaluation stages of decision-making algorithms creation and curation.

Additionally, Alaieri and Vellino (2016) argue that the ethical principles according to which the machines were programmed must be transparent and predictable. If the autonomy and self-learning abilities of some robots made their decisions non-predictable and difficult to explain, human trust would decrease. Thus, they call for further research and development in this area to ensure the ethical justification and trustworthiness of autonomous systems.

Despite the extensive research on trust in various contexts, there are still gaps and limitations in the literature related to how agency, responsibility and accountability impact trust in decision-making algorithms. Firstly, there is a lack of consensus on the definition of trust, which may lead to different interpretations and inconsistent findings. Moreover, trust is a complex and dynamic concept, influenced by many individual and contextual factors that may not be fully understood or controlled. Although this thesis cannot provide an exhaustive account of all aspects of trust, the examination of how agency, responsibility and accountability impact trust through transitivity analysis is under-explored in online discourses surrounding decision-making algorithms, potentially unveiling new perspectives for those developing and promoting such systems.

Additionally, the studies reviewed in this thesis provide some insight into how trust is impacted by agency, responsibility and account-

ability, but more research is needed to fully understand the mechanisms involved and how to design and evaluate decision-making algorithms that foster trust. Specifically, there is a need for further investigation into the role of communication, user feedback and ethical principles in building and maintaining trust in these systems. This research gap can begin to be filled by examining how trust in these systems is expressed via grammatical and social agency online. Moreover, the ethical implications of decision-making algorithms, especially in relation to autonomy and self-learning abilities, require further attention to ensure their trustworthiness.

### 2.3.2 *Blame and Decision-Making Algorithms*

Blame can be defined as the assignment of responsibility for a particular event or outcome, often with a negative connotation (Coates and Tognazzini, 2013). Blame can be directed towards individuals or groups and can have various functions, such as expressing disapproval, holding individuals accountable or seeking to assign causality (Baumeister, 1996; Coates and Tognazzini, 2013). Blame is often accompanied by moral judgments as it involves the evaluation of individuals' actions or omissions against certain norms or standards (Rozin, Markwith, and Stoess, 1997; Tetlock, 1992). However, the process of blaming is also influenced by various cognitive and motivational biases, such as the fundamental attribution error, which involves overestimating the role of dispositional factors and underestimating situational factors in explaining behaviour (Gilbert and Malone, 1995; Ross, 1977). Therefore, blame is a complex and multifaceted process that involves a range of cognitive, emotional and social factors and can have significant implications for individuals' self-esteem, social relationships and sense of justice. For this thesis, the concept of blame is particularly important as these public-facing systems depend on societal acceptance. Therefore, blame ultimately

impacts their integration into various domains, from healthcare to education and beyond.

The relationship between agency, responsibility and accountability in decision-making algorithms and their impact on blame assignment has been explored. Some scholars have noted high levels of agency in algorithms can lead to a reduction in accountability and make it difficult to assign blame for negative outcomes. For example, Mittelstadt et al. (2016) found algorithms with a high degree of agency can result in a 'responsibility gap', where neither the developers nor the algorithms are fully responsible for the outcomes produced. This study also emphasised the need to determine who should be held responsible for negative outcomes in decision-making algorithms (Mittelstadt et al., 2016). While developers and operators are typically seen as the most obvious targets of blame, others argue blame can be shared among all actors involved in the development and use of these algorithms.

Similarly, the fulfillment of ethical responsibilities by developers and operators is also an important aspect of blame. Jobin, Ienca, and Vayena (2019) found that fulfilling ethical responsibilities can increase user trust in algorithms, while failure to do so can lead to decreased trust and increased blame assignment. Similarly, Whittlestone et al. (2019) argue that ensuring ethical use of algorithms by fulfilling responsibilities is crucial for avoiding blame in technology and negative societal impacts, such as erosion of privacy, perpetuation of biases and exacerbation of social inequalities. This has been explored in the contexts of race and gender by Devlin (2023), who acknowledged that there is no 'quick fix', but that we must continually question who is creating and benefitting from such technologies.

Additionally, accountability is another key factor in the assignment of blame in decision-making algorithms. The accountability of developers and operators for the outcomes produced by algorithms they develop and use is necessary to ensure the ethical and responsible use of technology. Moreover, Taddeo and Floridi (2018) argue that

accountability is essential for holding developers and operators responsible for the ethical use of algorithms and building user trust in technology. However, blame is complex and depends on factors such as the degree of intention behind the actions, as noted by Coeckelbergh (2020b), which may mean that attributing blame solely to developers and operators may overlook systemic issues inherent in the design, deployment and regulation of decision-making algorithms.

The increasing use of decision-making algorithms has led scholars to grapple with the question of how to assign blame in cases where these algorithms produce negative outcomes. Jobin, Ienca, and Vayena (2019) found algorithms themselves can also be viewed as objects of blame, given they may perpetuate biases or produce negative outcomes due to the design of the system. However, they note that assigning blame can be challenging due to the complexity and opacity of these systems. On the other hand, Burrell (2016) argues that assigning blame is still important to ensure that decision-making algorithms are used ethically and responsibly. Nonetheless, the assignment of blame is complicated by the involvement of multiple actors, including developers, operators, data providers, regulators and the algorithms themselves (Mittelstadt et al., 2016). Therefore, this links to the aims of this thesis because understanding how blame is attributed, in cases involving decision-making algorithms, is crucial for assessing the ethical implications of their use. Moreover, it is useful for developing strategies to enhance accountability and trust in these systems.

Some research has been done to try and address these aforementioned challenges. Selbst et al. (2019) proposed incorporating fairness and abstraction in sociotechnical systems to ensure ethical use of decision-making algorithms, while Barocas and Selbst (2016) discussed the concept of disparate impact in big data . Additionally, Gunning et al. (2019) emphasised the importance of XAI to ensure transparency and accountability in decision-making algorithms. As this investigation unfolds, it will be interesting to see whether these

concepts are taken into consideration in views expressed on Twitter about these systems.

However, there are still several research gaps regarding the blame assignment process. Most notably, for instance, examining how decision-making algorithms are potentially blamed (or not) in social media discourses would provide insights into the public perception and discourse surrounding these systems. Understanding the dynamics of blame attribution in social media discussions can shed light on the factors influencing public trust, skepticism or criticism towards decision-making algorithms. By bridging this research gap, this thesis will provide a deeper understanding of the societal implications of decision-making algorithms, which may be used by the developers of these systems to inform strategies for fostering transparency, accountability and ethical use in their deployment.

Additionally, while the focus of this thesis is very defined, it is important to acknowledge other factors are still overlooked in this area of research. Firstly, current research lacks discussion on the cultural and societal factors that impact blame assignment. It acknowledges cultural differences can significantly influence how blame is assigned and that emotions such as anger or fear can lead to biased decision-making. Secondly, as this review has focused solely on blame assignment within the context of decision-making algorithms, it has not fully explored the roles of other actors, such as regulators and data providers. Examining their contributions and responsibilities can provide greater insight into how blame is assigned. This is partly done in this thesis in Chapter 4, but goes beyond the defined scope of maintaining focus on solely the systems.

In summary, the challenge of assigning blame in the context of decision-making algorithms is multifaceted, involving not only the developers and operators but also data providers, regulators and the algorithms themselves. Incorporating transparency, accountability, fairness and explainability in decision-making algorithms can promote ethical and responsible use and provide clearer guidelines for

assigning responsibility in cases where negative outcomes occur. By investigating how such systems are presented on social media, this thesis will provide an insight into blame assignment for decision-making algorithms so that developers and promoters are able to decipher which of these elements need to be addressed to foster continued usage or encourage adoption.

### 2.3.3 *Section Summary*

This exploration of trust and blame within decision-making algorithms has demonstrated their pivotal roles in the ethical deployment and societal acceptance of these systems (Heaton et al., 2023a). Trust, a multidimensional construct shaped by experience and context, is fundamental for users to perceive algorithms as reliable and autonomous (Bonnefon, Shariff, and Rahwan, 2016; Mayer, Davis, and Schoorman, 1995). Studies suggest that fulfilling responsibilities and ensuring accountability among developers and operators can enhance algorithmic trustworthiness, vital for widespread acceptance (Alaieri and Vellino, 2016; Lyons et al., 2017). Nevertheless, challenges persist, especially regarding transparency and ethical principles, highlighting the need for further research to fortify trust in these systems (Kwiatkowska and Lahijanian, 2016; Shahrdad and Amirani, 2018).

Conversely, blame assignment in decision-making algorithms involves a complex interplay between agency, responsibility and accountability. Higher algorithmic agency might result in a 'responsibility gap', complicating the attribution of blame for negative outcomes (Mittelstadt et al., 2016). Fulfilling ethical obligations by developers and operators emerges as a crucial aspect in preventing trust erosion and blame assignment (Jobin, Ienca, and Vayena, 2019; Taddeo and Floridi, 2018). However, challenges persist due to the complexity and opacity of these systems, urging the incorporation of fairness, trans-

parency and explainability to navigate the ethical terrain (Gunning et al., 2019; Selbst et al., 2019).

Overall, the complex nature of trust and blame within decision-making algorithms underscores the necessity for transparency and fairness to bolster societal trust and assign responsibility effectively. This may be possible to achieve through the examination of social media discourses for views expressed about decision-making algorithms. The evolving landscape demands interdisciplinary approaches to mitigate challenges and establish clearer ethical guidelines based on public response to these systems, crucial for navigating their ethical implications.

## 2.4 DECISION-MAKING ALGORITHM CASE STUDIES

This section of the review comprises an examination of literature relating to the three decision-making algorithm case studies for the thesis that will be detailed in Chapters 4, 5 and 6: the 2020 A Level algorithm (subsection 2.4.1), the NHS Covid-19 contact tracing app (subsection 2.4.2) and ChatGPT (subsection 2.4.3). As mentioned in Chapter 1, these systems were chosen as they were public-facing, generated significant public interest, and were popularly discussed on Twitter, as evidenced by the number of tweets collected for each case study and the richness of the discourse. Each of these sections examines the background of the system and its social impact.

### 2.4.1 *The 2020 A Level Algorithm*

#### 2.4.1.1 *Background to The 2020 A Level Algorithm*

On August 13th 2020, Ofqual (The Office of Qualifications and Examinations Regulation), the UK examinations regulations body, used a decision-making algorithm to replace the standard A Level qualifications, which had been cancelled that year due to the Covid-19

pandemic. The algorithm – defined here as the processing of data to produce a score through classification and filtering (Diakopoulos, 2016) – used prior centre attainment and teacher assessments to generate a grade for each qualification (Rosamond, 2020). In comparison to the predicted outcomes submitted by their teachers, 35.6 percent of students had qualification results lowered by one grade, 3.3 percent by two grades, and 0.2 percent by three grades (Whittaker, 2021). The conditions that their university offers or employment opportunities were required were unmet. Therefore, their career plans were irreparably compromised.

This became a highly contested issue to schools, regulators and the wider public (Kelly, 2021). The key aspect criticised was that prior assessment data and teacher-assessed grades had been submitted but not used in their sole form (Edwards, 2021). Instead, they were combined with previous assessment data. That rendered the calculation unfair to students and educators from highly deprived communities, especially.

The UK government defended the use of the algorithm initially, as it helped combat grade inflation. However, due to public outcry that these algorithmic-generated grades were unfair (Jiang and Pardos, 2021; Kelly, 2021), it retracted the algorithm-generated grades on August 17th 2020. Instead, all qualifications were awarded the teacher-submitted grades (BBC, 2020). The Education Secretary of State at the time, Gavin Williamson, appeared to place blame on Ofqual and emphasised he was not aware of the scale of the problem (Timmins, 2021). The public reaction also saw the resignations of Sally Collier, Chief Executive Officer and Chief Regulator of Ofqual, and Jonathan Slater, the most senior civil servant in the Department for Education. Therefore, the social impact of the choice went well beyond the class of 2020.

Ofqual reported there was no grading bias (Ofqual, 2020). However, it was found that the algorithm favoured students from more economically privileged backgrounds while others suffered more (Crisp

et al., 2024; Mallett, 2023; Smith, 2020). This was due to each school's historic results being a significant factor in the algorithm's grade calculation. This led to the algorithm being labelled as 'mutant' by UK Prime Minister Boris Johnson (Coughlan, 2020). Ofqual officials were quick to blame 'overly generous teachers', but not the algorithm itself or the decisions behind its deployment (Kelly, 2021), deflecting blame from human agents to the algorithm.

#### 2.4.1.2 *Social Impact*

Several studies examined the social impact of the algorithm. Bhopal and Myers (2020) surveyed 583 students and interviewed a further 53 students who were eligible to take A Level examinations, between April and August 2020. Their aims were to examine the impact (mental and academic) of predicted grades on A Level students, explore support systems in place for such students and analyse differences by race, class, gender and school type. Through quantitative and qualitative analysis, it was found that students had identified the significance of unfairness within their individual experiences. Students from all types of schools and backgrounds felt the deployment of the algorithm placed little or no value on individual students' experiences. Consequently, many students received results they perceived to be unfair (21% of those surveyed said they were happy with their results), which was in contrast to the official investigation report that concluded that there was no grading bias (Ofqual, 2020). This highlights further need to investigate responsibility in this area.

Additionally, Kolkman (2020) noted that the incident shone a light on algorithmic bias. However, he also noted that greater knowledge of algorithmic-driven decisions requires a better understanding of the functionality. More specifically, the author foregrounded the importance of critical reflection within the process of algorithm design and noted that, without intervention, there will be further unrest and distrust in algorithms that impact daily lives. Hecht (2020) further examined the social impact of using the algorithm. They stated that

public awareness, scrutiny and transparency are critical first steps to eliminate perceived bias from the algorithm but far from a guarantee. Therefore, these are important factors to consider when examining views expressed about the algorithm. Ultimately, the current literature demonstrates that different entities have been blamed for the algorithm's failure, yet limited research into how social media users reacted to the scandal, thus providing motivation for the focus on this case study, specifically, as part of this thesis on agency, trust and blame in decision-making algorithms.

### 2.4.2 *The NHS Covid-19 App*

#### 2.4.2.1 *Background to The NHS Covid-19 App*

The NHS Covid-19 App, the contact-tracing algorithmic-based system created by Serco on behalf of the UK government to track active cases of Covid-19, has impacted the United Kingdom on multiple levels, since its launch (Kretzschmar et al., 2020). The application is available on mobile phones and uses exposure logging, developed by Apple and Google (NHS England, 2021). This technology allows the application to send alerts, using a randomly generated identification number, when the user is near another application user who has logged a positive Covid-19 test. Despite its scientific-based intended functionality, its users reported issues regarding backwards incompatibility, incorrect alerts and false positive tests (Morales et al., 2021; Wee and Findlay, 2021). Such unexpected technical problems meant that users had to self-isolate for ten days even when the result was incorrect, with inadvertent consequences on their income and well-being (Bardosh et al., 2022; Kent, 2020).

Despite the UK government encouraging its adoption (Jacob and Lawarée, 2021), the uptake of the app was less than expected at 20.9 million downloads between September and December 2021, with 1.7 million notifications being sent out in England and Wales (Pandit et

al., 2022). According to Wymant et al. (2021), every 1% increase in the number of app downloads led to a 0.8–2.3% reduction in the number of Covid-19 infections, with their findings suggesting that anywhere between 100,000 and 900,000 cases were averted because of the information inputted by users into the system. However, Mbwogge (2021) claimed that a symptom-based contact-tracing system failed to meet the testing and tracing needs in the United Kingdom, which is further evidenced by the fact that cases of and deaths relating to Covid-19 increased to be the highest in Europe.

#### 2.4.2.2 *Social Impact*

Perhaps because of its technical challenges, a growing number of research projects investigated the public attitudes towards digital contact tracing in the UK. Williams et al. (2021) interviewed 27 participants over online video conferencing before the release of the Covid-19 app in the UK and found the response to be mixed and heavily influenced by moral reasoning. Analysis revealed five themes: lack of information and misconceptions surrounding Covid-19 contact-tracing apps; concerns over privacy; concerns over stigma; concerns over uptake; and contact-tracing as the 'greater good'. Samuel et al. (2021) conducted 35 semi-structured qualitative interviews in April 2020, showing interviewees' views about the potential of the app for contact tracing. Participants showcased a range of misconceptions and worries. However, as there was no follow-up to this study, it was impossible to discover which of the participants would then choose to download or not to download the app once it was launched in September 2020. These insights shall inform the investigation of the impact of the NHS Covid-19 App on British society throughout the pandemic and the perceptions of this system by its (intended, actual or former) users.

This possible evolution of the attitudes towards the app was instead monitored by Dowthwaite et al. (2021), who surveyed 1,001 UK adults and found that half of the participants had installed the

app, with 60% of these claiming to comply with it on a regular basis. They also found that there were issues surrounding trust and understanding that hindered the effective adoption of the app. Follow-on analysis showed that there were statistically significant correlations between lower trust amongst non-users, many aspects of the app and the wider social and societal context (Dowthwaite et al., 2022). A year after the app was launched, Pepper et al. (2022) identified five main themes during follow-up interview discussions: flaws in the app, usefulness and functionality affecting trust in the app, low trust in the UK government, varying degrees of trust in other stakeholders and public disinterest. According to the study results, these factors contributed to a drop in compliance over time. Similar findings were put forward by Paucar et al. (2022), who stated that responsibility and trust made the app better accepted by the public. Even though these were always relevant, other factors, like fear of infection, were contextual- and time-dependent. Arguably, this will be relevant when examining the presentations of social agents who tweeted about such an app system and its functionality, as perceived or evaluated by its self-proclaimed users or experts.

In July 2021, when the relaxation of government restrictions led to an increased amount of positive Covid-19 cases in the UK, the media scrutiny on the app intensified as a result of the numerous notifications sent through the app (Abbasi, 2021). As a result, this impacted the public's perception of the app and the pejorative blend 'pingdemic' was coined (Rimmer, 2021). This exemplifies the considerable impact that the deployment of the NHS Covid-19 App has had on British society and how this was reflected by (social) media and the terminology they used (Heaton et al., 2023d).

While it has been established that there is research interest in digital contact tracing from a sociological and epidemiological standpoint (Dowthwaite et al., 2021; Marsh et al., 2021; Paucar et al., 2022; Pepper et al., 2022; Smith et al., 2022), a gap in the presentation of the app itself was detected, which this thesis aims to address.

### 2.4.3 *ChatGPT*

#### 2.4.3.1 *Background to ChatGPT*

ChatGPT, developed by OpenAI, is an advanced AI chatbot designed to engage in human-like conversations with users (Rathore, 2023). Leveraging deep learning models and natural language processing techniques, ChatGPT is capable of understanding and generating human-readable text in a conversational manner (Hariri, 2024). It is trained on a vast amount of text data from diverse sources, enabling it to comprehend and respond to a wide range of queries and prompts (Haleem, Javaid, and Singh, 2022). At its core, ChatGPT utilises a transformer-based language model, which allows it to capture the contextual dependencies and semantic nuances in natural language (Ray, 2023). The model has been fine-tuned using reinforcement learning from human feedback, enabling it to generate coherent and contextually relevant responses (Hassani and Silva, 2023).

Users interact with ChatGPT through a user-friendly interface, engaging in real-time conversations with the chatbot (Firat, 2023). It aims to simulate natural conversations, offering assistance, entertainment and creative collaboration, marking a notable advancement in AI-driven conversational systems.

In terms of a timeline, ChatGPT was launched in chatbot form on 30 November 2022 (Haleem, Javaid, and Singh, 2022; Taecharungroj, 2023; Whalen, Mouza, et al., 2023). This built upon OpenAI's existing GPT-3 model and was set up as a conversational AI system capable of engaging with users, addressing follow-up questions, challenging erroneous assumptions and rejecting inappropriate requests. ChatGPT was trained using Reinforcement Learning from Human Feedback (RLHF) and fine-tuned based on the GPT-3.5 model (Chen et al., 2023).

In January 2023, ChatGPT achieved a significant milestone, surpassing 100 million monthly users at a faster rate than popular so-

cial media platforms like Instagram or TikTok (Ye, 2023). Its capabilities were showcased when the chatbot successfully passed prestigious graduate-level exams, garnering considerable attention (Ali et al., 2022). However, its popularity meant that it was sometimes difficult to access, with outages leading to frustration from users (Zhang, 2023).

By the end of January 2023, OpenAI introduced the AI Text Classifier, a novel tool intended to address concerns regarding academic dishonesty associated with the use of ChatGPT (Antaki et al., 2023; Dönmez, Sahin, and Gülen, 2023). The primary objective of this tool was to assist educators in identifying instances where a student or an AI system, such as ChatGPT, may have generated a specific assignment. Furthermore, OpenAI emphasised the potential of the AI Text Classifier in detecting disinformation campaigns and preventing the misuse of AI. However, the classifier was retired in July 2023 due to the low accuracy of the system (Hu, Chen, and Ho, 2023).

On 1 February 2023, OpenAI initiated the implementation of an experimental subscription plan, ChatGPT Plus, aimed at providing enhanced user experience and accessibility for ChatGPT, priced at $20 per month (Aiyappa et al., 2023). It was stated that ChatGPT Plus included expedited response times, priority access to novel features and enhancements and unrestricted availability to ChatGPT, even during peak usage periods (Xie et al., 2023). These developments highlight the rapid adoption and substantial societal impact of ChatGPT within a short timeframe.

On 1 March 2023, OpenAI launched a new Application Programming Interface (API) that facilitates the seamless integration of ChatGPT technology into a wide range of business applications, websites and services (Cao and Zhai, 2023). The pricing structure for this API was set at $0.002 per 1,000 tokens, corresponding to approximately 750 words, building on the 'GPT-3.5-turbo' AI model.

On 14 March 2023, OpenAI introduced GPT-4, an AI language model capable of analysing both text and image inputs, though lim-

ited to text output (Sanderson, 2023). Despite acknowledging shared limitations with earlier models, OpenAI partnered with organisations like Duolingo, Stripe and Khan Academy to integrate GPT-4, accessible to developers through an API, into various products (Gallifant et al., 2024). OpenAI provided GPT-4 to the public via the ChatGPT Plus subscription service, emphasising its improved creativity, collaboration and problem-solving accuracy (Rudolph, Tan, and Tan, 2023). Additionally, ChatGPT received an update incorporating the GPT-4 model, rendering it a multimodal system (Roose, 2023).

#### 2.4.3.2 *Social Impact*

Despite the short amount of time since its launch, the social impact of ChatGPT has been widespread (Abdullah, Madain, and Jararweh, 2022). The release of ChatGPT has garnered significant attention and public fascination, despite its limitations (Verma and Lerman, 2023). Journalistic reports have underscored the astonishment and intrigue from academics and technology professionals (Kelly, 2023). Moreover, concerns have emerged regarding the system's potential to generate and disseminate believable misinformation, leading to apprehension among users.

These assertions are founded on both observed and speculative use cases of ChatGPT and its predecessors, as documented by researchers and journalists. The potential applications of ChatGPT encompass a wide array of tasks, ranging from generating written content for various purposes such as minutes (Taecharungroj, 2023), websites (Kellerman, 2023), newspaper articles (Ray, Ghasemkhani, and Martinelli, 2024), reports (Kumar, 2023), poems (Michaux, 2023), songs (Zhuo et al., 2023), jokes (Kirmani, 2022) and scripts (Shafeeg et al., 2023). It can also facilitate code debugging (Feng et al., 2023), organise unstructured data (Hassani and Silva, 2023), generate queries and prompts (Wang et al., 2023), create 'no-code' automated applications for businesses (Taecharungroj, 2023), design ideation processes (Kocaballi, 2023) and provide therapeutic support (Kalla and Smith, 2023).

These diverse use cases vividly illustrate the extensive utility and perceived influence of ChatGPT.

One of the earliest studies regarding the social impact of ChatGPT was by Abdullah, Madain, and Jararweh (2022), who examined the multifaceted implications of ChatGPT across diverse domains, encompassing software development, media and news and education. Notably, they found that ChatGPT exhibited promising prospects in enhancing individuals' productivity and task completion efficiency. However, concurrent with the potential benefits, apprehensions arose concerning the potential misuse of ChatGPT, particularly within educational contexts. Moreover, the study highlighted the utility of ChatGPT in the analysis of user conversations and media interactions. By scrutinising these interactions, ChatGPT enabled the identification of both positive and negative trends within news content.

As research into ChatGPT has developed, there has been a focus on the 'panic' and concerns that have surrounded its launch and integration into society. Studies have shown that ChatGPT has the potential to fabricate information and present it as truth in contexts such as writing systematic reviews (Najafali et al., 2023) and healthcare warnings (De Angelis et al., 2023). Furthermore, the use of large language models in customer service could potentially lead to job loss in this particular industry, along with others (Aljanabi, 2023). Investigating this topic, Biswas (2023) asked ChatGPT to generate its own view on AI job displacement, where they found that customer service representatives, translators and interpreters, content writers and data analysts were most at risk.

With regard to ethical concerns, Zhou et al. (2023) found that some potential ChatGPT ethical concerns included bias in training data, privacy implications and the risk of malicious use and abuse. Looking specifically at ethics in scientific research, Ray (2023) outlined several areas of concern, including reliability, quality control, energy consumption, safety, privacy, intellectual property and authorship, responsibility, accountability, transparency, bias and discrimination. Re-

search has also shown that human oversight plays a vital role in providing context and ethical judgment that AI models may lack, which supports the identification and mitigation of potential biases, errors or unintended consequences (Ferrara, 2023). Building on previous assertions by Jasanoff (2020), who presented the idea that technological failures and societal harm are often depicted as unintentional outcomes or results of misapplication, Doshi, Bajaj, and Krumholz (2023) found that ChatGPT will instill awe but it needs to elicit appropriate action to evaluate its capabilities, mitigate its harms and facilitate its optimal use.

Researchers have also conducted studies into the educational impact of ChatGPT more specifically. For example, Tiwary, Subaveerapandiyan, and Vinoth (2023) aimed to explore the perspectives and sentiments of academics and information professionals towards ChatGPT. Through social media comments and a survey, they found ChatGPT-3's potential in research and writing tasks but highlighted the need for verification and fact-checking due to acknowledged limitations. Moreover, they revealed a noticeable shift in the attitudes of most of the academics surveyed, who were increasingly embracing ChatGPT despite initial resistance. This study offered valuable insights and guidance for academic professionals, content developers and librarians to navigate ChatGPT effectively. Additionally, Khalil and Er (2023) examined the effectiveness of ChatGPT in generating academic essays that can circumvent plagiarism detection mechanisms. Their findings indicated ChatGPT's potential for generating original content in diverse subjects, underscoring the importance for educational institutions to address potential plagiarism challenges resulting from AI technology integration.

Some studies have focused on the political nature of ChatGPT. For example, Hartmann, Schwenzow, and Witte (2023) analysed ChatGPT's political ideology through an extensive examination of its responses to 630 political statements. The study revealed ChatGPT's consistent pro-environmental, left-libertarian orientation, evident in

its support for policies like flight taxes, rent restrictions and abortion legalisation, highlighting the need to recognise and understand the potential impact of politically biased conversational AI on society and its ethical implications. These findings were, however, in direct contradiction to a piece of research by the BBC, which stated that ChatGPT should not 'express political opinions or engage in political activism' (Whannel, 2022).

Researchers have situated ChatGPT in the broader sphere of generative AI. For example, Fischer (2023) examined the implications of generative AI systems, such as ChatGPT, and highlights associated risks including false authorship, unreliable advice and job displacement in copywriting. This highlights a shift in the study of generative AI, focusing on its organisational and technological practices and its integration into human activities. It underscores the need for further research and user studies to explore individual vulnerability to AI-generated advice and address source attribution and citation concerns, emphasising the need for ongoing investigation and understanding.

However, as mentioned previously, Abdullah, Madain, and Jararweh (2022) found that, in terms of societal impact, the full extent of ChatGPT's impact is yet to be determined. They acknowledged the significant progress made in natural language processing and AI capabilities with the advent of advanced language models. The potential applications of ChatGPT can have wide-ranging implications, including improving conversations, providing deeper insights into humanity and facilitating tasks in fields such as programming, content generation, planning and more. However, they also raise concerns about the ethical use of ChatGPT and the need to address issues related to misinformation, biases and privacy.

#### 2.4.3.3 *Agency, Social Action and ChatGPT*

To date, there has been a small number of studies that have specifically investigated the agency and social actor status of ChatGPT. In

their study, Bran et al. (2023) found that four analysed news sources presented conflicting narratives about ChatGPT's competence as a social actor, with some depicting it as creative and healing, offering valuable novelty, companionship and the need for social acceptance. Conversely, some sources portrayed AI as incompetent, polluting human culture and replacing human skills and knowledge with stochastic, illusionary competence or even imposture. This dual representation leads to a contested social agency, vacillating between creative actors and essentially unthinking tools.

Additionally, Shijie, Yuxiang, and Qinghua (2023) underscored the importance of evaluating the credibility of AI-generated content by considering AI's role as a content generator and technological medium. They highlighted the need to incorporate AI's explainability and generative capabilities through third-party interface, whilst also keeping a significant focus on viewing AI through an anthropomorphic lens, treating it as a social actor in order to better assess its credibility.

Gutiérrez (2023) stated that AI systems, such as ChatGPT, function as social intermediaries within a network of actors and associations, where they play a role in generating outputs and intentions. In contrast to other findings, while AI lacks moral agency, its interactions have consequences, potentially including bias, accountability issues, transparency concerns and privacy implications.

The tendency to treat ChatGPT as a social actor, in the same manner as other examples of AI, has been explored in an educational context by Dai, Liu, and Lim (2023). Through discussing the potential benefits and challenges of ChatGPT and generative AI in higher education, including its potential to enhance student learning, propel student engagement, impact academic integrity and alter the role of educators, they found that ChatGPT has the potential to be a student-driven innovation. More of a focus was put on ChatGPT empowering students' epistemic agency, but it is clear that careful consideration must be given to its implementation and use.

Research has also been undertaken into the perception of ChatGPT's human-like traits in society by Al Lily et al. (2023), who analysed insights from 452 individuals worldwide, leading to the identification of two distinct categories of traits. The first category revolved around social traits, where ChatGPT assumes the roles of an 'author', mirroring human phrasing and paraphrasing practices, and an 'interactor', emulating human collaboration and emotional engagement. The second category revolved around political traits, with ChatGPT adopting the roles of an 'agent', replicating human cognition and identity, and an 'influencer', simulating human diplomacy and consultation. Interestingly, ChatGPT itself acknowledged the possession of these human-like traits, reinforcing its role in human society as a 'semi-human' actor that transcends its machine-based origins and technical essence.

Despite these studies, there is yet to be substantial research that focuses on how ChatGPT is presented on Twitter with regard to its perceived social agency, as this thesis intends to accomplish.

### 2.4.4 *Section Summary*

The examination of three decision-making algorithm case studies -— the 2020 A-Level algorithm, the NHS Covid-19 contact-tracing app and ChatGPT -— has showcased insights into their profound societal impact, ethical intricacies and challenges in user acceptance. To recap, Ofqual's deployment of the 2020 A Level algorithm triggered significant criticism due to its reliance on historical data, leading to perceived 'unfair' grade adjustments that disproportionately affected disadvantaged students (Rosamond, 2020). The subsequent reversal of this decision exposed the severe social consequences, becoming a pivotal moment in educational policy and public trust in these types of systems.

Likewise, the NHS Covid-19 app encountered issues including technical flaws, privacy concerns and varying public reception, illustrating the intersection between technological functionality and societal adoption (Kent, 2020). Despite its role in curbing Covid-19 cases, the app's limitations raised doubts about its effectiveness and public acceptance (Wymant et al., 2021). This highlighted the interplay between technological deployment and its societal impact, emphasising the critical need for user trust, transparency and ethical considerations within algorithmic systems.

Additionally, ChatGPT embodies the advancements and challenges in conversational AI, signifying a new era in human-machine interaction (Ray, 2023). Its rapid integration into society showcased its wide-ranging influence, from enhancing productivity to content generation in diverse domains (Abdullah, Madain, and Jararweh, 2022). However, concerns surrounding misinformation, ethical use and potential job displacement have emerged, posing significant ethical and social dilemmas (Fischer, 2023).

In summary, this section of the review has showcased examples of the intricate relationship between decision-making algorithms and society. They underscore the profound impact of algorithmic decisions on individuals, institutions and societal trust, prompting further exploration of algorithmic systems' social agency and their portrayal in online spaces to comprehend their reception and societal implications more comprehensively. With that in mind, despite the existing insights, the comprehensive understanding of how the social agency of all three of these decision-making algorithms is portrayed on platforms like Twitter remains largely unexplored, including the implications that this has on trust and blame.

## 2.5 CHAPTER SUMMARY

This literature review began by providing a multifaceted exploration of how grammatical agency reflects social agency within the context of decision-making algorithms in section 2.1, intertwining control, responsibility and social interactions. The section also introduces the pivotal concept of SAR and how it can be used to analyse grammatical structures, including active and passive constructions alongside transitivity patterns, meaning that it can shed light on how users assign responsibility to decision-making algorithms on Twitter (Van Leeuwen, 2008). Passive constructions and lexical choices emerge as pivotal influencers shaping decision-making potential and animacy perception.

In section 2.2, the complexity of agency, responsibility and accountability was explored alongside debates emerging on algorithmic autonomy, ethical considerations and transparency challenges. Efforts to enhance fairness and accountability through interdisciplinary collaboration underscore the quest to support societal agency. Yet, attributing responsibility and ensuring accountability demand further exploration in societal perceptions, especially online.

Section 2.3 illuminates trust's pivotal role in algorithmic acceptance, shaped by experience and context. Challenges persist in transparency and ethical principles, necessitating further research to fortify trust. Conversely, blame assignment involves a complex interplay between agency, responsibility and accountability, urging fairness, transparency and explainability in ethical navigation.

Examining case studies —- the 2020 A Level algorithm, the NHS Covid-19 app and ChatGPT —- in section 2.4 reveals societal impact, ethical intricacies and user acceptance challenges. These cases emphasise the need for trust, transparency and ethical considerations in algorithmic systems. This review prompts further exploration of societal perceptions in online spaces for comprehensive understanding.

In summary, this chapter unveils the intricacies of agency, responsibility, trust and blame within decision-making algorithms, highlighting their profound societal implications. Yet, the portrayal of these systems' social agency on platforms like Twitter remains largely unexplored. Further research is crucial to grasp societal perceptions and ethical implications surrounding these systems comprehensively.

# 3

# APPROACH

This chapter outlines the analytical approach taken within this thesis. To accomplish this, the chapter begins with section 3.1, which covers the approach to data collection and management in the context of the Twitter data collected for this PhD project. This chapter then discusses the overarching approach of this thesis in section 3.2. Within this, this section will discuss how individual approaches are deployed within this PhD thesis. This overarching analysis process is demonstrated visually in Figure 2.

Next, the following sections outline the state of the art of the computational and discursive linguistic approaches used in this thesis. Therefore, section 3.3 details the three NLP-based computational techniques (topic modelling in 3.3.1.1, sentiment analysis in 3.3.1.2 and emotion detection in 3.3.1.3) used to capture initial data trajectories. This is followed by comprehensive reviews of Corpus Linguistics (CL) in subsection 3.4.1 and Discourse Analysis (DA) in subsection 3.5.1. Each of these approaches will be used to analyse the case studies in Chapters 4, 5 and 6.

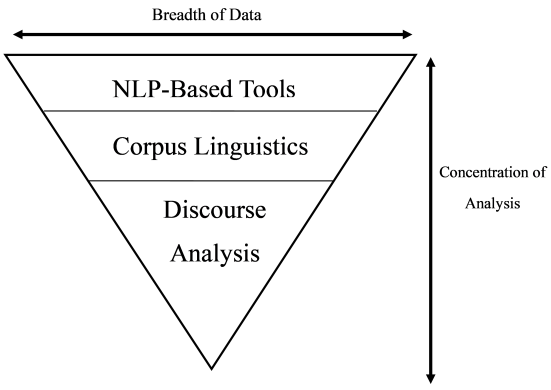


Figure 2: Figure to illustrate the proposed analytical approach.

## 3.1 DATA COLLECTION

Data collection occurred through using the Twitter for Academic Purposes Application Programming Interface; the code for data collection can be found in the University of Nottingham Research Data Management Repository. Twitter was used as a data source because of data availability, amount and real-time analysis reasons. Despite these considerable advantages, using Tweets as a specific social media dataset comes with both risks and rewards (Agarwal et al., 2011; Picazo-Vela, Gutiérrez-Martínez, and Luna-Reyes, 2012; Ventola, 2014). The fact that data can be collected from Twitter in real time supports the current computational linguistic analysis models that have the functionality to do the same (Gupta and Hewett, 2020; Kumar, Morstatter, and Liu, 2014). An important aspect of using Twitter is that its data can be pre-processed before analysis (Jianqiang, 2015; Symeonidis, Effrosynidis, and Arampatzis, 2018) and lends itself well to exploratory analysis principles (Chong, Selvaretnam, and Soon, 2014; Ottovordemgentschenfelde, 2017).

Complex ethical considerations are needed when scraping data from Twitter for data analysis. Although tweets are public (by default), Twitter 'data' is not provided by users for the purposes of research, and gaining explicit consent to use tweets for research is practically infeasible (Bishop and Gray, 2017; Williams, Burnap, and Sloan, 2017; Woodfield et al., 2013). Therefore, best practices recommended in social media research literature were followed, including not including any screenshots of tweets that may be later identifiable without first gaining consent from the tweet author (Bishop and Gray, 2017; Mason and Singh, 2022). This also included reporting only short, verbatim quotes, no longer than a few words, to illustrate study findings (Ahmed, Bath, and Demartini, 2017; Mason and Singh, 2022). Instead, tweets were anonymised to researchers as part of the data cleaning process. This PhD project was approved by the School of Computer Science's ethics committee (approval number CS-

2020-R33). The privacy notice and project information documents associated with this PhD project can be found in Appendices A and B respectively.

With regard to the data collection, this varied between each of the case studies and, thus, will be detailed in individual chapters. This is due to the various data collection terms used to capture the discourses on Twitter.

## 3.2 OVERARCHING ANALYSIS PROCESS

This section details the overarching analysis process, providing an overview of the individual processes undertaken and how they fit together within an interdisciplinary approach, adopted in this thesis project. While multidisciplinarity involves multiple disciplines working in parallel whilst maintaining distinct methodological boundaries (Arnold et al., 2021; Dwivedi et al., 2021), interdisciplinarity actively synthesises diverse perspectives and methodologies to create new unified frameworks (Arnold et al., 2021; Carr, Loucks, and Blöschl, 2018; Lang et al., 2012; Tobi and Kampen, 2018). This collaboration fosters the fusion of varied perspectives, methodologies and theories, often resulting in innovative solutions that cross between singular disciplinary boundaries (Palmer, 1999). In this thesis, an interdisciplinary approach facilitates a comprehensive examination of public discourse surrounding decision-making algorithms to provide nuanced insights into societal perceptions. As this thesis combines NLP, CL and DA in an integrated manner, the approach goes beyond applying multiple disciplinary lenses to the data. It involves using insights from one approach to inform the others, as the initial insights from NLP techniques will be further explored using CL and DA to understand the nuances of language use. Furthermore, the thesis is framed around a single overarching research question concerning agency, trust and blame in the discourse surrounding decision-

making algorithms. Each case study, and the combined approaches employed, are directed towards answering this question.

Contemporary challenges, such as expressions of social agency, trust and blame in decision-making algorithms, demand comprehensive approaches that transcend disciplinary confines (Dignum, 2020; Mohseni, Zarei, and Ragan, 2021; Piorkowski et al., 2021). Interdisciplinary research serves as a conduit for comprehensively understanding challenges such as the focus of this thesis, thereby facilitating the formulation of holistic strategies to effectively tackle them (Palmer, 2013; Weber-Lewerenz, 2021). This propels exploration into new research trajectories (Carr, Loucks, and Blöschl, 2018; Fallman, 2008; Lang et al., 2012), resulting in advancements that underscore how interdisciplinary research drives groundbreaking knowledge discovery (Haleem, Javaid, and Singh, 2022; Li, Chen, and Larivière, 2023) and understanding technologies with broad societal benefits (Dwivedi et al., 2021; Koohang et al., 2023). In this context, the interdisciplinary approach adopted in this thesis enables a thorough examination of public discourse surrounding decision-making algorithms, offering insights that uncover how social media users represent these systems and the implications for adoption and use.

Despite its merits, interdisciplinary research encounters challenges, including bridging communication gaps between disciplines and reconciling conflicting methodological approaches (Dwivedi et al., 2021; Palmer, 2013). Moreover, the prevailing preference for discipline-specific research within traditional academic structures and funding mechanisms poses impediments to sustained interdisciplinary (Lyall, 2019; Siedlok and Hibbert, 2014). Despite these challenges, the pivotal role of interdisciplinary research is to address complex problems, nurture innovation and expand knowledge across diverse domains (Lang et al., 2012; Palmer, 1999; Tobi and Kampen, 2018). Its potential to affect transformative breakthroughs renders it a suitable approach for this thesis as it aims to broaden the scope of understanding about the social impact of decision-making algorithms.

Table 1: The different approaches used in this thesis, including the data and analysis procedures and the intended insight.

| Approach | Analysis and Data | Insight |
|---|---|---|
| NLP-Based Approaches | Topic modelling, emotion dection and sentiment analysis;data sets divided into time periods relative to the size of the discourse (e.g., Covid-19 app by months, ChatGPT by weeks), allowing for a more manageable analysis. | Provided an initial investigation of each discourse, highlighting key themes, trends and changes over time, and setting the stage for more detailed analysis. |
| Corpus Linguistics | Splitting by time period (ChatGPT, Covid-19 app) or entity (A Level algorithm), a focus on keywords and collocation to investigate the grammatical structures and their presentation. This aided identification of how words and phrases were used in context and their associations with other words. | Offered insights into the underlying grammatical patterns and structures within the discourse, exposing stylistic and structural elements of language. |
| Discourse Analysis | Focusing specifically on agency, examination of areas of specific interest that were identified through the initial NLP-based analysis and the subsequent CL analysis. These starting points were crucial for a deeper examination. | Enabled a detailed analysis of agency in the discourses; provided a greater understanding of the implications for trust and blame. |

The rationale for using the three individual approaches is detailed in Table 1.

As a result, the amount of data that is able to be analysed using the NLP-based tools is higher. As points of interest are identified by viewing trends and trajectories, these can be used to inform the CL analysis. This meant that, as the analysis moved forward, the CL results informed the focus for the DA.

Within the following sections, the background of each individual approach and the way that it was deployed is detailed. This may differ somewhat between each case study; therefore, there are details that are specifically omitted here and included in Chapters 4, 5 and 6 instead.

## 3.3 NLP-BASED COMPUTATIONAL LINGUISTIC APPROACHES

### 3.3.1 *Background*

Three popular computational linguistic tools have been chosen as the means to explore these public discourses on Twitter. The three approaches – topic modelling, sentiment analysis and emotion detection – are united in their use of language to either describe or make predictions about a corpus. Thus, they are descriptive and predictive in their function. They are, furthermore, popular choices when there is a large amount of linguistic data to explore. Topic modelling is best used to uncover latent topics present within large bodies of text (Blei, Ng, and Jordan, 2003; Jelodar et al., 2019; Nikolenko, Koltcov, and Koltsova, 2017a). Sentiment analysis – otherwise known as opinion mining – uses predictive algorithms on a polarity scale to provide insight into the views expressed in text (Liu, 2010; Vyas and Uma, 2018). Emotion detection methods use similar predictive detection algorithms to sentiment analysis to ascertain emotions or states of being that may be expressed within a text (Mohammad and Turney, 2013; Sailunaz et al., 2018).

A number of alternative techniques could have been used, such as Named Entity Recognition, Convolutional Neural Networks (CNN) or deep learning, but these three approaches were chosen due to scope, suitability and accessibility. Ultimately, these three chosen approaches have consistently been used by researchers within this realm of social media discourses (Aribowo and Khomsah, 2021; González-Ibánez, Muresan, and Wacholder, 2011; Gupta and Hewett, 2020; Hu, Chancellor, and De Choudhury, 2019; Mathur, Kubde, and Vaidya, 2020; Sengupta, 2019; Villena-Román and Garcıa-Morera, 2013). They have also been chosen as their outputs can be compared in a diachronic way (Fernández-Cruz and Moreno-Ortiz, 2023; Wicke and Bolognesi, 2021).

#### 3.3.1.1 *Topic Modelling*

Automated topic modelling, particularly Latent Dirichlet Allocation (LDA), is seen as beneficial in qualitative text studies due to its focus on uncovering underlying topics present within a series of documents (Blei, Ng, and Jordan, 2003; Jelodar et al., 2019; Nikolenko, Koltcov, and Koltsova, 2017a; Nikolenko, Koltcov, and Koltsova, 2017b). The use of LDA as a topic modelling method with Twitter data has grown in interest in recent years (Arianto and Anuraga, 2020; Jelodar et al., 2019) and various techniques have been designed to undertake this investigation. There has been focus on using bigrams – pairs of adjacent words – to form topic models through LDA when investigating views expressed on Twitter (Jelodar et al., 2019; Srinivasan and Mohan Kumar, 2019). This is closely related to the work done by Yang and Zhang (2018), who used this with the Bag-of-Words (BoW) model. This aided the creation of topic models when working with Twitter data, especially when concerned with context dependence of short texts (the Twitter limit is currently set to 280 characters for unverified users). Moreover, Prihatini et al. (2018) combined, compared and contrasted LDA with Term Frequency Inverse Document Frequency (TF-IDF), a method regularly used for feature extraction of texts, when examining online news and related media articles. They found that the Precision, Recall and F-Measure values of LDA were higher than TF-IDF for predicting topics, thus the more suitable method for this scope.

One popular choice for topic modelling and LDA is the gensim package, using the analysis of co-occurring patterns to identify latent structures in plain text documents (Rehůřek and Sojka, 2011). Hidayatullah, Aditya, Gardini, et al. (2019) used gensim as the LDA model to investigate topics and trends regarding climate change and weather on Twitter, where five key topics were able to be defined through this method. Another relevant study was identified by Song et al. (2019), who investigated topics in media discourses regarding

illegal compensation given to victims of occupational injuries. The topics discovered were then used as recommendation points for ensuring efficient occupational health and safety schemes protect vulnerable employees from illegal practices, exemplifying the practical applications that using LDA can have.

This method has been used within contemporary social media studies, too. For example, Hu, Chancellor, and De Choudhury (2019) used gensim as their topic modelling tool when investigating discourses relating to homelessness on social media. They looked first at the blog posts of those who identify as homeless on the blogging site Tumblr and compared these to the blog posts of those who do not identify as homeless. By using LDA, thirteen latent topics were identified as part of the discourse of homeless blog posts and seventeen topics were uncovered to be part of the control group's discourse. Within this study, LDA was deemed successful in its identification of the different ways in which homeless people and non-homeless people discuss the topic of homelessness. It was recommended that organisations tap into these topical discourses to raise awareness and promote support, exemplifying the power of using this computational linguistic method. Additionally, Sengupta (2019) also used LDA to investigate latent topics in sub-Reddit forums. This combined manual inspection with LDA to validate findings.

How to best find the appropriate number of topics for a dataset when using LDA has been documented in studies, specifically. In their exploration of NLP-based techniques, Nguyen et al. (2020) suggested that the needs of the researcher must be examined: a small number of topics for a broad overview, and more topics for finer detail.

There have been critiques of topic modelling also. Maier et al. (2018) questioned the validity and reliability of LDA and offered an evaluative framework that enabled communication researchers to more effectively deploy this method. Their four recommendations were categorised as pre-processing the data, selecting parameters carefully,

evaluating the reliability of the model and, finally, checking the validation of the results via manual review. Researchers have also explored the limitations of 'off-the-shelf' topic modelling. In particular, stemmers (which are used to conflate several words to a shared meaning in topic modelling) were critiqued by Schofield and Mimno (2016), suggesting that they had no effect on the outcomes of topic models and often disrupted topic stability. Contemporary Twitter studies such as Saura, Ribeiro-Soriano, and Saldaña (2022) also expressed that labelling LDA topics is a manual process and, as such, this introduced bias into their results. Thus, this showcases the importance of using best practices and integrating other approaches to complement topic modelling analytics.

#### 3.3.1.2 *Sentiment Analysis*

Another popular investigation method to uncover the views expressed on social media is sentiment analysis. Sentiment analysis is defined as 'the computational treatment of opinions, sentiments and subjectivity of text' (Medhat, Hassan, and Korashy, 2014, p. 1). In its most common form, it uses a binary polarity scale from negative to positive, with neutral in between, initially examining the lexicon individually with a view to providing an overall sentiment of a text (Liu, 2010). It is often used to easily ascertain data that will provide insight into the opinions of others (Vyas and Uma, 2018), which enables investigation through a predictive element. Sentiment analysis has been used to investigate the meaning of Twitter discourses for this same reason (Liu and Zhang, 2012). For this thesis, it was chosen that the focus would be on dictionary-based approaches as these were the most common approaches seen in similar studies. Nevertheless, there are other approaches, such as CNN or deep learning methods, that are becoming popular within NLP-based social media analytics (Liao et al., 2017; Lu et al., 2017).

A staple of most Deep Learning (DL) sentiment analysis methods is the Naive Bayes classifier, which makes the simplifying assump-

tion that all words are sampled independently (hence the 'naive'). This produces a conditional probability of a document belonging to a category (Rish et al., 2001). Pak, Paroubek, et al. (2010) built a classifier based on Naive Bayes principles that classified Twitter data about a range of topics according to traditional sentiment polarity but claimed it outperformed other comparable classifiers in accuracy. Ulfa, Irmawati, and Husodo (2018) combined the Naive Bayes classifier with a Mutual Information method when examining tweets relating to tourism in Lombok and yielded a high classification accuracy. These approaches provide valuable insights into sentiment analysis techniques applied to Twitter data, which align with the sentiment analysis aspect of the approach of this thesis.

Two popular sentiment analysis modules that have been used for Twitter opinion mining are TextBlob and Valence Aware Dictionary and sEntiment Reasoner (VADER). TextBlob also uses the Naive Bayes model and also provides a subjectivity measure (Loria, 2018). Pokharel (2020) used this technique to analyse the response to the Covid-19 outbreak in Nepal and found that the majority of Nepalese citizens were taking a positive and hopeful approach, but there were instances of fear, sadness and disgust exhibited too. Additionally, Sivalakshmi et al. (2021) explored the sentiment towards the Covid-19 vaccine using TextBlob and concluded that the discourse was neutral-to-negative in polarity. An important factor here was that they identified that TextBlob was unable to read tokenised special characters as a limitation of the module, and factored this into their analysis. Tang et al. (2020) adopted the TextBlob sentiment software within their ConceptGuide system. The sentiment analysis here played a crucial element in the evaluation of their tool and future work proposes a learning efficiency analysis that may use similar principles, exemplifying the utilisation of TextBlob outside of a social media context. These examples showcase how TextBlob can be used to analyse social media data and beyond.

VADER classification module acts in a similar way to TextBlob, using bigrams to attempt to detect negation in syntactical structures. Various studies have ranged from mining emotions from online video comments (Chaithra, 2019a) to looking specifically at sentiment on Twitter. Chauhan, Bansal, and Goel (2018) recognised the changing and challenging formats of tweets might have had an impact on previous work undertaken, and, because of VADER's sensitivity to social media formats, it is a more suitable module to use and yielded more accurate results. Mustaqim et al. (2020) combined VADER's with the k-nearest neighbour algorithm, and found that the two of them together yielded greater insight into the Twitter discourse concerning Indonesian forest fires in 2019. Because of its suitability for social media research, VADER was used by Park, Ciampaglia, and Ferrara (2016) to investigate fashion trends on Instagram. Additionally, VADER has applications beyond social media discourse research. It was used in the study about student perceptions of a virtual teaching assistant in a study by Wang et al. (2021). This highlights the broad applicability of VADER, which may offer valuable insights into the sentiment expressed in Twitter discourses about decision-making algorithms.

Gupta and Joshi (2021) looked specifically at the role of negation in Twitter sentiment analysis, with a focus on negation scope detection and negation handling methods. Their main conclusion is that negation is not a trivial task but entails many challenges, such as implicit negations and negation exception cases. By exploring this within the healthcare domain on Twitter, they have been able to lay foundations for this to be applied to other areas of research on Twitter, stressing the importance of handling negation exception cases, where negation cues act as non-cue, through a deep learning model. In similar fashion, sentiment analysis modules have been critiqued for the lack of consideration for sarcasm. González-Ibánez, Muresan, and Wacholder (2011) investigated this by comparing machine learning techniques to human reviewers and found that neither performed

well in classifying sarcastic tweets. Villena-Román and Garcıa-Morera (2013) built on this when suggesting that sentiment analysis is so complex that humans will often disagree on the sentiment of a text in question. For this PhD thesis, it is important to consider both of these linguistic phenomena when working with automatically classified data, especially when performing a human review.

There have been developments in the standard sentiment analysis models in recent years, with social media researchers going beyond the standard approach. For example, Watanabe (2021) set forth an alternative to sentiment analysis entitled Latent Semantic Scaling (LSS), which used principles of latent semantic analysis (word embeddings) to improve the traditional sentiment model. In addition, Atteveldt, Velden, and Boukes (2021) deployed a survey of many different sentiment analysis techniques and concluded that sentiment dictionaries were not of acceptable standards of validity and, while machine and deep learning outperformed dictionary-based methods the best performance was attained by trained human coding. This further shows that work to combine quantitative and qualitative methods here may be of use. All the studies exemplify that, to make the best use of these computational language analysis methods, they must be combined with other approaches to validate results, such as CL and DA, as in this thesis.

Some studies have begun to offer ideas about how to overcome the limitations of sentiment analysis. Some of these have been technical, such as Ribeiro et al. (2020) evaluating the number of actionable bugs by using an agnostic testing methodology for NLP models. However, some have focused on interpretation. For example. the study by Agarwal et al. (2015) offered the idea of using contextual information alongside sentiment results to better interpret them. This goes alongside other suggestions, such as the research into presenting sentiment over a period of time as a trajectory by Howard (2021). This allows researchers to see how sentiment trends begin to form and develop, thus offering sentiment change as an interpretation strand.

These ideas were important when considering the approach to analysis, and informed the decision to present not only sentiment trajectories, but also topic and emotion trajectories.

#### 3.3.1.3 *Emotion Detection*

Emotion detection from text could be seen as a sister method to sentiment analysis in the sense that it attempts to assign to documents a multidimensional vector representing its emotional valence (resp. sentiment valence) across a set of pre-defined emotion categories (resp. sentiment polarity), based on observation of text. One of the earliest pieces of work from the profile of mood states set forth by Bollen, Mao, and Pepe (2011). They used a psychometric instrument to extract six mood states (tension, depression, anger, vigour, fatigue, confusion) from the aggregated Twitter content and found that social, political, cultural and economic events are correlated with significant, even if delayed fluctuations of mood. In the context of this PhD thesis, incorporating emotion detection alongside sentiment analysis can provide a more comprehensive understanding of the emotional dimensions of the Twitter discussions.

A popular contemporary emotion detection module for Python is EmoLex. The algorithm takes a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), which was manually done initially by crowdsourcing (Mohammad and Turney, 2013). This model has been applied to Twitter investigations, such as the analysis by Aribowo and Khomsah (2021) of Indonesian Twitter users' response to the Covid-19 pandemic. Mathur, Kubde, and Vaidya (2020) again used a similar process to find high levels of trust and fear in tweets relating to Covid-19 from all over the world. Facebook data has also been investigated using EmoLex, with notable examples being the two 2019 studies by Balakrishnan et al. that investigated the emotions in the online diabetes community. EmoLex was refined using string-based Multinomial Naïve Bayes algorithm, with results indi-

cating a 6.3% improvement (i.e. 82% vs. 75.7% for average F-score) when compared to the EmoLex alone (Balakrishnan and Kaur, 2019; Balakrishnan et al., 2019).

The examples here illustrate the benefits of using EmoLex and emotion detection models within a social media context. However, within most of these studies in a social media context, there has been limited focus on the shortcomings of using these methods. Some studies have commented on the need for more complex emotional categories (Jiang, Brubaker, and Fiesler, 2017), which may solve overlap between categories (De Silva et al., 2018; Leung et al., 2021). Also, some have attempted to mitigate human review biases seen when classifying texts as 'neutral' because of the annotator's uncertainty around the best-fit category, rather than it actually being neutral (Fujioka et al., 2019). Additionally, EmoLex was used by Fast, Chen, and Bernstein (2016) to develop Empath, a tool that can generate and validate new lexical categories on demand from a small set of seed terms to use in texts. This offered some critique of EmoLex, as they stated that it correlates imperfectly with Linguistic Inquiry and Word Count analytical procedures. This is an important starting point for a more open discussion on the practical advantages and limitations of using each of these computational linguistic approaches when investigating social media discourses, a consideration pertinent to the exploration conducted in the thesis.

### 3.3.2 *Application*

#### 3.3.2.1 *Topic Modelling*

To prepare the existing data for the analysis in this thesis, the gensim module's 'simple_preprocess' function was used to tokenise the data. Additionally, bigram and trigram models were created using the 'phrases' function in gensim. The process involved generating meaningful bigrams and lemmatising the text using the Natural Lan-

guage Toolkit (Cushing and Hastings, 2009). The id2word dictionary was then constructed by combining the input data with the gensim corpora, assigning a unique ID to each word in the document. Based on this dictionary, a corpus was created, representing the mapping of word IDs to their respective frequencies (Rehůřek and Sojka, 2011). Finally, the topics were generated and displayed using the 'gensim.models.ldamodel.LdaModel' function within gensim. Determining the appropriate number of topics for LDA remains a challenge, prompting researchers to recommend considering the researcher's objectives. A smaller number of topics can provide a broad overview, while a larger number allows for more detailed analysis (Nguyen et al., 2020). All code for the topic modelling can be found in the University of Nottingham Research Data Management Repository.

#### 3.3.2.2 *Sentiment Analysis*

For each study, VADER (Hutto and Gilbert, 2014), a sentiment classification module that detects negation in syntactical structures, was used. VADER has proven effective in analysing sentiment on social media platforms like Twitter (Chauhan, Bansal, and Goel, 2018; Wang et al., 2021). The 'sentiment_analyzer_score' function was utilised, configuring the parameters to classify each tweet as 'positive', 'negative', or 'neutral'. Tweets with a score of 0.05 and above were labelled as 'positive', while those with a score of -0.05 and below were classified as 'negative', according to VADER guidelines (Hutto and Gilbert, 2014). Contextual information alongside sentiment results was incorporated to improve interpretation (Agarwal et al., 2015), whilst also presenting sentiment as a trajectory over time, allowing for the capture of sentiment trends and changes (Howard, 2021). The VADER code used can be found in the University of Nottingham Research Data Management Repository.

In two of the studies, namely Chapters 4 and 5, TextBlob sentiment analysis module was also used to perform part of the analysis (Loria, 2018). The TextBlob model was programmed with a training set about

Covid-19 (Lamsal, 2021) to become familiar with Covid-related lexicon, that appears frequently in the A Level algorithm and Covid-19 app discourses, using the 'train' and 'test' commands. This was not done for VADER as it is a pre-trained analyser. For this, a Comma-Separated Values (CSV) file containing tweets only was imported into the Python library and the command 'blob = TextBlob(sentence)' executed. The TextBlob code used can be found in the University of Nottingham Research Data Management Repository.

#### 3.3.2.3 *Emotion Detection*

EmoLex, a popular Python module for emotion detection, associates English words with eight basic emotions through manual crowdsourcing (Mohammad and Turney, 2013), which was utilised to analyse emotions in the dataset. It has been successfully applied in various Twitter investigations (Aribowo and Khomsah, 2021; Balakrishnan and Kaur, 2019; Balakrishnan et al., 2019; Mathur, Kubde, and Vaidya, 2020) The 'top.emotions' command was employed, exporting a CSV table that showcased each tweet's correlation to various emotions such as fear, anger, anticipation, trust, surprise, sadness, disgust, and joy. Additionally, a separate column was included to label the dominant emotion in each tweet. Additionally, effort was made to mitigate biases in human review when classifying texts as 'neutral' and to address the imperfect correlation between EmoLex and Linguistic Inquiry and Word Count analytical procedures (Fujioka et al., 2019). The EmoLex code can be found in the University of Nottingham Research Data Management Repository.

#### 3.3.2.4 *Best Practices for NLP Tools*

To formalise the process for using the NLP tools, a framework was searched for that would enable engagement with the methods and critically reflect on the use of them in each of the case studies. However, this search was unsuccessful – while many offered frameworks

that were specific to the methods being used, missing elements included how to frame critical reflection within this and ensure its universal relevance to the three chosen methods. As a result of not being able to find a well-defined whole model, existing recommendations for best practice were combined in a process that could be replicated by other social media researchers.

The five steps in this process helped guide the use of each of the methods. This is a combination of borrowed best practices and enables critical reflection through the use of the model by Maclean (2016). This process is illustrated in Figure 3. Once the tool for the analysis was chosen, depending on what is being examined and the aim of the research (Hartmann et al., 2023) (step 0), the steps followed were:

1. **Set expectations:** record what you hope to find in the discourse by using computational linguistic methods. Setting expectations is advocated by Post, Visser, and Buis (2017), who suggest that, by writing down expectations prior to the start of the data collection and analysis, the reflection after this is complete will be much more fruitful.

2. **View as trajectories:** present data chronologically to show which topics are discussed, the sentiment of views expressed or the emotions detected. This is a good place to begin to see patterns and areas of interest in the data. Presenting longitudinal data as a trajectory is advocated by Howard (2021) and compliments how trends can be seen quickly through real-time data collection (Alamoodi et al., 2021).

3. **Human review:** according to similar studies (González-Ibánez, Muresan, and Wacholder, 2011; Maier et al., 2018), it is important to manually human review a sample of the tweets. This offers the opportunity to not only classify the tweets according to the categories defined by each tool but also annotate instances

of potential inaccuracy, such as sarcasm or negation. It was decided that the reviewers would review 10 tweets per period to ensure a fair but manageable sample. The human review was undertaken by two different reviewers and the inter-annotator agreement was calculated.

4. **Examine items of interest with context:** whether they are turning points, extreme polarities or suggest they have been questionably categorised, examining these with contextual data, such as knowledge about events that move the public at the time, may help create more meaning from the results, as per the suggestions of Agarwal et al. (2015).

5. **Conduct formal critical reflection:** formally conduct critical reflection using Maclean's weather model. Use the expectations recorded before using the method to measure its success and suitability for analysis on this occasion.


Figure 3: A diagram to illustrate the borrowed best practices analysis process, first set out by Heaton et al. (2023b).

#### 3.3.2.5 *Critical Reflection Model*

After exploring each of the methods using the process, a critical reflection takes place. Finlay (2008) defines critical reflection as 'learning through and from experience towards gaining new insights of self and practice'. Critical reflection is not a new concept within HCI and social media research (Sengers, McCarthy, and Dourish, 2006) but a greater focus has been placed on the design process of new or developing technology. Within this work, critical reflection was employed to examine how suitable the tools have been to investigate the social media discourses revolving around decision-making algorithms by

applying a simple four-strand critically reflective model outlined by Maclean (2016). Using principles from other models, such as the one by Gibbs and Unit (1988), Maclean provides the following stages:

- Sunshine – what went well?
- Rain – what did not go well?
- Lightning – what came as a shock or surprise?
- Fog – what wasn't understood or could be a further challenge?

Although initially used for educational practitioner reflection, this model allows for simple yet robust critical reflections. This would provide a concise format to present lessons learnt in an accessible form for other social media researchers who are not experts in computational linguistics. Maclean's model's focus on aspects that were surprising or shocking makes it different from most models and presents the opportunity to carve out future plans and work from the final reflective stage. It is important to note that the critical reflection is for the use of the method itself, rather than how successful the process was in aiding the analysis. For the purpose of demonstrating this process, critical reflections will take place after reviewing the results of the method for all three case studies, allowing the drawing of comparisons between the two examples offered. When reflecting, items of particular interest included in the speed, clarity and accuracy of the processes and outputs.

### 3.3.3 *Section Summary*

This section has reviewed NLP-based computational linguistic tools (topic modelling, sentiment analysis, emotion detection) to illustrate their potential to analyse and understand social media discourses. The chosen NLP-based computational linguistic tools stand united in their descriptive and predictive functionality, making them suitable for large-scale linguistic data exploration in social media (Liu, 2010;

Nikolenko, Koltcov, and Koltsova, 2017a; Sailunaz et al., 2018; Vyas and Uma, 2018). This review has shown their potential to unveil latent topics, sentiments and emotional valence within large textual corpora, exemplified by studies scrutinising various discourses on platforms like Twitter (Arianto and Anuraga, 2020; Hidayatullah, Aditya, Gardini, et al., 2019; Prihatini et al., 2018; Srinivasan and Mohan Kumar, 2019; Yang and Zhang, 2018). However, these tools have also faced scrutiny, with studies critiquing their limitations, such as topic stability, difficulty in interpretation and lack of nuance (González-Ibánez, Muresan, and Wacholder, 2011; Maier et al., 2018; Schofield and Mimno, 2016).

Subsection 3.3.2 detailed NLP-based approaches encompassed topic modelling, sentiment analysis, emotion Detection and adopting best practices. For topic modelling, the gensim module's functionalities were used for tokenisation and model generation, with LDA being employed to generate topics (Chauhan, Bansal, and Goel, 2018; Wang et al., 2021). Sentiment analysis utilised VADER and TextBlob modules for classifying sentiment in tweets, considering the context for improved interpretation (Agarwal et al., 2015; Chaithra, 2019b; Park, Ciampaglia, and Ferrara, 2016). Emotion detection employed EmoLex to categorise tweets into eight basic emotions and emphasised the need to address biases in human review (Aribowo and Khomsah, 2021; Fujioka et al., 2019; Mohammad and Turney, 2013). However, despite the comprehensive approach, a formalised framework for NLP tool usage was not found, leading to the creation of a five-step process combining existing best practices and facilitating critical reflection (Hartmann et al., 2023; Heaton et al., 2023b).

The critical reflection model by Maclean (2016), which will be employed post-NLP exploration, involves four stages – sunshine, rain, lightning, and fog -— to evaluate the successes, failures, surprises and challenges faced during method application, aiming to provide insight into strengths and limitations of the NLP-based techniques (Finlay, 2008; Sengers, McCarthy, and Dourish, 2006).

## 3.4 CORPUS LINGUISTICS

### 3.4.1 *Background*

One suitable approach to provide insight is Corpus Linguistics (CL). A corpus is defined as a body of written text or transcribed speech, which can be linguistically or descriptively analysed (Kennedy, 2014). CL takes this idea of further investigating the corpus through a multitude of different analytical tasks. This is the study of language data on a large scale (McEnery and Hardie, 2011) and is concerned with language use in real contexts (Adolphs and Lin, 2011). CL allows for the comparison of multiple corpora (more than one dataset) to identify trends and patterns in texts, which is particularly helpful when comparing data from different time periods, such as in this thesis.

Data is tagged according to the part-of-speech (noun, verb, adjective, etc.). Once this tagging occurs, an analysis process called collocation can begin. Collocation is defined as the co-occurrence of two or more words within a defined word span (Jaworska, 2017). When using frequency as the sole measure, Baker (2006) states that it might not be possible to verify whether a co-occurrence is a true reflection of a semantic relationship or whether chance played a part. Instead, LogDice, a statistical significance measure, becomes a useful indicator of lexical and grammatical associations between textual elements, as well as themes (Mautner, 2007). In this sense, concordances, which display whether a related word appears more frequently before or after an identified 'key' word (Hoey, 2007), help identify collocations as they can show how adjacent or in close vicinity the related words are together. Therefore, concordance lines can display the context surrounding a word of interest (Hoey, 2007).

There are advantages to using CL to analyse social media datasets. According to Jaworska (2017), CL offers ease in how large amounts of data can be automatically scanned to uncover patterns in frequency and keywords. This is echoed by Tognini-Bonelli (2001), who

states that CL allows access to real-world, authentic texts and a high processing speed. Given its efficiency and capacity to process large datasets, CL facilitates diachronic comparisons across corpora through lexical usage (Baker, 2010). Because of its capacity to identify language patterns in large datasets, CL has been frequently deployed to carry out analyses on social media.

Jaworska (2017) also categorises media research involving CL into two strands: the first focuses on structural, pragmatic and rhetorical features of text, and the second on how language shapes representation. Similarly, Nugraha, Sujatna, and Mahdi (2021) concentrated on both whilst investigating a Twitter corpus about the 2020 Charlie Hebdo shootings, and the terrorist attacks on the headquarters of the French satirical magazine. While '#JeSuisCharlie' was used to most frequently express sympathy, '#CharlieHebdo' featured in messages dealing with a wider variety of topics and emotions. Through using keyword and concordance analysis, and building on the previous CL findings of Kopf and Nichele (2018), they found that there were 13 categories of keywords – such as place, the weapon, and the attacker. These categories are connected to each other: for example, many tweets linked the attacker to Islam, his religion, and discussed Pakistan and Islamic culture generally, framed by this incident. These studies all constitute examples of using CL to analyse Twitter discourses of social interest or having an impact on society.

Despite its key advantages, CL can pose analytical challenges with social media data. For instance, Baker and Levon (2015) found that CL, used in isolation, provides a focus on collocation and word frequencies, which is descriptive in functionality, and thus focus is drawn away from interpretation or critique. Rose (2017) also criticises the restricted explainability of CL-derived results, despite the large evidence these could provide. In this sense, the author calls for an integration of CL with other qualitative approaches to ensure more meaningful insights. These recommendations appear supported by Sulalah (2020), who investigated the semantic prosody of 'increase'

in Covid-19 discourses. Additionally, Liimatta (2020) states that CL analysis can be problematic when dealing with short texts because of its normalised counts – usually calculated on a base of either 1,000 or 10,000. The calculations could generate unreliable values when applied to very short texts – such as tweets – due to the excessively small lexical samples these allow to consider. As a result, very short texts, which are especially common on certain social media platforms, should be interpreted carefully when compared.

### 3.4.2 *Application*

Corpus Linguistics (CL) was chosen for its ability to analyse large datasets efficiently and uncover language patterns (Jaworska, 2017). CL offers various language-focused perspectives, including diachronic comparisons focusing on lexical usage (Baker, 2010). Its effectiveness in revealing language patterns in substantial datasets (Anthony, 2013; Hunston, 2010; Kopaczyk and Tyrkkö, 2018) makes it a common choice for social media analysis (Kopf and Nichele, 2018; Nugraha, Sujatna, and Mahdi, 2021; Russo and Grasso, 2022). CL also facilitates the comparison of multiple corpora, making it suitable for analysing data from different time periods, such as in the datasets used in this thesis.

In this analysis process, CL-computerised tools were used to examine collocation, the co-occurrence of words within a defined span (Jaworska, 2017). Rather than relying solely on frequency, it is possible to use statistical significance measures like LogDice and Log Likelihood to identify lexical and grammatical associations and themes (Mautner, 2007). Concordances, showing the proximity of related words, helped identify collocations (Hoey, 2007).

The CL software used for this thesis was The Sketch Engine (Kilgarriff et al., 2008), chosen for its practicality and availability to academics. It also provided a series of reference corpora for comparison.

The analysis consisted of several stages. First, keyword analysis identified keywords in the discourse, comparing them to the enTenTen20 reference corpus (Suchomel, 2020), chosen due to being comprised solely of internet texts, including social media texts. Keyness scores, a value generated by comparing the word frequency in a target corpus compared to an appropriate reference corpus (Jaworska, 2017), were generated, offering an overview of the lexicon in the analysed tweets.

Next, concordance lines with key search terms that were suspected as potential social actors, related to each individual discourse, were examined to initiate collocation analysis. Active and passive constructions were explored. LogDice was chosen as the statistical measure of collocational strength as it not only measures the statistical significance of a collocation but also factors in the size of the subcorpus, making comparisons between subcorpora of different sizes easier. LogDice compared the observed co-occurrence of words to their expected co-occurrence based on their individual frequencies. A high LogDice score indicated a strong association between two words, suggesting they often appear together, while a low score implied a weaker association. The strongest collocates for each period were reported based on LogDice scores, with a minimum threshold of three occurrences to determine significance. The top ten words with the strongest collocations from each time period were analysed using DA, which will be explored methodologically next.

### 3.4.3 *Section Summary*

This section has reviewed Corpus Linguistics to illustrate its potential to analyse and understand social media discourses. Firstly, subsection 3.4.1 has showcased Corpus Linguistics (CL) as an additional approach concerned with the analysis of vast textual data to uncover patterns, themes and trends. CL allows for comparison across cor-

pora, aiding diachronic analyses and facilitating investigations within social media research contexts (Baker and Levon, 2015; Hoey, 2007; Jaworska, 2017; Kennedy, 2014). However, CL presents challenges with short texts and may prioritise descriptive aspects over critical interpretation, necessitating integration with qualitative methods for more meaningful insights (Liimatta, 2020; Rose, 2017).

Subsection 3.4.2 details how the CL tool The Sketch Engine, chosen for its efficiency in large dataset analysis, will be used to uncover language patterns through keyword analysis and collocation exploration (Baker, 2010; Jaworska, 2017).

## 3.5 DISCOURSE ANALYSIS

### 3.5.1 *Background*

Considering the challenges posed by CL, discussed in the previous section, Discourse Analysis (DA) was chosen as a complementary approach. Whilst CL analysis tools struggle to pinpoint different perspectives and meaning shades, DA examines texts for nuance and pragmatic opinion (here meaning an examination of implied meanings of language). Therefore, these approaches were deemed especially effective together in exploring trust and blame in decision-making algorithm Twitter discourse.

Discourse surpasses the sentence boundaries (Schiffrin, 2001) and comprises language stretches that are interlinked and create meaning, thus they carry an inscribed sociolinguistic value (Cook, 1989). In this sense, questioning the social significance of language can uncover how it influences — and is influenced by -– the world around us (Johnson, McLean, and Kobayashi, 2020). Therefore, Discourse Analysis studies language use beyond the sentence level, examining how meaning is constructed through connected speech or written text in social contexts (Brown, 1983). It investigates patterns of language in

use and the relationship between language and socio-cultural contexts. As Gee (2014) notes, Discourse Analysis examines how language enacts social and cultural perspectives and identities. Therefore, DA is an interpretative qualitative approach to text analysis that draws upon related theoretical frameworks.

In fact, there are several focal points that can be adopted when approaching DA and Reeves, Kuper, and Hodges (2008) label them as descriptive, empirical and critical. While descriptive addresses solely how language and grammar work together to create meaning in isolation, empirical and critical variations account for context and even include it as part of the data collected from discourses. Empirical analysis has been used successfully in studies where there is still a microanalytical focus on language. However, critical analysis places even greater emphasis on contextual information through macroanalysis, which focuses on the power and perspectives of individuals and institutions. As this is relevant to the aims of this thesis, the DA used here will be influenced by Critical Discourse Analysis (CDA).

CDA specifically examines how discourse reproduces and maintains power relations, social inequalities and dominance in society (Fairclough, 1993). It aims to reveal hidden power dynamics and ideologies embedded in language use. Van Dijk (1997) emphasises that CDA focuses particularly on how discourse structures enact, confirm, legitimate or challenge relations of power and dominance in society. As a result, CDA can be used as a tool to better understand meanings implied by the context of a text or series of texts (Van Dijk, 1997). Fairclough (1993) identifies three CDA layers: micro, meso and macro. The micro analysis examines syntax (sentence construction), metaphorical meanings and rhetoric. The meso analysis looks at the interpretation of the relationship between discursive processes and the text. The macro analysis examines the explanation of the relationship between the discourse and the socio-cultural context.

Another significant contribution with regards to contextual meanings of the discourse is put forward by Van Dijk (2001), who offers

a socio-cognitive perspective. Accordingly, discourse can be viewed as socially shared representations of societal arrangements, as well as interpreting, thinking, arguing, inferencing and learning. Although different, the two contributions are similar in regards to transitivity (Amoussou and Allagbe, 2018). For example, an examination of transitivity patterns may uncover who is acting as the agent – thus, performing the action – over whom and whether passive verbal constructions exclude and background social actors. Therefore, this shows existing studies employing CDA show that a specific focus on the agency is possible to unveil blame and responsibility.

Despite CDA's relevance to this exploration of agency, trust and blame, it is important to note, however, that the DA employed in this thesis is not always CDA. As CDA focuses heavily on the political implications of power imbalances, this did not necessarily fall within the bounds on this PhD investigation. CDA aims to deconstruct dominant discourses and expose the underlying ideologies that shape social practices and power relations (Fairclough, 1993; Van Dijk, 1997). This critical perspective is not the primary focus of the thesis, which is more concerned with objectively understanding how agency is attributed to decision-making algorithms. Therefore, the DA used does draw upon – and is influenced by – the principles of CDA, yet not enough to make it the lone type of DA approach used. While CDA offers valuable insights into the political dimensions of language, its explicit focus on power and ideology might not be the most suitable sole approach for this particular thesis.

As previously outlined in 2.1, transitivity analysis in linguistics examines how agency is emphasised, manipulated or concealed in text, considering active and passive voice and nominalisation (Oktar, 2001; Richardson, Mueller, and Pihlaja, 2021). DA can be used to examine how an agent, defined as an entity with internal energy to perform actions (Leslie, 1993), can be obscured through passive constructions, potentially shifting responsibility (Clark, 1998). DA assists the exploration of how agency is also conveyed through lexical choices and

metaphors (Morris et al., 2007). These linguistic features can reveal attitudes towards automated decision-making algorithms when analysed in context.

Metaphors have also been used to personify inanimate entities and increase the dramatic effect and intensity of a statement (Goatly, 2007). Additionally, vocabulary can be examined to unearth how words are used to show ideology, including the use of euphemisms and metaphors. It is also important to factor in how implicit information can be inferred and deduced through the examination of these aspects of language. Given its relevance, this work will use transitivity and agency as a focus of the analysis.

Similar studies have used DA to examine Twitter data, whilst addressing other social aspects such as gender and origins. Among them, Aljarallah (2017) investigated perspectives on women driving in Saudi Arabia, finding specific hashtags that supported or opposed women driving. Their results showed, among others, that tweets with the hashtag #Womencardriving presented significant support towards the movement. However, opposing reactions emerged from the hashtags #Iwilldrivemycar and #Iwillentermykitchen. In another study by Sveinson and Allison (2021), representations of gender and stereotyping have also been explored, including an overwhelming dislike for hyper-feminised items marketed to women and girls through detailed linguistic analysis. This study demonstrated that clothing serves as more than just a reflection of consumer preferences, as it can also embody the cultural identity of an organisation. Also, Kreis (2017) investigated the hashtag #refugeesnotwelcome, unearthing that users deployed a rhetoric of inclusion and exclusion to depict refugees as unwanted, criminal outsiders. Her findings showed that this discourse reflected a prevailing political climate in Europe, where nationalist-conservative and xenophobic right-wing groups were gaining influence and promoting a discourse that is prominent on social media. Overall, these studies demonstrate the

benefits of using DA on Twitter discourses specifically, highlighting the depth of understanding that it can uncover.

Notably, DA brings several advantages as it can reveal unacknowledged aspects of human behaviour and support new or alternative positions on social subjects (Mogashoa, 2014; Morgan, 2010). In this sense, DA is naturally interdisciplinary (Wodak, 2007) and requires an abductive approach, where a symbiotic relationship between theory and empirical data is necessary (Mogashoa, 2014). As DA and CDA examine the intricate relationships between text, social opinion, power, society and culture, it provides a lens to better understand urgent social implications (Van Dijk, 1997). Additionally, the incorporation of an epistemological aspect into DA means that, while the researcher brings their own beliefs and perspective, reflection upon findings has its place within the approach. Bucholtz (2001) claims this to be reflexivity with a heightened self-consciousness. Therefore, DA – influenced by CDA – is an appropriate choice to explore social action, blame and agency, as in this thesis.

As with any methodological approach, DA has shortcomings, too. Firstly, it requires considerable effort and time required to perform DA on a large dataset (Wetherell and Potter, 1988). Additionally, the subjective nature of DA, approaching data with a personal perspective and lens, may limit its validity and decrease the objectivity and applications of the findings (Gill, 2000). Therefore, combining findings from computational linguistic analysis may help mitigate this. Also, Morgan (2010) notes DA is not fixed and is always open to interpretation and negotiation. The lack of objective measures available to analysts may result in inaccurate or misrepresentative findings. This complements the view of Olson (1995) that it is not a 'hard science' and more of an insight through examination and discussion.

These shortcomings provide a rationale for using CL with DA to increase processing efficiency. This also aids the mitigation of the potential subjectivity of DA: using a semi-automated approach first means comparisons can be organised according to the research focus. Al-

though combining CL and DA does not grant ultimate objectivity, it is less prone to exclusive subjective analysis.

### 3.5.2 *Application*

Subsequently, Discourse Analysis (DA) was employed. DA, as an interpretative qualitative text analysis method, draws on relevant theoretical frameworks (Hart, 2008; Johnson, McLean, and Kobayashi, 2020; Kendall et al., 2007). It aided in discerning implied meanings within the textual context (Bloor and Bloor, 2013; Tenorio, 2011; Van Dijk, 1997). This approach aligns with this data-driven strategy for addressing research questions regarding all three systems presentation and their impact on society, a method well-demonstrated in numerous studies involving Twitter discourses (Aljarallah, 2017; Kreis, 2017; Sveinson and Allison, 2021), showing that a specific focus on agency fosters the identification of blame and responsibility.

By combining DA with the results from the Sketch Engine CL-analysis tool (Kilgarriff et al., 2008), the agency and social action conveyed in concordance lines featuring individual search terms was able to be explored. This collaborative approach has shown its effectiveness in similar studies (Abbas and Zahra, 2021; Baker, 2012; Brookes and McEnery, 2020; Nartey and Mwinlaaru, 2019). Baker (2012), examining the representation of Islam and Muslims in the British press, saw that corpus-driven procedures revealed patterns such as Muslims being linked more often to extreme belief than moderate or strong belief. These findings underscore the need for critical interpretation alongside quantification. Meanwhile, Abbas and Zahra (2021) explored the role of Twitter in political campaigns, focusing on tweets from Donald Trump and Joe Biden during the 2020 US presidential election. They identified themes related to Covid-19 policies, environmental issues, racial unrest and others. This research sheds

light on the discursive practices and political ideologies of the candidates through a combined CL and DA approach.

Contemporary research continues to use a corpus-based DA approach (Boucher et al., 2024; Brookes, 2023; Brookes and McEnery, 2020; Deignan and Love, 2021; Love, 2021; Semino et al., 2024). Specifically, for example, Brookes and McEnery (2020) employed correlation statistics to examine how violent jihadist texts appropriate and redefine Islamic terminology. Their analysis revealed dependent relationships between terms that extend beyond simple collocation into broader forms of textual cohesion. Their analysis demonstrated how these cohesive patterns vary between violent and non-violent texts and showed that linguistic manipulation functions as symbolic capital through which extremist groups construct identities that legitimise violence. Moreover, Semino et al. (2024) analysed Mumsnet Talk threads about HPV vaccination, comparing responses to stories versus factual posts. Their corpus-driven approach found that stories received more supportive engagement, whilst factual posts triggered more challenges. Additionally, Love (2021) used a corpus-based DA approach to compare swearing patterns between the 1994 and 2014 Spoken British National Corpus by examining 16 swear words across demographic variables. He found that the results showed an overall decrease in swearing by 2014 with 'fuck' replacing 'bloody' as the most common swear word. Traditional patterns persisted regarding gender and age distribution, yet socio-economic patterns proved more complex than anticipated. Most crucially, Nartey and Mwinlaaru (2019) presented a meta-analysis of 121 studies using corpus-based DA, analysing their chronological development, domains of engagement, topical issues and regional coverage. Overall, all these studies exemplify that corpus-based DA offers a robust methodology for analysing discursive reflections of social issues.

### 3.5.3 *Section Summary*

This section has reviewed Discourse Analysis to illustrate its potential to analyse and understand social media discourses. DA examines the nuances, pragmatic meanings and social implications of language usage, exploring transitivity, agency and contextual meanings within texts (Cook, 1989; Fairclough, 1993; Leslie, 1993; Schiffrin, 2001; Van Dijk, 1997, 2001). This approach has unveiled societal perceptions and ideologies reflected in social media discourses, from gender representations to perspectives on social and political events (Aljarallah, 2017; Kreis, 2017; Sveinson and Allison, 2021). Nevertheless, DA exhibits limitations, such as the time-intensive nature of large datasets and the subjectivity inherent in interpretative analyses (Gill, 2000; Wetherell and Potter, 1988). Hence, it may be possible to analyse the chosen discourse with a combined approach of computational linguistic (both NLP-based and CL-based) analysis with DA to enhance efficiency, reduce subjectivity and enable methodological triangulation (Morgan, 2010; Olson, 1995). This integration may provide an avenue for more robust, nuanced and insightful analyses of social media discourses by amalgamating the strengths of computational linguistic tools with the interpretative depth of DA.

Subsection 3.5.2 illustrates how DA was employed in conjunction with the CL results to explore agency and social action conveyed in the discourse (Johnson, McLean, and Kobayashi, 2020; Van Dijk, 1997). Moreover, the use of Social Actor Representation (SAR) within DA delved into grammatical structures' role in conveying social agency, attributing responsibility through linguistic features (Morris et al., 2007; Van Leeuwen, 2008; Weber, 1978). The primary rationale is to uncover themes in decision-making algorithm presentations, elucidating power dynamics within opaque operational mechanisms (McGlashan, 2020; Razis, Anagnostopoulos, and Saloun, 2016). This comprehensive approach facilitated a nuanced understanding of social media

discourses and their presentations in the context of decision-making algorithms (Bernard, 2018; Kitishat, Al Kayed, and Al-Ajalein, 2020).

## 3.6 CHAPTER SUMMARY

The chapter has presented the comprehensive analytical approach adopted in this thesis, starting in section 3.1 with the approach to data collection using the Twitter for Academic Purposes API, chosen due to Twitter's data availability and real-time analysis potential. The chapter highlights the ethical considerations and the need to anonymise data to align with best practices in social media research.

Subsequently, the section 3.2 details the integration of the individual approaches into the overarching methodology of the thesis. It elaborates on how NLP-based tools were used to capture initial data trajectories, followed by CL analysis and DA, providing insights into grammatical structures, social agency and power dynamics within the Twitter datasets. This section also discusses the pivotal role of interdisciplinary research in addressing complex issues like those surrounding decision-making algorithms. The benefits of interdisciplinary research, such as collaborative frameworks and innovative solutions, are discussed alongside the challenges, including communication gaps between disciplines and funding barriers.

Further to this, each of sections 3.3, 3.4 and 3.5 reviewed various computational linguistic tools like NLP-based approaches, Corpus Linguistics and Discourse Analysis, delineating their strengths and limitations. These approaches together offer the ability to analyse large datasets (NLP-based tools), uncover language patterns (CL), and explore nuances and social implications within texts (DA). The integration of these approaches is proposed for more robust analysis, capitalising on computational tools' strengths and the interpretative depth of DA.

Overall, the chapter outlines a comprehensive analytical roadmap that amalgamates diverse approaches to explore the complexities of the specific Twitter discourses on decision-making algorithms. This multifaceted approach aims to provide nuanced insights and understandings that bridge disciplinary boundaries.

# Part II

## EMPIRICAL WORK

This part contains the results from the analysis of the Twitter discourses pertaining to the selected decision-making algorithms case studies. Chapter 4 is concerned with the 2020 A Level algorithm, chapter 5 features the Covid-19 contact-tracing app and chapter 6 investigated ChatGPT.

# 4

# 2020 A LEVEL ALGORITHM

## 4.1 STUDY BACKGROUND

As previously explained in Chapters 1 and 2, blame and agency in relation to automated decision-making is an emerging topic in academia (Floridi et al., 2018; Mayer, Davis, and Schoorman, 1995; Wagner, 2019). Although currently under-explored, studying this has shown to be important when forming interventions for when decision-making algorithms do not perform in a way that the public expects of them (Olhede and Wolfe, 2020). A recent example of this is the case of the 2020 A Level algorithm in England and Wales, where examinations during the Covid-19 pandemic were replaced by automatically calculated grades, based on factors such as the historical performance of the institution in previous years' exams (Kelly, 2021). Although initially defended, the algorithm-decided grades were abolished and teacher assessment grades were used instead due to an outpouring of public dismay (BBC, 2020; Jiang and Pardos, 2021).

Although public perspectives about the A Level algorithm have already been collected (Bhopal and Myers, 2020; Hecht, 2020; Kelly, 2021; Kolkman, 2020), there is a research gap regarding public views expressed on the social media website Twitter, which could provide valuable data, as it hosts a plethora of opinions relating to current affairs (McCormick et al., 2017; Weller et al., 2013). Therefore, addressing this research gap could offer a fuller and more detailed picture of the wider public's response to the event.

This chapter uses the combined approach of Natural Language Processing (NLP), Corpus Linguistics (CL) and Discourse Analysis (DA). This quantitative and qualitative analysis is underpinned by Social

Actor Representation (SAR), a branch of Social Action Theory (SAT), where grammatical and transitivity structures play a crucial role in the representation of social actors (Van Leeuwen, 2008). More specifically, transitivity analysis – the examination of active and passive agents in texts – may uncover who is reported or depicted as the agent over whom and how passive verbal constructions impact social actions, possibly deleting or masking social actors. There are various SAR techniques that indicate whether an agent in a text constitutes a social actor, including *exclusion*, *backgrounding*, *individualism*, *assimilation*, *personalisation* and *impersonalisation*, which will all be explored. Thus, using SAR is helpful when examining blame and responsibility in discourse, as this thesis intends to do.

This PhD research is based on the assumption that combining NLP, CL and DA to Twitter discourses can mitigate some of the potential shortcomings of NLP-based computational linguistic tools that usually are utilised for social media research (Mogashoa, 2014). This is due to the high emphasis on context and how language is used, underpinned by SAR. In fact, studies into Twitter discourses using these methods have yielded insightful and meaningful results on women driving in Saudi Arabia (Aljarallah, 2017), refugees (Kreis, 2017) and the dislike for hyperfeminised items being marketed to women and girls (Sveinson and Allison, 2021). These examples showcase how this approach can be used effectively in the wider context of social media research to achieve greater understanding of social phenonema, which will be examined in this chapter, whilst adding a specific focus on blame and agency.

This chapter intends to address whom Twitter users blamed for the disruption to A Levels. Ultimately, using this combined approach will add to the current discourse regarding which entities have been blamed for the algorithm's failure, particularly illuminating ideas about how social media users reacted to the scandal and the social repercussions of enforced algorithmic implementation.

Summarising, this chapter will use NLP-based approaches, CL and DA to examine how blame is implied in relation to automated decision-making, through agency and transitivity, in Twitter discourses regarding the A Level algorithm. From a practical perspective, the entities will be identified through the aid of SAR. From a theoretical perspective, complementing NLP-based computational linguistics with CL and DA will illustrate the proposed combined language analysis approach outlined in Chapter 3.

### 4.1.1 *Study Research Question and Objectives*

The sub-research question for this chapter is as follows:

> What insights into agency, trust and blame in the Twitter discourse surrounding the 2020 A Level algorithm can be achieved through combining language analysis approaches?

In turn, the following objectives will be addressed:

1a Demonstrate how Natural Language Processing techniques (sentiment analysis, topic modelling and emotion detection) provide insight into Twitter discourses surrounding the 2020 A Level algorithm.

1b Demonstrate how Corpus Linguistics, particularly collocation, provides insight into public Twitter surrounding the agency of the 2020 A Level algorithm.

1c Demonstrate how Discourse Analysis provides insight into Twitter discourses surrounding the agency, trust and blame of the 2020 A Level algorithm.

1d Identify the strengths and limitations of using the three approaches to investigate Twitter discourses surrounding the 2020 A Level algorithm.

## 4.2 STUDY APPROACH

### 4.2.1 *Data Collection and Processing*

As previously mentioned, using Twitter has allowed the collection of a large, readily available dataset. Twitter data can be processed before analysis (Jianqiang, 2015), lending itself well to exploratory analyses (Chong, Selvaretnam, and Soon, 2014). For convenience, data was collected using the Twitter for Academic Purposes Application Programming Interface (API) and Tweepy (Roesslein, 2009). It was ensured that the collection and analysis method complied with the terms and conditions for the source of the data and the API. The data were sourced from the United Kingdom and only tweets in English were selected, meaning the analysis investigated views expressed in English only. Since retweets indicated agreement or support, duplicate tweets were expected, although eliminated from the corpora not to bias counts.

For the specific tweets used in this dataset, the 18,239 tweets composing the dataset were published from 12th August 2020, the day before A Level results were released to students, until 3rd September 2020, after Ofqual's chair appeared at the Education Select Committee. Tweets containing 'Ofqual algorithm', 'ofqualalgorithm', 'A level algorithm', 'alevelalgorithm', 'a levels algorithm', 'a-level algorithm' or 'a-levels algorithm' were gathered. These search terms were chosen on the basis of their relevance to the algorithm, rather than the A Level results in general. The tweet IDs and other associated information can be found in the University of Nottingham Research Data Management Repository.

### 4.2.2 *NLP-Based Techniques*

#### 4.2.2.1 *Topic Modelling*

The topic modelling technique chosen was LDA and the gensim module was chosen to perform this. The existing data was tokenised using gensim's 'simple pre-process' function and bigram and trigram models were created using the 'phrases' function. The process also allowed bigrams to be made and lemmatised using Natural Language Toolkit (Bird, Klein, and Loper, 2009). The id2word dictionary was inputted and combined with the gensim corpora to create the dictionary specific to this dataset. Each word in the document was given a unique ID. Then, using the dictionary, the corpus, which was a mapping of the word ID and the word frequency, was created (Wang et al., 2020). The topics were then generated and printed using the 'gensim.models.ldamodel.LdaModel' function within gensim.

#### 4.2.2.2 *Sentiment Analysis*

Two different sentiment analysis approaches were chosen – one that used TextBlob and the other using VADER. This was done to make comparisons between the two systems – especially with the overt claim that VADER accounts for negation within its algorithm. The TextBlob model was programmed with a training set about Covid-19 (Lamsal, 2021) to become familiar with the Covid-related lexicon using the 'train' and 'test' commands. This was not done for VADER as it is a rule-based analyser. For TextBlob, a CSV file containing tweets only was imported into the Python library and the command 'blob = TextBlob(sentence)' executed. For VADER, the function was 'sentiment_analyzer_score' and parameters were set so that each tweet could labelled 'positive', 'negative' or 'neutral' – with 0.05 and above being 'positive' and -0.05 and below being 'negative', as explained in 3.3.2.2.

#### 4.2.2.3 *Emotion Detection*

The emotion detection module run was EmoLex, also in Python. The 'top.emotions' command was deployed, Which exported a CSV table with each tweet's assignment to each emotion – fear, anger, anticipation, trust, surprise, sadness, disgust and joy – declared, along with labelling the dominant emotion in a tweet in a separate column.

### 4.2.3 *Corpus Linguistics and Discourse Analysis*

The next step concerned CL. Using the CL software The Sketch Engine (Kilgarriff et al., 2008), a keyword analysis was conducted to investigate frequently featuring social actors. The reference corpus used was the English Web 2020 (enTenTen20) (Suchomel, 2020), which comprises 36 billion words of internet texts. Since it contains texts from social media, this was believed to be a suitable reference corpus for this study. Comparing the collected corpus to the reference corpus was used to generate a keyness score, which was calculated by comparing the frequency of the words in the target corpus to the frequency of the words in the reference corpus. Secondly, concordance lines featuring potential social actors were examined to prompt the collocation analysis. This included using LogDice as a statistical measure of collocational strength. Thirdly, DA was used to examine agency and blame as expressed in the concordance lines, where the selected keywords appeared in context.

Additionally, the focus was placed on transitivity, through the examination of social actors in sentence structures, vocabulary choice and the use of metaphor and possession. Specifically, principles of Leeuwen's SAR (Van Leeuwen, 2008) were employed to provide insight into these social representations. Therefore, items of interest that could be related to blame, agency and social action were examined through the collocation analysis of their concordance lines.

## 4.3 NLP-BASED TECHNIQUES ANALYSIS

This section details the findings from the NLP-based techniques – notably, topic modelling, sentiment analysis and emotion detection. It follows the step-by-step best practice process that was outlined in 3.3.2.4.

### 4.3.1 *Topics*

#### 4.3.1.1 *Expectations and Initial Findings*

For this case study, the expectation of applying these methods was to, once again, see whether there were any broad themes within this online discourse relating to the Ofqual algorithm. It was anticipated that there would be a smaller number of latent topics in comparison to the other case studies due to the reduced time frame and dataset size. It was believed that the lexical items would clearly indicate the overarching topic labels.

Using gensim LDA, four latent topics were discovered through gensim LDA, each with ten key lexical items that are associated with that topic, presented in table 2. Once again, the number of topics was decided on through manual topic inspection and regeneration. Initially, there were three topics, but it was concluded that a fourth was necessary due to one topic containing many generic words and dominating the discourse, as per the suggestions of Nguyen et al. (2020). All topic findings are available in the University of Nottingham Research Data Management Repository.

Topic trajectories are presented in Figure 4.

One of the most featured words of the most prominent topic is 'government', which may foreground the importance of the role of the UK government in the decision to use and then withdraw the algorithm. 'Flaw' is the most featured word in Topic 2, potentially highlighting

Table 2: Ranking of the top 10 lexical items associated with each latent topic

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|---|
| 1 | level | flaws | exam | students |
| 2 | government | statistics | gcses | levels |
| 3 | data | father | fiasco | schools |
| 4 | algorithms | foresaw | mutants | teachers |
| 5 | williamson | punishment | boris | grade |
| 6 | like | exams | johnson | school |
| 7 | wrong | labour | news | government |
| 8 | education | unlawful | blames | downgraded |
| 9 | blame | controversial | bbc | based |
| 10 | gavin | williamson | ofqual's | teacher |

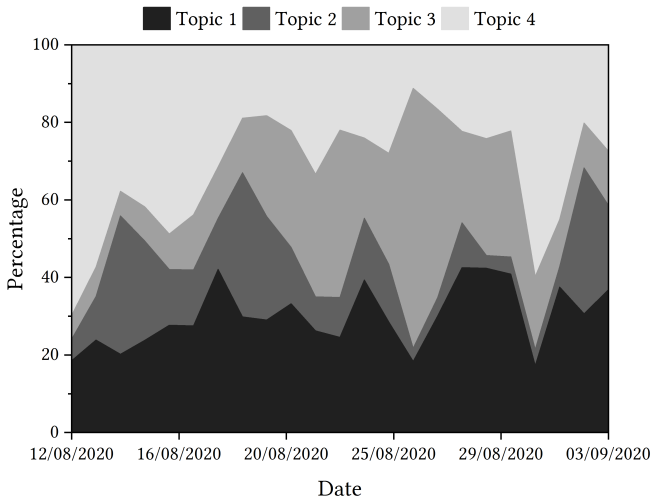


Figure 4: Trajectories of topics detected in tweets relating to the A Level algorithm.

there is a significant amount of discussion about whether the algorithm was fit for purpose.

When looking at how the trajectories of the topics fared over the sample period, Topic 4, with words such as 'students', 'schools', 'teachers', etc., dominated the discourse initially. This aligns with the first reporting of the story, where lexical items such as these may have been popular. It might also point to intense focus on the feelings of students, teachers and schools affected by the algorithm. The topic was not predominant again until the end of the month, although, when examining the contextual factors of the algorithm in mainstream media, there is no clear reason why this may have happened.

Discussion around the government and officials within the Department for Education, including Gavin Williamson, as seen in Topic 1, proves popular throughout the sample period according to the data. However, this is very different to how Topic 2, including lexical items such as 'flaw' and 'statistics', is discussed. This topic is popular at the start of the sample period but declines in proportional popularity until the Educational Select Committee at the beginning of September. Instead, the discourse shifts to the discussion of Topic 3, featuring words such as 'fiasco' and 'mutant'.

#### 4.3.1.2 *Topic 1: Government*

The trajectory of Topic 1 revealed subtle fluctuations in its prominence throughout the sample period, marked by words such as 'government', 'Williamson' and 'education'. This topic revolved around governmental involvement and decision-making regarding the algorithm's implementation and eventual withdrawal.

Throughout August, Topic 1 experienced fluctuations in its prevalence. Initially, it maintains a relatively moderate level of prominence, but as the month progresses, noticeable spikes in its frequency occur, particularly in mid to late August. It might be thought that these peaks coincided with significant events and developments related to

the algorithm, such as the substantial increase in Topic 1's prominence on August 24th. However, limited significant contextual information specific to this date could be found.

#### 4.3.1.3 *Topic 2: Technical Flaws*

Topic 2 illustrated fluctuations in its proportional popularity. Initially, Topic 2 garnered attention with discussions centred around 'flaws' and 'statistics' associated with the algorithm. This suggests a focus on the statistical aspects and potential shortcomings of the algorithm. This resonates with the lexical items linked to Topic 1, particularly 'government' and 'data', indicating an early emphasis on the technical intricacies and decision-making processes concerning the algorithm within governmental spheres.

Initially, Topic 2 had a relatively low proportion of tweets compared to other topics, indicating that algorithmic flaws might be of less importance. However, it gradually increased in prominence, reaching its peak on 14th August, the day after results were released, and maintaining relatively high proportions until 20th August, after the teacher assessment grades were instated. This rise in popularity coincides with heightened discussions about flaws and statistical intricacies related to the algorithm, suggesting increased scrutiny and attention to these aspects during this period. Interestingly, Topic 2 experienced a decline in proportion after this even though discussions about statistics and flaws persisted. This fluctuation could be attributed to shifts in focus within the broader discourse surrounding the algorithm.

#### 4.3.1.4 *Topic 3: Johnson's Response*

The trajectory of Topic 3 underwent discernible shifts in prominence. Initially, this topic registered a relatively modest level of significance compared to other topics. However, its importance gradually ascended in the ensuing days, notably peaking on 26th August 2020, coinciding with the trajectory of Topic 3, as revealed by the provided

data, unfolded as a compelling narrative of fluctuations over the observed period. This topic, characterised by terms such as 'exam', 'gcses', 'fiasco', and 'mutants', according to the lexical analysis, underwent discernible shifts in prominence, reflecting evolving public sentiment or media coverage surrounding educational assessments.

Initially, on 12th August 2020, Topic 3 registered a relatively modest level of significance compared to other topics, recording a value of 0.0278. However, its importance gradually ascended in the ensuing days, notably peaking on 26th August 2020 with a value of 0.3829. This topic is seen to gain in popularity after the government u-turn on the use of the algorithm and is very prevalent in tweets up until the end of August. Upon the announcement of Boris Johnson labelling the algorithm as 'mutant', this became the overwhelmingly most discussed topic according to the data. The labelling of the event as a 'fiasco' as the story gained in media popularity may have been a contributing factor in the rise of this latent topic, which was reflected in the language seen in the human review sample tweets too.

Subsequent to this peak, Topic 3 experienced a somewhat subdued trajectory, although it maintained a relatively elevated status compared to its initial levels. Notably, on 31st August 2020, there was a dip in the significance of Topic 3, aligning with a resurgence in discussions pertinent to Topic 1, particularly concerning governmental actions and decision-making processes as the attention of the discourse turned to the Educational Select Committee.

#### 4.3.1.5 *Topic 4: Impact on Education*

The final topic, closely associated with discussions in education, indicated a particular focus on the ramifications of the algorithm on educational institutions and stakeholders. Fluctuations in the prominence of Topic 4 throughout August and early September became apparent. Initially, Topic 4 commanded a significant presence in the discourse. This mirrored the initial reporting of the algorithm issue,

where concerns regarding its impact on students, schools and teachers predominated discussions.

As the days progressed, the trajectory shifted. For example, the day after the results were released, there was a notable decline in Topic 4's prominence, which coincided with a surge in Topic 2. This shift suggested a temporary diversion of attention from educational concerns to the technical aspects of the algorithm. However, once the algorithm-calculated grades were abolished, attention turned back to this topic. Despite these slight fluctuations, Topic 4 maintained a notable presence. This consistency underscored the enduring significance of education-related discussions in the context of the algorithm

#### 4.3.1.6 *Human Review and Critical Reflection*

10 tweets were randomly sampled for each day of the discourse (230 tweets total). The human reviews matched the automatically assigned topic 36% of the time. Inter-annotator agreement was 0.476, again indicating moderate agreement. A reason behind this might have been the separation between two topics that occurred in the early stages of the process. The most common error seen here was the human reviewer labelling a tweet associated with Topic 1 as Topic 4 (and vice-versa). This low classification accuracy when comparing the human reviews to the automated topic might indicate that some of the topics are unclear or not fit for purpose.

The critical reflection for this was as follows:

SUNSHINE Upon reflection, the positive aspects of using LDA and topic modelling for both datasets have been the ability to see which lexical items appear frequently with one another, in an attempt to discover latent topics. The gensim tool for topic modelling was also easy to use and compatible with both datasets. Additionally, combining this method with the context of the chosen case studies illuminated topics for potential further analysis and follow-up.

RAIN Using gensim's LDA topic modelling, there were limited specific guidelines on how the output (e.g., the lexical items) should be interpreted. As a result, this interpretation means it is important to consider what the topics might be. This means that comparing these findings to those conducting similar studies may be more difficult.

LIGHTNING A surprise that occurred through using LDA was the recurring lexicon that appeared in different topics. This foregrounds the importance of context in this process.

FOG An aspect that caused slight confusion through using gensim's LDA is, once again, the interpretation strand. Although interpretation is encouraged after compiling a lexicon associated with topics, it is difficult to pin down how interpretation occurs. Given that LDA is an automated process based on frequencies, a challenge is how a human interprets the results to create meaning.

### 4.3.2 *Sentiment*

#### 4.3.2.1 *Expectations and Findings*

The aim of using the sentiment analysis techniques was to see the general feeling of sentiment detected in the discourse and how co-occurring events may have impacted sentiment. This sentiment may manifest in whether individual parts of the discourse are predicted to be positive, negative or neutral.

From the TextBlob sentiment analysis, Figure 5 shows that overall sentiment ranged from 0.088 to -0.052, indicating that overall sentiment is neutral. However, from the VADER sentiment analysis, Figure 5 shows that overall sentiment ranged from 0.03 to -0.5, indicating that overall sentiment is negative. All sentiment findings are available in the University of Nottingham Research Data Management Repository.

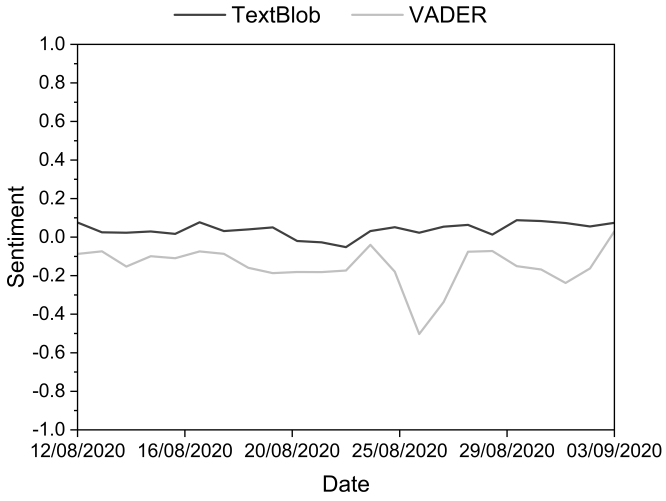


Figure 5: TextBlob and VADER sentiment analysis of tweets relating to Ofqual A Level algorithm in August and September 2020.

When presenting the results as a trajectory, the sentiment polarity exhibited considerable variability. Initially, on 12th August, the sentiment polarity suggested a marginally positive sentiment. However, in the following days, the sentiment polarity experienced fluctuations, indicating a shift towards a more neutral sentiment. The general trend saw an increase in negativity on 14th August, with a steady rise in positivity detected in tweets for the next few days. Throughout the remainder of August, the sentiment polarity continued to fluctuate, albeit generally maintaining a relatively neutral range. Notable deviations occurred, particularly on 24th and 25th August, suggesting a slightly positive sentiment.

Towards the end of August and into September, the sentiment polarity displayed more pronounced fluctuations. Notably, on 26th August, there was a significant drop in sentiment. However, on the 30th and 31st of August, the sentiment polarity rose, signifying a return to a more positive sentiment trend. The trend persisted into September, with sentiment scores remaining relatively positive, perhaps exemplifying a continued positive trajectory.

To interpret the results as best as possible, they were examined with key dates in the chronology of the Ofqual algorithm, as per the next

step in the process. For example, a negative change in sentiment was detected on 14th August, the day after the results were shared with students. A rise in positive sentiment detected on 17th August came on the same day as the government announced that the algorithm-calculated grades would be replaced with teacher-assessed grades. The sharp negative change in sentiment on 26th August came on the same day as UK Prime Minister Boris Johnson told students that their results had been affected by a 'mutant algorithm'. A rise in positive sentiment on 3rd September came the day after Ofqual Chair, Roger Taylor, apologised to students when appearing at the Educational Select Committee at the House of Commons.

There were instances where the algorithm classification involving negation and sarcasm could be seen as incorrect. Within the dataset, tweets that were categorised as positive (a score of 1.0) did include negation, pairing 'not' with 'believe' and 'no' with 'accident', for example. Once again, there were many tweets that could be deemed to be sarcastic that were classed as positive by TextBlob. With these tweets, hyperbolic adjectives such as 'sophisticated', 'great' and 'flawless' were all seen.

As a result, the VADER module was deployed. With the inclusion of the VADER sentiment data, as displayed in Figure 5, overall sentiment ranged from 0.03 to -0.5, indicating that sentiment is negative. At the start of the discourse, the VADER score registered a slightly negative sentiment. This negativity persisted over the subsequent days, with fluctuations but generally remains below zero. On 23rd August, there was an abrupt change in sentiment detected in tweets, as the data in Figure 5 shows – sentiment detected increased in positivity. Remarkably, on 26th August, a significant plummet in the VADER score was observed, descending to -0.50273. This sharp decline suggests a notable surge in negative sentiment surrounding the algorithm, possibly reflecting the comments by Boris Johnson about the 'mutant algorithm'.

However, towards the latter part of the sampling period, sentiment started to show signs of a shift towards positive, which may indicate a change in the narrative or a resolution to some of the concerns raised earlier.

When revisiting the tweets in the sample that included negation but were potentially incorrectly classified, new VADER scores were 0.8372, 0.6546, -0.1655 and 0.4019. For tweets deemed sarcastic, the VADER scores were 0.9739, 0.3612, 0.1548 and 0.9081.

#### 4.3.2.2 *Human Review and Critical Reflection*

Again, 10 tweets per day (230 total) were randomly sampled and classified by two reviewers according to whether they were positive, negative or neutral. The human review score matched the computer-assigned sentiment category on 40% of occasions. The inter-annotator agreement was 0.547, again indicating moderate agreement. This agreement may have been slightly higher than the topic modelling agreement levels due to the more straightforward classifying process (positive, neutral, negative only). Nevertheless, there were still disagreements between reviewers. Between them, classifying tweets that the algorithm deemed 'neutral' caused the most disagreement, with the reviewers not matching on 29/72 occasions.

26% of the tweets were found to contain negation structures. According to the human reviewers, 74.6% of these tweets containing negation structures were classified incorrectly by TextBlob. Considering sarcasm, 6% of the tweets reviewed were labelled as sarcastic. 71.4% of these tweets were classified incorrectly as positive by TextBlob according to the human review.

The critical reflection was as follows:

SUNSHINE One advantage of sentiment analysis was its efficiency in analysing large amounts of data quickly and easily with simple coding. Integrating VADER into the sentiment analysis process simplified the comparison between the two approaches. Tracking senti-

ment scores over time offered an instant overview of sentiment trends, facilitating the identification of turning points for further qualitative investigation.

RAIN One aspect of the investigation that proved challenging was the unresolved issues surrounding negation and sarcasm in sentiment analysis. These challenges may lead to decreased accuracy of techniques and the robustness of findings. Additionally, interpreting individual sentiment scores was difficult as they felt somewhat meaningless in isolation.

LIGHTNING A notable finding was the significant variation in outcomes, raising questions about the reliability of each tool for large-scale analysis. Particularly surprising was the considerable fluctuations in sentiment, especially during periods like August 24th, where contextual understanding failed to offer much insight into the changes.

FOG Despite serving as a starting point, the sentiment scores often lacked clear guidance on their significance and how they could deepen understanding of the discourse. Presenting scores without context further exacerbated the lack of clarity. For instance, discerning the difference between scores like -0.18091 and -0.1815, observed on August 21st and 22nd in the VADER classification, remained unclear without proper guidance.

### 4.3.3 *Emotions*

#### 4.3.3.1 *Expectations and Findings*

Our expectation was to see which emotions were the most common in the discourse and how the emotions detected may change over time. With this being only a three-week time period, it was expected that

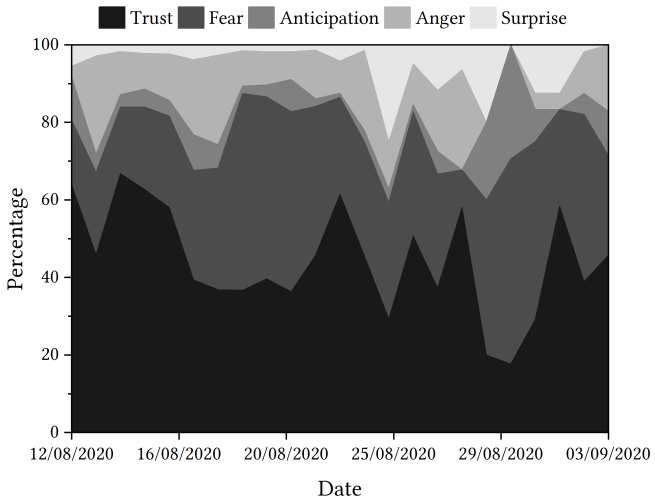


Figure 6: Emotions detected in tweets relating to the Ofqual A Level algorithm.

perhaps only a few different emotions would be detected or possibly only some fluctuation in the proportion of emotions would occur.

As shown in Figure 6, 'trust' was the emotion most often detected, followed by 'fear'. Looking at the trajectory of emotions detected in the sample period, there are key dates that appear to have greater spikes in 'fear', and thus a downturn in 'trust' detected, including the days following the u-turn announcement, the statement about 'mutant algorithm' and 30th August. However, it remains unclear whether tweets relating to 'trust' were indeed positive indicators of trust or indicators of mistrust. All emotion findings are available in the University of Nottingham Research Data Management Repository.

One potentially interesting emotion to explore is 'anticipation'. While there were some tweets throughout the discourse that were categorised as expressing the emotion 'anticipation', this is a relatively low amount of tweets until 30th August, when 11.63% of tweets were classified in this category, rising sharply from 4.17% the day before and no tweets detected two days before. Once again, 30th August does not appear to be a significant date in the timeline of events relat-

ing to the Ofqual algorithm, so the potential rationale for this increase is not immediately clear.

'Anger' and 'surprise' are the two other emotions detected in the dataset. 'Anger' appears to be more prevalent in the discourse at the start of the sample period, particularly on the day that results were released and the day of the government u-turn. With 'surprise', however, this emotion is detected in relatively few tweets at the start of the discourse, which seems to change on 26th August – the date that Johnson made a statement on the 'mutant algorithm'. This 'surprise' might be attributed to the use of the word 'mutant', or the general unexpected nature of the statement. In what might seem an apparent consequence of this statement, 'surprise' is generally detected more frequently than previously in the majority of the remaining days in the sample period.

There are questions as to whether the emotion 'trust' indicates the direction of trust. For example, tweets containing 'criticised', 'unfair', 'unequal' and 'failure' were all attributed to 'trust', despite being potentially more indicative of distrust instead. Additionally, when presented with a tweet that contained vocabulary such as 'great' and 'similar' (used in a literal and not sarcastic context), there was limited opportunity to classify this as 'happy' or 'supportive'. For context, EmoLex had classified this tweet as 'anticipation'. Going forward, this might be a consideration for developing expectations when working with emotion detection algorithms.

#### 4.3.3.2 *Human Review and Critical Reflection*

For the Ofqual discourse, 10 tweets per day (230 in total) were randomly sampled to be reviewed. Once again, the categories to be assigned were 'trust', 'fear', 'anticipation', 'anger', 'surprise', 'sadness', 'disgust', 'joy' and 'no emotion'. Reviewers matched the EmoLex assigned category on 37% of occasions. The inter-rater reliability was 0.481, again indicating moderate agreement. Once again, these agreement levels may be explained by the range of emotion options avail-

able. Tweets reviewed in this process will form the examples of the following section. Between the reviewers, classifying tweets that the algorithm deemed as 'fear' caused the most disagreement, with the reviewers not matching on 17/60 occasions. Instead, reviewers categorised these tweets as 'anger'.

The critical reflection was as follows:

SUNSHINE One of the aspects that worked well here was the speed of automated detection with a large dataset and that each tweet was able to be classified in some way.

RAIN As with sentiment analysis, the accuracy of the detection may have been an issue. Additionally, insights into the discourse are presented without context – as would have been helpful in the analytical process. With limited context, the emotions could be seen as arbitrary.

LIGHTNING A surprising element was the presence of 'positive' and 'negative' within the initial emotion set in EmoLex. This might have meant that vital information may have been missed as the tweets may have been closely aligned with other emotional categories, yet this has not been included in the results.

FOG The distinction between emotions could have been clearer. For example, should tweets associated with trust only feature those that actively support this emotion, rather than oppose it? As seen in further exploration, there were tweets categorised as containing the emotion 'trust' that may have been categorised as the opposite. If this had been clearer before using the tool, the picture of the discourse may have been more accurate.

### 4.3.4 *Section Summary*

This section of the chapter summarises the NLP-based results of the analysis of the A Level algorithm discourse. Through topic modelling using gensim LDA, as seen in subsection 4.3.1, four latent topics emerged, which ranged from governmental involvement and technical flaws in the algorithm to responses from education stakeholders and the algorithm's impact on educational institutions. The trajectories of these topics over time highlighted fluctuations in their prominence, with shifts in focus corresponding to significant events and developments.

The sentiment analysis, conducted using both TextBlob and VADER and detailed in subsection 4.3.2, provided further depth to the understanding of the discourse about the algorithm. While TextBlob indicated a predominantly neutral sentiment, VADER revealed fluctuations indicating shifts towards negativity, particularly during critical moments such as the government's u-turn on the algorithm and Boris Johnson's remarks on the algorithm being 'mutant'. The sentiment trajectories aligned with key events in the algorithm's timeline, offering insights into how these events influenced public sentiment. However, challenges with accurately classifying tweets containing negation and sarcasm underscored the limitations of sentiment analysis tools and the need for nuanced interpretation.

Finally, subsection 4.3.3 explored the detected emotions using EmoLex. This revealed differing patterns, with 'trust' and 'fear' being prevalent. The trajectory of these emotions fluctuated, perhaps reflecting the fear and uncertainty at the beginning of the discourse and the anger in the decision afterwards. However, challenges in accurately categorising tweets highlighted the limitations of automated emotion detection tools and the importance of contextual understanding. For example, negation and sarcasm stood out as linguistic phenomena that tools struggle with. Therefore, it can be stated that the classifi-

cation of tweets as expressing a certain sentiment or emotion is not always accurate.

Moreover, the threads of trust and fear showed that further exploration is needed to ascertain the cause of these emotions and the degree to which they are felt; this is possibly where a more discursive sociolinguistic approach may be helpful, like DA. Finally, the use of the reflective toolkit by Maclean (2016) enabled concise evaluations of the performance and suitability of each of the methods and raised potential concerns with interpretation and lacking transparency regarding limitations such as dealing with negation and sarcasm.

## 4.4 CORPUS LINGUISTICS AND DISCOURSE ANALYSIS

This section first comprises the CL keyword analysis, which enabled the identification of potential social actors for investigation. Based on this first list, four potential social actors (the algorithm, Ofqual, the government and students) were investigated through the examination of collocational strength and a focus on agency and blame, using DA.

### 4.4.1 *Keyword Analysis of Potential Social Actors*

Table 3 shows the top ten words with the highest keyness score when compared to EnTenTen2020. From this analysis, four potential entities were identified: the algorithm itself, Ofqual, students and the government, as they all appeared as top keywords. These were identified as they were all nouns that had the potential to be presented actively in a grammatical construction and thus could be a social actor. These entities are now explored in the following following sections, which detail how blame is placed or not placed on the entity of concern through the main events of the discourse. A sample of the concordances, examined in conjunction with collocational findings,

Table 3: The top ten words with the highest keyness score.

| Item | Relative frequency (per million) | | Score |
|---|---|---|---|
| | Focus corpus | Reference corpus | |
| algorithm | 28,339.45 | 0.51 | 29.3 |
| a-level | 9,881.38 | 1.27 | 10.9 |
| ofqual | 8,598.08 | 0.08 | 9.6 |
| results | 6,261.83 | 14.79 | 7.2 |
| grades | 5,717.25 | 7.94 | 6.7 |
| students | 4,730.1 | 94.14 | 5.2 |
| a-levels | 4,175.65 | 1.61 | 5.2 |
| by | 7,584.61 | 471.41 | 4.9 |
| exam | 2,826.54 | 30.25 | 3.7 |
| government | 2,518.88 | 45.21 | 3.4 |

is available in the University of Nottingham Research Data Management Repository.

### 4.4.2 *The Algorithm*

The collocational strength of the top ten lexical items associated with *algorithm* is shown in Table 4 (after stopword associations were removed, identified through NLTK's list by Bird, Klein, and Loper, 2009). The trajectory of the collocations over time can be seen in Figures 7 and 8. Both *a level* and *ofqual* appeared as adjectival modifiers to *algorithm*. *Flaws* collocates strongly with *algorithm* at the start of the discourse, pertaining to one particular tweet that had been retweeted many times about a father (hence the strong collocation with this word too) that points out 'algorithm flaws'. This returned towards the end of the discourse, where there were many tweets discussing how Education Secretary Gavin Williamson 'knew of the flaws of the algorithm'. Words with high collocational strength that are in the semantic field of education, such as *results*, *grades* and *exam* were also present, but could not say much about how the algorithm was pre-

Table 4: Collocational strength of *algorithm*.

| Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|
| a-level | 3405 | 6006 | 12.2297 |
| ofqual | 2393 | 5226 | 11.7701 |
| results | 1324 | 3806 | 11.0105 |
| flaws | 1090 | 1149 | 10.9247 |
| a-levels | 1110 | 2538 | 10.8458 |
| foresaw | 997 | 1015 | 10.8066 |
| father | 991 | 1025 | 10.7971 |
| exam | 1004 | 1718 | 10.7622 |
| grades | 1053 | 3475 | 10.703 |
| level | 903 | 1305 | 10.641 |

sented. Therefore, from this analysis alone, it is not clear whether the algorithm itself was given grammatical agency or perceived social agency by tweet authors.

However, through the manual examination of other concordances, the algorithm itself is presented as having agency and potentially being blamed for the events that occurred. In this section, the key findings relate to the active presentation of the algorithm, its metaphorical agency and personalisation, and how this changes through the timeline as tweets show an undetermined responsibility for the actions.

On August 12th 2020, tweets show the algorithm performing a task as the social actor in grammatical constructions. Tweets that contain structures such as 'that algorithm is going to screw you' and 'this algorithm appears to be cementing that bias towards the wealthy' received 235 total engagements (combined likes and retweets). The active syntactical structures imply that social agency is with the algorithm. On 13th August, the day results were released to students, there were also many tweets that gave the algorithm social agency, presented in a similar way, illustrated by the active statements that the algorithm 'caused today's chaos' (5795 engagements). Here, personalisation is seen. This is in addition to a tweet that contained 'the

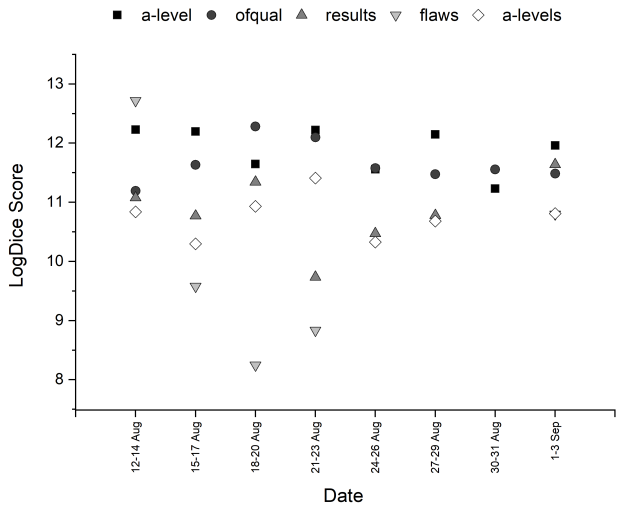


Figure 7: Temporal trajectory of LogDice scores of collocates of *algorithm* - part A.

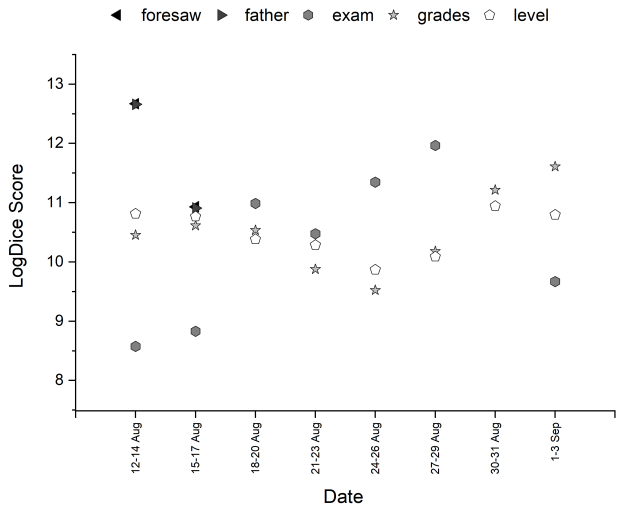


Figure 8: Temporal trajectory of LogDice scores of collocates of *algorithm* - part B.

algorithm used by Ofqual can't be applied to small cohorts' (5517 engagements), here foregrounding the importance of the algorithm, despite a lack of agency, through this passive construction. This could be seen as the *backgrounding* of Ofqual and a foregrounding of the algorithm.

Prior to the government change, transitivity analysis showed more cases of the algorithm being presented in an unfavourable way. Regarding pathways to university, one tweet says that it is 'intolerable that an algorithm is denying this to others' (7774 engagements), a clear active grammatical construction that places agency with the algorithm. Another tweet states that 'this racist, discriminatory and downright evil algorithm is ruining lives' (2595 engagements) – overtly stating that the algorithm has the power to have a substantial impact on humans, thus being *personalised*. Additionally, a tweet on 16th August stated that '97% of GCSE results fully decided by an algorithm' (1490 engagements). This implies that the algorithm has the capacity to make decisions on the outcome of the General Certificate of Secondary Education (GCSE) qualifications of students. Another well-engaged tweet on 16th August stated that the 'algorithm has given them Us and fails' (13256 engagements) – placing agency with the algorithm through *personalisation*.

This sentiment continued into the date of the reversed decision, 17th August 2020. One tweet with 2136 engagements included the clause 'your future should be based on your abilities not an algorithm', continuing the notion that the algorithm has the potential to change lives. Another tweet with 7126 engagements said that 'private schools had done better with the Ofqual algorithm'. Despite being part of a prepositional phrase in this context, the algorithm is still mentioned when the foregrounded part of the tweet is concerned with inequality of results. However, the algorithm is nominally labelled as 'the Ofqual algorithm' - thus, despite the active presentation of the algorithm, it is owned by Ofqual, thus potentially blurring the boundaries of blame and accountability.

There are occasions when the algorithm is referred to as being 'used' by an unknown actor. This is first seen on the most engaged-with tweet on 12th August, the day before results were released to students, which stated 'the algorithm used to grade A-level results is incredibly sophisticated' (4513 engagements). The fact that a transitive verb 'used' is chosen here without a named active social actor creates the impression that authors believe the algorithm is not to blame for the results, but the anonymous 'user' is. There are further instances where this occurs, such as the 'algorithm used for A-level grades' on 17th August (1695 engagements).

The algorithm is also presented passively, implying removed agency. One tweet with 1329 engagements states that people 'benefited from [the] algorithm' on 13th August. Additionally, the most engaged-with tweet on 15th August (10311 engagements) discussed the importance of rectifying the situation prior to the release of GCSE results the following week, stating that the qualifications would also be 'assigned *solely* by another Ofqual algorithm'. While this presents Ofqual as the possessor of the algorithm and could imply blame, the algorithm itself is performing the task of 'assigning' despite being an inactive entity. This is in addition to a tweet on the same day that explains '1/4 state school students were downgraded by the algorithm versus 1/10 private school students' (2931 engagements). Here, again, while a passive construction is used, the algorithm is not the focus of the construction; instead, the focus is shifted to the inequality of the 'decisions' that the algorithm made. Thus, while blame is not attributed to the algorithm through syntactical structures here, the subject matter of the tweet places blame on it through the foregrounding of this comparison. This *backgrounding* limits the agency that the algorithm has as a social actor but still implies blame.

Passive constructions continued on 18th August when a UK university tweeted about supporting students 'who have been disproportionately affected by the A-level algorithm' (298 engagements). Again,

while this is a passive construction, agency may still be attributed to the algorithm as it has performed an action that affected a human. However, it must be noted that the construction of the sentence foregrounds the students in this case.

Further on in the discourse, on 25th August, there are tweets that imply the algorithm is doing a 'job', an activity usually performed by a human. One author wrote 'Ofqual guidance doesn't require them to moderate – that was the job of the algorithm'. This personification and *personalisation* of the algorithm could place further blame and agency on it as a distinct social actor. This is in addition to a user who details that the algorithm had 'failed [their] daughter', thus implying that the algorithm had agency to perform such an action.

To summarise, the algorithm is mostly seen in active constructions that indicate the agency is with it as a social actor. The personalisation and agency metaphor strategies seen in tweets also add to the indication that people see the algorithm as a social actor too. There are, however, instances where the algorithm is portrayed in passive constructions, although blame could still be interpreted. In the final dates of the dataset explored, more tweets directed blame through agency at Ofqual and the UK government. There are some active constructions that involve the algorithm, but the majority are centred around organisations or individuals. These social actors will now be explored in more detail sections 4.4.3, 4.4.4 and 4.4.5.

### 4.4.3 *Ofqual*

This section explores Ofqual as a potential social actor, with a specific focus on active and passive agency, agency metaphor and individualism of a defined entity within Ofqual, Roger Taylor. The collocational strength of the top ten words associated with *Ofqual* is shown in Table 5. The trajectory of the collocations over time can be seen in Figures 9 and 10. Once again, the lexicon associated with educa-

Table 5: Collocational strength of *Ofqual*.

| Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|
| algorithm | 2396 | 17225 | 11.7719 |
| exam | 299 | 1718 | 10.4625 |
| results | 330 | 3806 | 10.2255 |
| exams | 227 | 1110 | 10.1972 |
| have | 299 | 3726 | 10.096 |
| ignored | 182 | 308 | 10.0737 |
| regulator | 182 | 347 | 10.0636 |
| used | 206 | 1339 | 10.0059 |
| unlawful | 169 | 387 | 9.94632 |
| not | 226 | 2961 | 9.82106 |

tion was present. Collocations of interest included *ignored*. This was seen throughout the discourse, such as 14th August ('Ofqual ignored offers of expert help with its algorithm') and 20th August ('Ofqual ignored exams warning a month ago'). The use of the word 'ignored' here could be seen as significant as it places Ofqual as the active social actor in the tweet. *Have* was also collocationally strong, often performing as an auxiliary verb where Ofqual is the social actor ('Ofqual have created an algorithm which just doesn't work', 'Ofqual have downgraded', 'Ofqual who have ruined young lives' and 'Ofqual have favoured the unadjusted small cohorts'). *Used* is seen in constructions that are active ('Ofqual has used an unequal algorithm') and passive ('the algorithm used by Ofqual) throughout the discourse. There was a great deal of engagement with a tweet that stated 'Ofqual exam results algorithm was unlawful, says labour'. Although not an examination of agency, the use of the adjective *unlawful* might be an indicator of blame.

Through further concordance examination, users showed other ways in which they blamed Ofqual. Immediately, it is clear that the process of *assimilation* is present in tweets pertaining to Ofqual due to it being a group. One of the most common situations that this

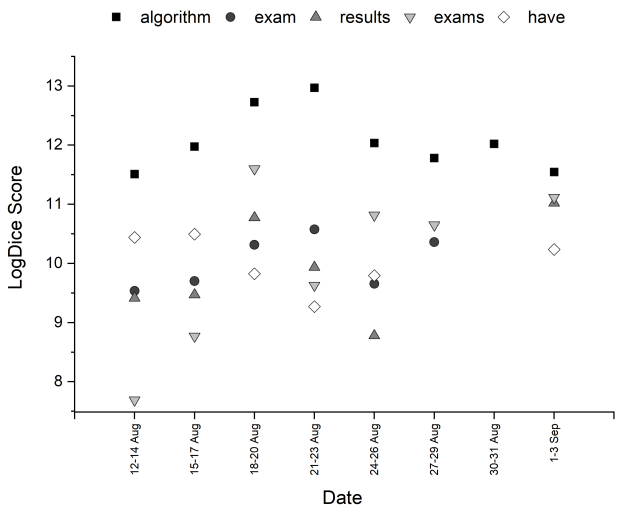


Figure 9: Temporal trajectory of LogDice scores of collocates of *Ofqual* - part A.

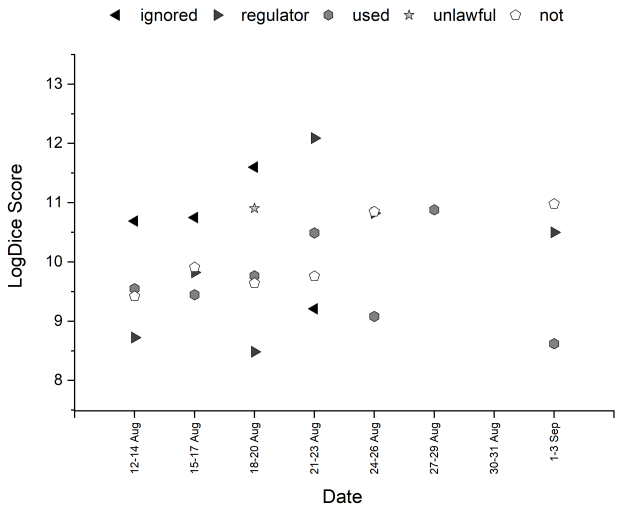


Figure 10: Temporal trajectory of LogDice scores of collocates of *Ofqual* - part B.

occurred was by attributing ownership of the algorithm to Ofqual, as seen in tweets that contained the phrases 'its algorithm', found throughout the discourse.

Upon the revision of grades, Ofqual was mentioned more in the discourse as a social actor. This is seen in tweets that involve the possession of the algorithm and some that talk about Ofqual as a separate social actor. In tweets that do discuss Ofqual as owners of the algorithm, such as 'experts question how their algorithm could so blatantly favour private schools', seen on 17th August with 4274 engagements, this possession is clear. However, the algorithm here still has some sort of agency as it is the social actor doing the 'favouring'. This blurs the lines between who the social actor is and, therefore, who is to blame. This implication of multiple entities that present the algorithm as the social actor but Ofqual as the possessor continues the following day. This is seen in a tweet with 270 engagements that states 'the government knew Ofqual's algorithm would disadvantage the disadvantaged'. This may result in blurred blame.

As previously alluded to, there are tweets that foreground Ofqual as the social actor, rather than as the owners of the algorithm. For example, one tweet with 2029 engagements on 20th August contains 'it's their faith in these one-dimensional metrics that bedevills education', with the possessive pronoun 'their' referring to Ofqual. This hyperbolic use of language to heighten emotion and impact intensifies the focus on Ofqual as a blameworthy social actor. This is exemplified further in a tweet with 135 engagements on 22nd August, stating 'Ofqual [...] applied the algorithm'.

In later parts of the corpus, this continues. One tweet with 106 engagements on 2nd September expresses exasperation with Ofqual by stating 'how did the Ofqual people not realise that what they did with the algorithm would not be acceptable'. Ofqual is clearly presented as an implicated social actor here, with the algorithm part of the prepositional subject phrase. This emphasises Ofqual's agency and, thus, implies blame to them. These tweets coincide with Ofqual

Chair, Roger Taylor, speaking directly to the Educational Select Committee.

Users also placed agency and blame on Taylor himself through *individualism*. This is seen especially in early September 2020, when Taylor spoke to the Educational Select Committee. As early as 13th August, the day results were released to students, Taylor is actively implicated. In the same tweet that stated that the 'algorithm caused today's chaos', the tweet author goes on to state that 'Ofqual chair Roger Taylor also chairs the Centre for Data Ethics & Innovation', which is heavily linked with Dominic Cummings, former advisor to Boris Johnson. This active construction, and use of the verb 'chairs', which is indicative of status and power, could implicate Taylor, especially with the high engagement with the tweet (5,795 engagements). There are other tweets from around a similar time that could place blame on Taylor through agency. For example, one tweet on 16th August states 'Roger Taylor, [...] responsible for the #algorithm, flunked his own A-levels but was given a "second chance" after passing the entrance exam' (117 engagements). Several verbal phrases in this tweet are attributed to Taylor – including that he is 'responsible' for the algorithm, and, potentially, the failure of the process. Additionally, blame is further implied through the idea that Taylor 'flunked' his exams and 'was given' (a passive construction) a second chance. Similarly to Ofqual, there are times throughout the discourse when the algorithm is attributed to his possession – such as 'benefit from grade inflation under his algorithm' (4164 engagements).

On 24th August, Taylor is presented in both an active and passive way. For example, a tweet with 518 engagements states 'Roger Taylor's company was criticised' for failures concerning algorithms in the past. This passive construction removes the social actor from the construction and foregrounds the importance of Taylor. This is further emphasised by the active role he is given later in the same tweet when the author writes that 'he's chair of the body charged with overseeing algorithms', and in another tweet that states 'Roger Taylor

chairs both Centres for Data Ethics and Innovation (CDEI) & Ofqual'. As well as overtly critiquing Taylor's conflicts of interest by holding multiple senior roles, the use of the lexical item 'chair' (in both noun and verb word classes) reinforces the status, power and responsibility that Taylor has.

On 2nd of September, Taylor appeared at the Educational Select Committee to discuss the algorithm's impact. Tweets placed agency and blame on Taylor. An example includes 'Roger Taylor [...] admits the decision to use an algorithm to award results was a "fundamental mistake"' (105 engagements). Taylor is clearly the focal social actor in the construction, with intensity heightened through the use of 'admits'. However, there are other tweets on this date that do implicate Taylor as a blameworthy social actor, but do so by using the word 'tells' in place of 'admits', thus softening the potential blame on Taylor.

To summarise, Ofqual is seen to be presented as a key social actor in this discourse, attracting blame from Twitter users by using active agency and possession. Taylor, here, is seen to be blameworthy through repeated individualism.

### 4.4.4 *The UK Government*

In this section, the UK government is explored as a potential social actor, focusing on assimilation and individualism for senior government figures. The collocational strength of the top ten words associated with *government* is shown in Table 6. The trajectory of the collocations over time can be seen in Figures 11 and 12. There are words that might be expected to be related to the government (*uk*, *tory*) and also words that are particularly associated with this specific discourse (*ofqual*, *algorithm*, *a-level*). *U-turn*, the word with the highest collocational strength, appears as both a noun ('should the government perform a u-turn'), a verb ('Ofqual want the government

Table 6: Collocational strength of *government*.

| Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|
| u-turn | 147 | 582 | 11.1546 |
| after | 80 | 764 | 10.1577 |
| must | 56 | 401 | 9.89148 |
| uk | 58 | 548 | 9.83631 |
| ofqual | 176 | 5226 | 9.73726 |
| tory | 45 | 281 | 9.66849 |
| algorithm | 396 | 17225 | 9.43429 |
| a-level | 139 | 6006 | 9.23917 |
| not | 80 | 2961 | 9.18879 |
| have | 93 | 3726 | 9.17913 |

to u-turn') and, later in the discourse, a noun phrase ('even with the government algorithm u-turn'). The majority attributed the action of the 'u-turn' to the government, as seen in excerpts such as 'the government has u-turned', 'government u-turn on exam results' and 'we welcome the government's u-turn'. *After* is frequently used as a prior conjunction to clauses such as these, discussing the need for teacher-assessed grades. Unlike the first two entities, this collocation analysis implies the government could be blameworthy.

*Must* is used as a modal verb in a variety of constructions that call on the government to address the situation, such as 'the government must u-turn', 'the government must apply cags' and 'the government must learn from the shambolic handling of a-level results'. All of these constructions place the government as blameworthy social actors.

Further concordance examination places blame on the UK government as a collective entity, as well as some individual figures. Once again, *assimilation* is found in many constructions. Tweets throughout the discourse refer to the algorithm as 'the government's algorithm', which is expanded upon as a noun phrase by different tweet authors,

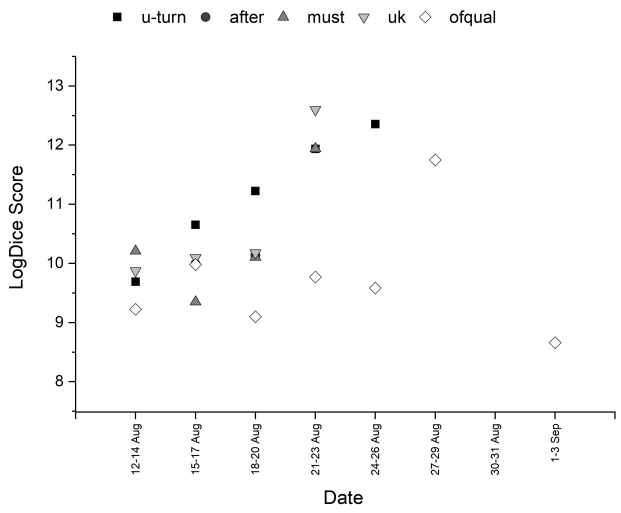


Figure 11: Temporal trajectory of LogDice scores of collocates of *government* - part A.

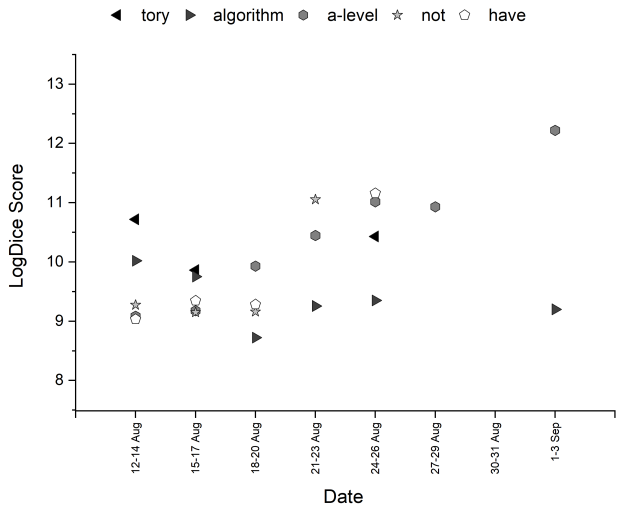


Figure 12: Temporal trajectory of LogDice scores of collocates of *government* - part B.

such as referring to it as the 'hastily-built government algorithm' (665 engagements).

In a direct address to A Level students on 13th August, one author said 'I am sorry this government has failed you' (1599 engagements). Blame is placed on the government as the implicated social actor. Further implications of blame could come from the active statements 'government refusing to learn from a level fiasco' (619 engagements) and 'this government really don't like teachers' (1490 engagements). Another tweet stated that the choice of using the algorithm was 'devastating by the UK government' (512 engagements). Although passive, this construction might attribute blame to the government through the foregrounding of the particularly emotive word 'devastating'. This is again seen in 'negatively hurt by the Tory algorithm' (434 engagements), where the emphasis is on the emotion (the 'hurting') rather than the government. Although this is *backgrounding*, the implication of blame remains.

There are some instances of support, rather than blame, early on in the discourse, too. A tweet with 362 engagements contains 'the government never trusts teachers but in this v unusual situation it is the fairest way'. The author implies the government is a social actor but in a positive way, despite the verbal phrase 'never trusts' usually being associated with negativity.

Upon the revoking of the use of the algorithm, tweets imply blame is with the government, including one example with 429 engagements that states 'time for the government to hold up their hands'. The implied imperative, the government as the subject of the clause and the colloquialism 'hold up [...] hands' may imply blame. A tweet with 1684 engagements from 18th August says 'the government will blame Ofqual', with the active construction perhaps showing that the government is attempting to distract blame from themselves. This is coupled with tweets that expand the possessive noun phrase, such as 'their rigged algorithm' (4105 engagements). Later, on 25th August, active constructions further implicate the government, such as 'the

government ignored red flags' (35 engagements). This links to the idea that ministers put their 'faith' in the algorithm.

On 26th August 2020, the day that UK Prime Minister Johnson announced that results had been jeopardised by a 'mutant algorithm', Twitter users placed blame on the government. The most engaged-with tweet on this day, which had 5884 likes and 1762 retweets, used a series of rhetorical questions to imply that the government was to blame for the results scandal. Part of the tweet reads, 'Who set the parameters for Ofqual's algorithm? ministers! Who didn't ask the right questions? Ministers! Who didn't ask for a simulation of the impact? Ministers!! So who should resign?' This tweet's use of effective tripling as a rhetorical device is noteworthy, but it also has aspects of agency to explore. The interrogative pronoun 'who' could be substituted for the government (or 'ministers' in this case), making them an implied active social actor in the fault of the algorithm. Although the responses to the tweet were not part of the original dataset, there were other tweets within the dataset that linked the same BBC article, thus acting as a springboard for conversation and framed contextually around this specific piece of information. These tweets presented the government as implicated social actors.

Once again, there are individual social actors within this body, explored as *individualism*. Firstly, there are specific instances where blame is attributed to UK Prime Minister Boris Johnson. Upon the release of results, structures in tweets indicated that he had ownership of the algorithm, such as 'clever Boris' algorithm' (96471 engagements), implying blame is with Johnson. Additional tweets also indicate blame with Johnson, specifically on 26th August. One user tweeted about Johnson that 'he can't wriggle out of responsibility with bluster and distortion' (32 engagements). This presents Johnson as the active social actor and the verb phrase 'wriggle out' may indicate he is to blame.

Other tweets discuss Gavin Williamson, UK Education Secretary of State at the time of the A Level results in 2020. On the day of

the government u-turn, one tweet stated that Williamson had 'signed off on' the algorithm (700 engagements), showcasing him as a blameworthy social actor and decision-maker. On 18th August, after the reversal, constructions included 'Williamson is trying to blame Ofqual' and 'he admits he didn't even bother checking it' (224 engagements). These constructions show his active agency. However, Williamson is also presented in passive constructions, with one tweet with 524 engagements saying that he 'was badly advised'. This reduces blame towards Williamson, especially through the obscuring of an unknown social actor in the construction through *exclusion*.

In summary, the findings here indicate that elements of blame through active agency and social action for the government can be derived from the tweets. Passive constructions use emotive language that still implies blame is with the government). There are times when assimilation occurs and, as the discourse continues, individualism is more apparent for Johnson and Williamson.

### 4.4.5 *Students*

The collocational strength of the top ten words associated with *students* is shown in Table 7. Again, there are anticipated semantically-related words present (*a-level*, *grades*, *gcse*, *england*). Many of the occurrences of *their* relate to how well teachers know their students (seemingly in retaliation to the decision to use an algorithm to calculate grades, rather than teachers, and discussions about their futures in the wake of the decisions made.

The strength of the relationship of *students* and *downgraded* can also be examined. These are a mix of passive ('students getting downgraded results by some algorithm') and active ('algorithm that downgraded many disadvantaged students') constructions, where students were the object in either. There were instances where the verb 'downgraded' was intransitive and the social actor performing the action

Table 7: Collocational strength of *students*.

| Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|
| a-level | 651 | 6006 | 11.23 |
| their | 395 | 2757 | 11.1663 |
| grades | 373 | 3475 | 10.9105 |
| gcse | 238 | 1344 | 10.8521 |
| have | 336 | 3726 | 10.7039 |
| downgraded | 155 | 914 | 10.3885 |
| england | 132 | 767 | 10.2139 |
| many | 115 | 613 | 10.0773 |
| all | 139 | 1425 | 10.0488 |
| given | 108 | 473 | 10.0458 |

was not included in the tweet ('40% of a-level students being downgraded'). While this reduces potential blame for students, it does not implicate another social actor. It is also important to note that this is another example of *assimilation*. Upon further DA examination, it appeared that students were presented as passive in the majority of constructions, regardless of the verb used, including *given* when the decision was reversed ('students in England will be given grades estimated by their teachers' - a tweet with many retweets). This may suggest that students are not as heavily implicated.

### 4.4.6 *Section Summary*

The analysis, using CL and DA, explored the evolving portrayal of key social actors – namely the algorithm, Ofqual, the UK government and students – throughout the timeline of events surrounding the A Level results controversy. Initially, the algorithm was depicted as an active agent, with blame attributed to its perceived flaws in educational outcomes. However, as discourse progressed, the direct blame subsided, and responsibility became more diffuse, often with external actors being implicated. Ofqual, initially painted as an active decision-

maker, saw its role transition to that of an owner or possessor of the algorithm, leading to blurred lines of accountability. Conversely, the UK government consistently featured as an active social actor, with blame intensifying as discourse unfolded, particularly towards individuals like Boris Johnson and Gavin Williamson. Meanwhile, students were predominantly portrayed passively. This analysis illuminates the complex interplay of blame and agency among the various stakeholders involved in the controversy.

## 4.5 DISCUSSION

The following section discusses the implications of blame being attributed to the algorithm itself, Ofqual and the UK government through the combination of collocation, transitivity and social action analysis. Although these are three different aspects, in this chapter they are explored in an intertwined way. This is then related to previous research into the algorithm and the A Level results of 2020 to contribute to existing analysis concerning blame and responsibility for the issuing of results. After, considerations as to how the results work in a complementary way are presented, building on the previously identified research gap.

### 4.5.1 *Topics, Sentiment and Emotions*

Overall, four latent topics were identified. The most prominent topic featured the word 'government,' indicating its significance in the algorithm's usage and withdrawal decision. Topic 2, with 'flaw' as the most prominent word, sparked discussions about the algorithm's suitability. Topic 4 initially dominated discussions, focusing on students, schools, and teachers affected by the algorithm. Topics 1 and 2 had discussions around government officials, while Topic 3 gained popularity after the government's u-turn, marked by terms like 'fiasco'

and 'mutant'. Once again, the labelling of these topics may not have been accurate. TextBlob sentiment analysis indicated neutral sentiment (0.088 to -0.052), while VADER showed negativity (0.03 to -0.5). Analysing results with key dates in the Ofqual algorithm's chronology revealed corresponding sentiment changes. Classifying negation and sarcasm in tweets presented challenges, leading to the deployment of the VADER module. Again, similar sentiment scores were difficult to interpret. 'Trust' and 'fear' were the most detected emotions, with key dates showing spikes in 'fear' and a downturn in 'trust', such as the u-turn announcement and the mention of the 'mutant algorithm'. However, it remains unclear whether tweets expressing 'trust' truly indicate trust or mistrust.

One significantly challenging aspect of this analysis was the determination of the number of topics chosen for the discourse. Initially, a manual decision was made to expand from three topics to four, although employing a measure for topic coherence, such as Hierarchical Dirichlet Process, could have aided in arriving at an optimal number (Teh et al., 2004). This decision-making process influenced the subsequent errors encountered with classification. The human review revealed discrepancies between the algorithm-generated results and human classification. While following the principles of minimising topics to reduce inaccuracies, it may have been more beneficial to refine topics into smaller, more defined themes (Sengupta, 2019).

Additionally, sentiment analysis highlighted inaccuracies in classifying negated and sarcastic tweets. A considerable proportion being seemingly misclassified. 74.6% of negated tweets and 71.4% of sarcastic tweets were classified inaccurately when compared to human classification. These findings echo previous studies on sentiment analysis challenges, indicating potential limitations of off-the-shelf tools in capturing nuanced sentiment in social media data (González-Ibánez, Muresan, and Wacholder, 2011; Gupta and Joshi, 2021).

Furthermore, the examination of items of interest with context sheds light on the challenges of interpreting results accurately. The

limitations in interpreting results, especially in sentiment analysis, were noted due to the lack of guidance on the meaning of scores, suggesting a need for a more nuanced understanding of sentiment analysis outputs (Pokharel, 2020; Sivalakshmi et al., 2021). Similarly, the emotion detection output trajectories raised questions about the categorisation of emotions, advocating for more distinct emotional categories to improve accuracy (Balakrishnan and Kaur, 2019; Balakrishnan et al., 2019; Fast, Chen, and Bernstein, 2016; Jiang, Brubaker, and Fiesler, 2017).

In the critical reflection phase, Maclean's weather model facilitated the identification of both the advantages and limitations of the NLP tools used. While the tools were found to be easy to implement and provided a starting point for investigation, their accuracy was questioned, especially concerning diverging interpretations when examining linguistic features in context. This reflects the broader critique of computational linguistic methods and emphasises the need for a more nuanced understanding and application of these tools in social media research (Saura, Ribeiro-Soriano, and Saldaña, 2022).

### 4.5.2 *Blame for The A Level Results*

Through the analysis of transitivity in concordance lines, collocation and DA, underpinned by SAR, it was possible to see how blame was attributed to social actors throughout this Twitter discourse. The algorithm itself was most commonly presented as having active agency. The tweets that supported this seemed to imply that the algorithm was a social actor, despite its inanimate state, and so blame was shifted to the algorithm. Tweets implied that the algorithm was able to make decisions independently. This is in line with expectations of agency and blame that are outlined by Richardson, Mueller, and Pihlaja (2021) and personalisation by Van Leeuwen (2008).

Through personification and agency metaphor, the algorithm is depicted as carrying out human-like actions. This appears to support the idea of Goatly (2007) that this is done for increased dramatic effect and implies the algorithm has the capacity to make independent decisions, such as removing pathways to university.

Although less frequently, there are also times when the algorithm is included in passive constructions. This is especially true when the algorithm is being referred to as being used by an unknown social actor, thus shielding the 'user', and may take agency away from the algorithm and obscure blame. There are times when more intense verbs are used in passive constructions, still implicating the algorithm. This relates to the notions of agency specified by Clark (1998) and could seen to be obscuring agency through *backgrounding*, according to principles of SAR (Van Leeuwen, 2008).

However, considering verb choices, there are passive constructions that contain the verbs 'assigned' and 'graded'. Thus, a small portion of tweets using passive constructions appear to imply that the algorithm can still be blamed. This can be categorised as an agency metaphor according to Morris et al. (2007).

This builds upon existing research that found that students thought that the algorithm's result generation was unfair, thus implicating the algorithm (Bhopal and Myers, 2020) and ties into the potential backlash against algorithms that was reported to have occurred – and predicted to intensify – by Kolkman (2020) and Hecht (2020). This, in turn, supports one of the other findings: that students were not blamed through agency and transitivity in this Twitter discourse due to their passive presentation.

The UK government and the regulation body Ofqual were also presented as responsible social actors by Twitter users. For both social actors, active statements were seen that could implicate them as agents of blame. This was less frequent than the algorithm being implicated at the start of the sampled discourse and more frequent towards the

end of the discourse. *Assimilation* and *individualism* were both seen here.

Some tweets showed how blame was attributed to social actors through the possession of another. For example, Ofqual and the UK government were, in many tweets, seen to be the owners of the algorithm, which implies that they are to blame for the failures of the algorithm. This occurs throughout the discourse, especially on dates of significant events, such as the algorithm belonging to Roger Taylor on the date he appeared at the Educational Select Committee, the algorithm belonging to Boris Johnson on the date he called it a 'mutant algorithm', and the algorithm belonging to Gavin Williamson on the date of the u-turn. The idea of another entity possessing the implicated entity of the algorithm also blurs blame. The examination of how context affects language plays a crucial role in finding how blame is expressed through transitivity and, also, possession (Johnson, McLean, and Kobayashi, 2020).

### 4.5.3 *CL and DA to Complement NLP-Based Tools*

As per the objectives of this chapter, specifically objective 1d, it was important to examine how the qualitative findings from CL and DA, in addition to statistical collocation measures, provided further nuance to the quantitative findings from the NLP-based tools.

Overall, using the sentiment, topic and emotional trajectories provided a sound starting point for analysis. An example of this is the analysis conducted on 26th August, where the examination of VADER sentiment analysis pinpoints 26th August as the date with the largest sentiment change and the lowest sentiment value in the discourse. Through using CL and DA, it was clear that the majority of blame – through active agency, agency metaphors, hyperbole, possession, assimilation and individualism – on this date was directed towards the UK government and Boris Johnson. This was the date he

declared the algorithm to be 'mutant'. The combination of analyses may suggest that Johnson's actions implicated him as responsible for the failure of the algorithm's deployment due to the fact that the previous sentiment scores were low and tweet authors portrayed him as an implicated social actor.

CL was used primarily to identify potential social actors of blame and uncover patterns of transitivity (Jaworska, 2017). Combining these analytical perspectives enhances the findings beyond sentiment analysis.

There were, however, some issues with the data collection process. Upon reviewing tweets, it was clear that there were many replies to tweets that formed part of the discourse. But, due to the specific parameters of the search criteria used to collect this data, these replies were not part of the dataset. This potentially limits findings, especially as DA is underpinned by the analysis of interaction between others (Johnson, McLean, and Kobayashi, 2020). However, other tweets used the same news articles to provide context to their tweets. This is still a response to a main source and connects tweets to one another, therefore mitigating some of these shortcomings.

Above all, this demonstrates that the combination of NLP, CL and DA is a suitable mechanism to be deployed on Twitter discourses surrounding social and topical issues (Aljarallah, 2017; Kreis, 2017; Sveinson and Allison, 2021). It also demonstrates value for a combination of qualitative and quantitative measures being used to analyse social media (Wodak, 2007). This echoes the findings of previous studies that have done this successfully with different qualitative methods(Atteveldt, Velden, and Boukes, 2021; González-Ibánez, Muresan, and Wacholder, 2011) and showcases that this combination can be applied to Twitter discourses too. Ultimately, using CL and DA provided a more detailed lens to explore urgent social ideas and, in this case, blame and social actors (Van Dijk, 1997).

### 4.5.4 *Limitations and Future Work*

As there were over 18,000 tweets, it was not possible to examine all of these individually. Although the use of NLP and CL may have mitigated this, more insight may be waiting to be unearthed in this discourse. As previously expressed, the search criteria used to form the initial dataset may be missing aspects of the discourse due to its strict lexical conditions. Finally, using DA meant that the analysis was approached with an individual's own subjective perspectives, potentially questioning the validity of the insights (Gill, 2000; Morgan, 2010), although this was mitigated as much as possible through the using the three distinct approaches.

When considering future work, there is potential to use NLP, CL and DA to investigate related threads or themes. For example, this exploration could be enhanced by investigating thematisation, which would link to the latent topics found using computational linguistics (Halliday, 1994) and the use of structural-functional linguistics and social-semiotics, which may enrich the analysis by considering the socio-cultural context and meaning-making processes inherent in Twitter discourses (Mpofu, 2022; Osei Fordjour, 2021; Tucker et al., 2020). This allows for greater depth of research into the views expressed about the algorithm and could be done by multiple researchers to mitigate subjective biases. On a related note, a further suggestion may be to continue to use SAR to examine how the different social actors interact with one another.

Another suggestion would be to effectively integrate the approach of 'quantitative first, qualitative second' into a more iterative cycle. Considering principles of iterative data science, such as the 'epicycles of data analysis' (Peng and Matsui, 2016), a process could focus on the cyclical development of expectations, analysis of data, and matching of expectations to data, which repeats. This might mitigate not being able to analyse the replies excluded from the original dataset. In this model, the discourse becomes a 'moving feast', where NLP-

based tools can then be re-deployed to capture replies to key tweets, which are further analysed using CL and DA. Similarly, Social Network Analysis could be used with NLP and CL approaches to explore language patterns in this discourse, in a similar way to McGlashan (2020).

## 4.6 CHAPTER SUMMARY

In summary, the NLP-based results unveiled four latent topics, with a notable focus on government involvement and algorithmic flaws. These themes evolved over time, initially centred on discussions around government officials before shifting to a heightened scrutiny of algorithmic shortcomings post the government's u-turn. However, the classification of these topics may have lacked precision, signalling a necessity for refinement. For the sentiment findings TextBlob's trajectory suggested a neutral stance, whereas VADER indicated negativity, with sentiment fluctuations mirroring pivotal events in the algorithm's timeline, such as a significant dip when Boris Johnson labelled the algorithm 'mutant'. Challenges were apparent in categorising negated and sarcastic tweets, leading to a significant proportion of misclassifications.

Furthermore, the examination of emotions detected 'trust' and 'fear' as predominant, with fluctuations reflecting shifts in public perception. However, uncertainties persist regarding the interpretation of tweets expressing 'trust' in relation to the algorithm. Reflecting on the analysis process, challenges in determining the most suitable number of topics and accurately classifying tweets were noted. The limitations of sentiment analysis tools and emotion detection algorithms were evident, indicating the need for a more nuanced understanding and application of computational methods in social media research. Overall, while computational methods offer valuable insights, their limitations underscore the importance of critical reflection and refine-

ment in research methodologies, hence the suitability and contribution of CL and DA to investigate how the algorithm was portrayed in terms of agency.

The CL and DA findings reported and discussed in this chapter show that many Twitter users blamed the algorithm as a standalone social actor in the conetxt of the A Level results. This reaction was expressed through active agency, including agency metaphor (such as 'that algorithm is going to screw you) and personalisation of the algorithm (such as 'the job of the algorithm'). This suggests a tendency among Twitter users to personify the algorithm, treating it as a sentient entity with deliberate intentions and actions. By doing so, users assigned it agency and responsibility, framing it as a primary cause of the perceived negative outcomes.

Additionally, the UK government and Ofqual, and devolved social actors within these organisations like Taylor and Johnson, were also blamed by Twitter users through similar constructions and elements of possession (such as 'benefit from grade inflation under his algorithm'). This was seen less frequently at the start of the discourse and more frequently towards the end. This was mainly done through assimilation in earlier tweets and individualism in later tweets. This pattern indicates a change in Twitter users' views of responsibility. Initially, blame was largely placed on the algorithm, reflecting a general frustration or outrage. However, as the conversation progressed, users increasingly targeted figures from Ofqual and the UK government, holding them directly responsible for the perceived failures. The shift from a broad attribution of blame to specific individuals may suggest a move towards pinpointing individual culpability within these larger institutions.

Furthermore, passive constructions could be seen for all of these social actors, with some indicating more blame than others (such as 'the algorithm used by Ofqual'). Techniques to obscure and shift blame were also seen, like backgrounding (such as 'devastating by the UK government') and exclusion (such as 'he was badly advised').

The use of passive constructions may serve to deflect direct accountability, while techniques like backgrounding and exclusion further obscure the roles and actions of specific actors.

Ultimately, although it could not be determined which social actor out of the algorithm, Ofqual and the government were blamed the most, this chapter concludes by stating that these entities were presented as blameworthy social actors throughout the discourse. As well as providing insights into the online response to this particular event, there is potential for a broader impact too. Despite the disruption of the pandemic coming to an end in the UK, this contribution provides insights into how members of the public may react to future decision-making algorithm interventions.

In addition, the methodological conclusions illustrate how CL and DA can be used in a complementary way with NLP-based computational linguistic tools like sentiment analysis. More specifically, using quantitative data as starting points allows for more focused qualitative analysis. For example, the previously reported significant negative shifts in sentiment coincided with more authors suggesting blame was with the UK government and Boris Johnson. To ensure the application of 'epicycles of data science' creates an iterative computational and discursive methodological process, a more in-depth investigation of blame attribution and expression could be undertaken.

This chapter makes several significant contributions to the overarching thesis. Firstly, it provides a detailed analysis of Twitter discourse surrounding the A Level results controversy, shedding light on the attribution of blame to various social actors, including the algorithm itself, government officials and regulatory bodies like Ofqual. Twitter users blamed the algorithm, personifying it and attributing active agency. Additionally, blame shifted towards government and Ofqual, indicating a transition from broad attribution to specific culpability. Passive constructions and blame-shifting techniques were evident, suggesting nuanced expressions of responsibility. This depiction of blame is especially important for addressing the overarching

research question of the thesis and comparing the findings to those of the other case studies, which will occur in Chapter 7.

Methodologically, this chapter has demonstrated that CL and DA complement the NLP-based computational linguistic tools in researching the 2020 A Level algorithm; however, there is further scope for how these approaches can be used in an iterative manner.

# 5

# COVID-19 CONTACT-TRACING APP

## 5.1 STUDY BACKGROUND

As previously outlined, the agency of automated decision-making algorithms is a long-standing debate in academic research (Araujo et al., 2020; Pepper et al., 2022; Wagner, 2019). Decision-making algorithms can mitigate human errors or inaccuracies (Bullock, 2019; Busch and Henriksen, 2018; Panagiotopoulos, Klievink, and Cordella, 2019; Young, Bullock, and Lecy, 2019). However, when they operate decisions in lieu of individuals, the algorithms can be seen to develop a social agency and be perceived as having human-like characteristics (Crang and Graham, 2007; Ziewitz, 2016). This is an issue because it blurs the lines between the role of technology and human agency (Burrell, 2016; Wagner, 2019), potentially leading to situations where responsibility and accountability become ambiguous or misplaced (Mittelstadt et al., 2016; Selbst et al., 2019). When decision-making algorithms are not performing their assigned tasks as expected, though, investigating algorithmic agency can mitigate additional problems, such as reinforcing biases or producing outcomes that undermine trust in automated decision-making systems (Feier, Gogoll, and Uhl, 2021; Olhede and Wolfe, 2020; Peeters, 2020; Velkova and Kaun, 2021). Of course, investigating this agency and its implications for trust and blame is the core aim of this PhD thesis.

Like the A Level algorithm discussed in Chapter 4, another exemplary decision-making algorithm that has had significant societal impact is the NHS Covid-19 App, which altered individual working patterns and the ability to socialise (Kent, 2020; Kretzschmar et al., 2020). This app was designed and released to mitigate the spread

of Covid-19 in the UK by tracing contact with infected individuals and notifying people to self-isolate (Dowthwaite et al., 2021; Jacob and Lawarée, 2021; Kretzschmar et al., 2020). Launched in September 2020, six months after the Covid-19 virus began to circulate in the UK, this app gained public attention due to an array of issues and concerns (Dowthwaite et al., 2021; Mbwogge, 2021; Paucar et al., 2022), with social media websites, like Twitter, voicing the views of many users. Several studies investigated the sociological and epidemiological impact of the app (Kent, 2020; Marsh et al., 2021; Smith et al., 2022; Wymant et al., 2021), yet a gap still persists with regards to how the app has been discussed on Twitter, specifically.

When considering the perceived agency of decision-making algorithms, these expressions of agency about any given entity can be investigated in multiple ways, such as through interview or observation (Ahearn, 1999; Grillitsch, Rekers, and Sotarauta, 2021). However, on social media specifically, Corpus Linguistics (CL) and Discourse Analysis (DA) were purposefully deployed to examine the relationship between grammatical agency and social agency (Richardson, Mueller, and Pihlaja, 2021). Grammatical agency – or transitivity – can show whether an entity is presented actively performing an action or passively having an action performed on/to them (Leslie, 1993). Deconstructing the agency of decision-making algorithms in the discourse can shed light onto the perceived power relations between entities (Clark, 1998) and how these can ultimately indicate social actors in discourses (Van Leeuwen, 2008). To address the lack of examination of grammatical agency and transitivity in social media discourses, the social agency of decision-making algorithms was uncovered by examining tweets mentioning the NHS Covid-19 app.

By applying three methodological approaches, Natural Language Processing (NLP), CL and DA, underpinned by Social Actor Representation (SAR) (Van Leeuwen, 2008), the keyword 'app' was examined in context to examine whenever it featured as a common grammatical subject of perceived agency.

Thus, this chapter shows how the social agency of the app is implied or established through how users presented it, via grammatical constructions. For clarity, this chapter primarily comprises detailed specifics of the approach utilised for this study, as delineated in section 5.2. Following this, section 5.3 conducts an analysis of the findings derived from the automated topics, sentiment, and emotion trajectories. Section 5.4 undertakes an exploration of discourse through the perspectives of CL and DA. Finally, these findings are scrutinised in conjunction with pertinent literature on digital contact-tracing in the UK and beyond, as well as public-facing decision-making algorithms more broadly in section 5.5, offering overall findings on the online representation of the app and its perceived agency, responsibility, trustworthiness and blameworthiness.

### 5.1.1 *Study Research Question and Objectives*

Stemming from the main research question that this thesis addresses, the key sub-research question for this chapter is as follows:

> What insights into agency, trust and blame in the Twitter discourse surrounding the NHS Covid-19 app can be achieved through combining language analysis approaches?

In turn, the following objectives will be addressed:

2a Demonstrate how Natural Language Processing techniques (sentiment analysis, topic modelling and emotion detection) provide insight into Twitter discourses surrounding the NHS Covid-19 app.

2b Demonstrate how Corpus Linguistics, particularly collocation, provides insight into public Twitter surrounding the agency of the NHS Covid-19 app.

2c Demonstrate how Discourse Analysis provides insight into Twitter discourses surrounding the agency, trust and blame of the NHS Covid-19 app.

2d Identify the strengths and limitations of using the three approaches to investigate Twitter discourses surrounding the NHS Covid-19 app.

## 5.2 STUDY APPROACH

### 5.2.1 *Data*

Data extraction occurred using the Tweepy module in the Python programming language Roesslein, 2009. The key search criteria for this were tweets containing '@NHSCovid19App', which is the official Twitter handle for the UK's contact-tracing app, and the related hashtag '#NHSCovid19App'. The reason for this choice was to ensure that tweets were directly related to the experience of the contact-tracing app itself, rather than the wider NHS Test and Trace system or the Covid-19 pandemic generally. Although aspects of the discourse may not be revealed through this search term alone, it provided a starting point for investigating views expressed about the app.

In total, 180,281 tweets (1,797,052 words) were collected from 23rd September 2020, the day before the app launched in the UK, to 31st July 2021. Further to this, a second dataset was collected using the search term 'pingdemic' to capture relevant tweets relating to the surge in self-isolation notifications in July 2021. This dataset contained 36,022 tweets (831,579 words). After this, tweets were condensed down to remove advertisements from the dataset, resulting in a final corpus of 118,316 tweets over an eleven-month period. The data was sourced from the United Kingdom, as this is where the app was deployed, and only tweets in English were selected. Therefore, the analysis investigated views expressed in English only. The tweet

IDs and other associated information can be found in the University of Nottingham Research Data Management Repository.

Following the best practices recommended in the social media research literature, explored in section 3.1, any screenshots of tweets that may later identify their author were not included. Instead, as part of the data-cleaning process, tweets were anonymised, and only short extracts from tweets were reported verbatim (therefore, including typographical or grammatical inaccuracies). This project design was approved by the university department's ethics committee (approval number CS-2020-R33). Data was pseudonymised during extraction, with a unique number automatically generated to refer to each tweet. Stopwords were removed from the dataset using gensim, along with the removal of all long and short URLs and the indication 'RT' (retweet) at the beginning of any tweet. Twitter handles that appeared within tweets were also redacted, using gensim, for anonymity.

### 5.2.2 *NLP-Based Techniques*

#### 5.2.2.1 *Topic Modelling*

The chosen technique for topic modelling was LDA, implemented using gensim. Initially, the existing data underwent tokenisation via gensim's 'simple_preprocess' function. Following this, bigram and trigram models were created using the 'phrases' function. This process also facilitated the creation of bigrams and their lemmatisation using the Natural Language Toolkit (Bird, Klein, and Loper, 2009). Integration of the id2word dictionary with the gensim corpora produced a dataset-specific dictionary, assigning unique IDs to each word in the document. With this dictionary, a corpus was generated, mapping word IDs to their respective frequencies (Wang et al., 2020). Finally, topics were derived and displayed using the 'gensim.models.ldamodel.LdaModel' function within the gensim framework.

#### 5.2.2.2 *Sentiment Analysis*

Using TextBlob and VADER allowed the comparison of outputs, particularly focusing on VADER's claimed ability to account for negation within its algorithm. While the TextBlob model was trained using a Covid-19 dataset (Lamsal, 2021) with the 'train' and 'test' commands to familiarise it with Covid-related vocabulary, no such training was required for VADER as it was pre-trained. For TextBlob, a CSV file containing tweets was imported into the Python library and executed the 'blob = TextBlob(sentence)' command. Meanwhile, for VADER, the 'sentiment_analyzer_score' function was used. This set parameters to label each tweet as 'positive', 'negative' or 'neutral', where a score of 0.05 and above signified 'positive' sentiment, and -0.05 and below denoted 'negative' sentiment. These thresholds have been consistent throughout the case studies to provide a clear distinction between positive and negative sentiment, whilst also allowing for a margin of neutrality to accommodate tweets with more ambiguous sentiment expressions.

#### 5.2.2.3 *Emotion Detection*

Through EmoLex, the 'top.emotions' command was used. A CSV table indicating the correlation of each tweet with emotions such as fear, anger, anticipation, trust, surprise, sadness, disgust and joy was generated. Furthermore, a separate column was included to identify the dominant emotion in each tweet. enabling a more obvious identification of the emotion associated with a tweet.

### 5.2.3 *Corpus Linguistics*

With the aid of CL-computerised tools, the study was focused on collocation, i.e. the co-occurrence of two or more words within a predefined word span (Jaworska, 2017). The CL software used to undertake this analysis was The Sketch Engine (Kilgarriff et al., 2008),

which was chosen for practical and analytic reasons. Indeed, as it is freely available to many academics, it allows to upload of ad-hoc corpora and it provides a series of reference corpora which can be used for comparisons.

The analysis performed in this study contains different stages. First, keyword analysis was used to identify keywords in the discourse, using the embedded English Web 2020 (enTenTen20) (Suchomel, 2020) as the term of comparison. EnTenTen20 has over 36 billion words of specifically internet texts, including social media, and so acts as a suitable reference corpus. Additionally, keyness scores were generated by comparing the frequency of the words in the target corpus to the frequency of the words in the reference corpus. This has allowed examining the key characteristics of the corpus compiled, providing an overview of the tweets collected for the analysis.

Secondly, concordance lines featuring 'app', were examined through collocation analysis to investigate its role as a potential social actor. In order to verify active constructions, the collocation criteria were 'app' and one verb to the right (R1). To ascertain passive constructions, the search criteria was 'by the app'. Passive constructions, such as 'the app was x-ed by...' and 'the app has been x-ed' were examined by including verbs to the right (R1). To be specific, these were removed from the active constructions and added to the passive constructions.

Moreover, the LogDice was considered as the statistical measure of collocational strength. LogDice was included as it not only measures the statistical significance of a collocation, but it also factors in the size of the subcorpus, making comparisons between subcorpora of different sizes easier, as explained in 3.4.2. To take advantage of this capacity, this study split the corpus into five subcorpora that reflected the key moments in the evolution of the pandemic in the UK, in chronological order:

- Period 1: App Launch (September 2020)

- Period 2: Early Months (October to December 2020)
- Period 3: Second National Lockdown (January to February 2021)
- Period 4: Later Months (March to June 2021)
- Period 5: 'Pingdemic' (July 2021)

The analysis showcases the strongest collocates for each time period, ranked by LogDice score. A minimum threshold of three occurrences was used for the collocate to be significant enough to report, hence the variation of collocates in each time (previously explained in subsection 3.4.2).

### 5.2.4 *Discourse Analysis*

Finally, from this, Discourse Analysis (DA) was applied to examine agency and social action as expressed in the (concordance) lines, where 'app' appeared as a (key)word in context. In this study, DA was used complementarily to The Sketch Engine CL-analysis tool (Kilgarriff et al., 2008) to pinpoint different perspectives and meaning shades, referring to the various subtle nuances that a word or phrase can convey (Cruse, 1986). Therefore, these approaches were deemed especially effective together, as accomplished in studies (Abbas and Zahra, 2021; Baker, 2012; Nartey and Mwinlaaru, 2019) with similar purposes to this analysis, as previously illustrated in subsection 3.5.2. Additionally, several studies demonstrate the benefits of using DA on Twitter discourses, specifically (Aljarallah, 2017; Kreis, 2017; Sveinson and Allison, 2021). Therefore, this methodological approach was deemed to fit with the data-driven approach to analysis in an attempt to answer the research questions around the presentation of the app on Twitter and its implications regarding responsibility, trust and blame.

Once again, Social Actor Representation (SAR) was used as the theoretical underpinning. Specifically, this chapter focuses on socio-semantic categories such as *excluding* and *backgrounding* of grammatical agents, *personalisation*, *impersonalisation* and agency metaphor, all of which offer insights into users' attribution of responsibility. Additionally, representation structures like *genericised* and *specified* social actors reflect power dynamics within discourse, vital for comprehending responsibility attributed to it in tweets.

By analysing all these characteristics in the Twitter discourse collected, it was intended to identify common presentation traits of the NHS Covid-19 app, ultimately displaying how power relations are communicated in real-life data dealing with algorithmic-operated decisions, even when mechanisms are not fully clear. After establishing these, similar semantically-related thematic groups (as seen previously in Razis, Anagnostopoulos, and Saloun, 2016 and Kitishat, Al Kayed, and Al-Ajalein, 2020) were identified to aid the analysis of the presentation and perceptions of the app over time.

## 5.3 NLP-BASED TECHNIQUES ANALYSIS

This section presents the results from the NLP-based analyses. Notably, subsection 5.3.1 presents the findings from the LDA topic modelling, subsection 5.3.2 details sentiment trends and subsection 5.3.3 outlines the emotions present in the dataset.

### 5.3.1 *Topics*

#### 5.3.1.1 *Expectations and Initial Findings*

One of the expectations of using these methods for this case study was to see the broad themes associated with the NHS Covid-19 app that were being discussed online. It was anticipated that each topic would be generated with a distinct set of words that would be asso-

Table 8: Ranking of the top 10 lexical items associated with each latent topic

| | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| 1 | isolate | serco | download |
| 2 | positive | government | protect |
| 3 | contact | phone | help |
| 4 | work | work | store |
| 5 | notification | data | risk |
| 6 | code | download | google |
| 7 | phone | iphone | apple |
| 8 | results | private | play |
| 9 | self | phones | love |
| 10 | tested | good | available |

ciated with it to make it clearly defined. It was also expected that the lexical items found within each theme would make it easy to label the topics. These expectations were rooted in the understanding of the capabilities of the approach utilised and the nature of the data.

Three latent topics were discovered through gensim LDA. Each topic contained ten key lexical items. These words are presented in descending order of association with the latent topic in Table 8. The number of topics was decided on through manual topic inspection and regeneration, examining the ten key words each time, to ensure minimal lexical item overlap. All topic findings are available in the University of Nottingham Research Data Management Repository.

With regard to how the topics presented themselves in the tweets from each month of the research time frame, Figure 13 details the percentage of tweets relating to each topic per month.

The discovery of these topics through the use of gensim's LDA function provided starting points for further focus. As shown, the most featured word of the most prominent topic was 'isolate'. This foregrounded the importance of the topic of self-isolation as part of the discourse surrounding this specific contact-tracing app, underlining that the disruptive effect the app could have on people's lives

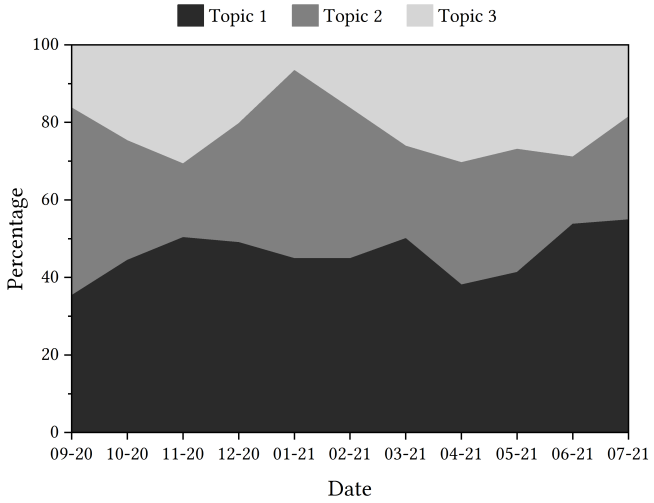


Figure 13: Trajectories of topics detected in tweets containing 'NHSCovid19App'.

was dominating the social media discourse about it too. It was also of note that 'serco' was the most common word associated with Topic 2, which could show concern for who was responsible for the design and implementation of the app. With the plotting of each topic's prevalence in the discourse for each month, it could be seen that Topic 1 was the topic that had been detected in tweets most consistently, although tweets were more concerned with Topic 2 at the time of the app's launch. In November 2020, there appeared to be a rise in tweets that discussed Topic 3, and this was the same again in April 2021.

The sharpest monthly increase in a topic's discussion was the rise in tweets discussing Topic 2 in January 2021. When attempting to relate this contextually to the state of the pandemic in the UK, the shift to a more prominent Topic 2 came at the same time as the second nationwide lockdown and, thus, tweets that discuss the government may have increased. This was evidenced when looking at the human review sampled tweets. As shown in Figure 13, the greater discussion of Topic 2 came at the expense of Topic 3, which was less prominent in tweets in January 2021. This suggests that tweets were less concerned with downloading the app and more concerned with the develop-

ment of the app by the government and Serco at this point in time. Again, this was seen in the human review sample.

#### 5.3.1.2 *Topic 1: How the App Functions*

In the context of the associated lexical items, detailed in Figure 13, it becomes apparent that the functionality of the app was discussed around themes such as isolation, positive test reporting and contact tracing. Topic 1 underscored a focus on measures to control the pandemic through the app, encompassing protocols for isolation, management of positive cases and contact tracing endeavours.

The trajectory of Topic 1 depicted fluctuations in its prominence over time. Initially, during September 2020, Topic 1 held a moderate level of significance compared to other topics. However, its importance gradually escalated in the subsequent months, reaching its zenith in July 2021. This trajectory indicated a growing relevance of Topic 1 within the discourse, mirroring the evolving nature of discussions during the pandemic period. As the pandemic unfolded and various other were implemented and adapted, discussions pertaining to these measures gained increasing prominence.

#### 5.3.1.3 *Topic 2: App Development*

The discussions around the app development appeared to be concerned with the development of the app, including those who were responsible for it. Specifically, words associated with the topic underscored discussions concerning government policies and data management practices, particularly in the context of technology and digital infrastructure. Notably, the term 'serco' stands out, suggesting discussions potentially related to government contracts or outsourcing, which may have influenced the fluctuations in Topic 2's trajectory over time.

The trajectory of the discussion of the app development reveals fluctuations in its prominence over the observed months. Initially, Topic

2 exhibited a relatively high level of prevalence, indicating significant discussions centred around governmental involvement. However, this prominence gradually diminished in subsequent months, with Topic 2 experiencing fluctuations in its proportional representation compared to other topics. Despite this, the topic retained a noticeable presence throughout the period under analysis. This trajectory underscores the enduring relevance of these themes in public discourse, highlighting the importance of continued scrutiny and analysis in these areas.

#### 5.3.1.4 *Topic 3: Obtaining the App*

An analysis of the lexical items linked with the idea of obtaining the app indicated that they were concerned with how to utilise technology to accomplish this. These terms suggest a discourse centred around technological aspects, particularly around downloading, protecting and managing data. Moreover, this indicated discussions potentially about technology companies.

The trajectory of Topic 3 depicted fluctuations in its significance over time. Initially, in September 2020, Topic 3 maintained a relatively low level of importance compared to other topics. However, its significance gradually increased in the subsequent months, notably peaking in March 2021. This rise in prominence corresponded with more companies using the app as a primary method of contact-tracing, potentially forcing citizens to use the app if they wished to use an organisation's services.

#### 5.3.1.5 *Human Review and Critical Reflection*

After this analysis of the topics, two blind human reviews were completed. A random sample of 10 tweets per month (110 total) was selected and categorised according to the pre-defined topics that were generated. The reviews found a 57% match between the human reviews and the automated topic labelling. Inter-annotator agreement

(measured by Cohen's Kappa) was 0.525, indicating moderate agreement according to Viera and Garrett (2005). This agreement rate may have been due to the broad topics. In this, common errors included labelling of Topic 1 when the automated labelling suggested it would be Topic 2 (and vice-versa).

The critical reflection was as follows:

SUNSHINE This tool facilitated the identification of frequently co-occurring lexical items, aiming to unveil latent topics. Moreover, the gensim tool proved to be user-friendly.

RAIN Limited explicit guidance on interpreting the output meant that researchers may need to speculate on the potential themes. This could potentially hinder comparisons with findings from similar studies.

LIGHTNING Certain words recurred across various topics, underscoring the significance of context. For instance, the word 'work' may signify 'function' or 'paid labour' depending on its context within different topics.

FOG While the compilation of a lexicon associated with topics encourages interpretation, the process itself remains elusive. Given that LDA operates on frequencies autonomously, the challenge lies in how humans derive meaning from the results.

### 5.3.2 *Sentiment*

#### 5.3.2.1 *Expectations and Findings*

The expectation of using sentiment analysis was to gain an overview of the discourse and see whether trajectories aligned themselves with contextual factors occurring simultaneously. The sentiment analysis

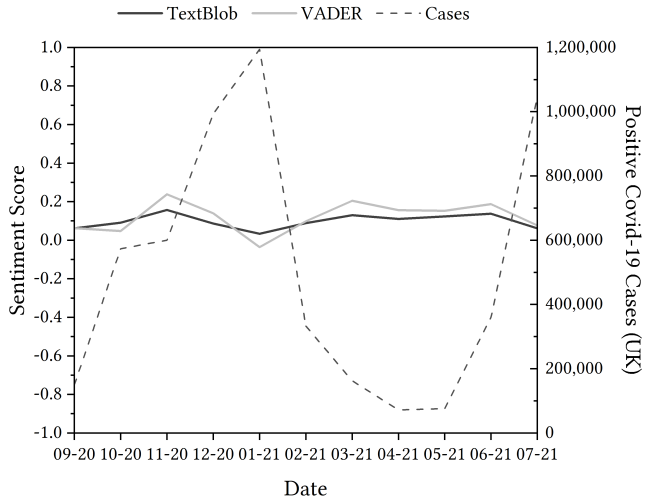


Figure 14: Evolution of the sentiment of tweets containing 'NHSCovid19App' using TextBlob and VADER from September 2020 to July 2021.

results should say whether the overall feeling of this discourse is positive, negative or neutral.

From the TextBlob sentiment analysis, Figure 14 shows that overall sentiment was 0.03 to 0.16, indicating that overall sentiment is slightly above neutral. All sentiment findings are available in the University of Nottingham Research Data Management Repository.

After applying TextBlob sentiment analysis, data was presented as a chronological trajectory. The general trend saw positivity detected within tweets rise from September to November 2020, only for tweets to be categorised as more negative in both December 2020 and January 2021. Positive sentiment detected rose again in February and March 2020, dipping slightly in April, but rising again in May and June. Tweets were deemed less positive comparatively in July, with sentiment changing from 0.14 in June to 0.06 in July.

Comparing sentiment detected in tweets relating to the app to the wider context of the UK's history with the pandemic was the next step in the process. For this, an examination of positive Covid-19 cases was undertake and shown against the sentiment detected in tweets, depicted in Figure 14. These showed similar inverted trends –

sentiment for the app is more negative as the number of cases rise in January, and sentiment for the app is more positive as the cases lower in the spring months. Note that this is not to suggest causation here, but to highlight the importance of wider contextual data which are vital to interpret sentiment analysis results at the time at which the data was recorded. This is of particular importance as a sole reliance on computational analysis may overlook underlying factors influencing public perceptions, potentially leading to misinterpretation of the findings.

Additionally, from the human review sample selected, there were a number of tweets which had been classed as positive (1.0). Of these tweets, some had this classification but, upon inspection, it is likely that other human classifiers would agree regarding making an alternative decision, and decide that they were expressing mainly negative sentiment instead. An aspect of language that these tweets have in common is the inclusion of negation within their syntactical structures (e.g., "I don't"; "can't say"; "not to be")', together with words that are taken without negation could be interpreted by an algorithm as very positive (e.g., "trust", "best", "impressed", "proud"). Of course, tweets that could be interpreted as sarcastic were also of note here for similar seemingly inaccurate detection reasons. Additionally, several tweets in the data sets were all categorised as tweets that had strong positive sentiment detected when classified using the TextBlob module but were categorised as negative when examined in context.

The seemingly incorrect categorisation of these tweets could mean that the reliability and accuracy of the TextBlob sentiment analysis tool may be questioned. Therefore, another sentiment analysis module was deployed for further investigation. Due to the fact that the VADER sentiment analysis module states that it accounts for negation, it was a logical next step to see whether there was a difference in the categorisation of the dataset, but also with these focused tweets more specifically.

When comparing the two, as shown in Figure 14 the general trend of the data when comparing the VADER sentiment analysis to the TextBlob sentiment analysis is similar, but varies in two different places: October 2020 and May 2021. There are opportunities to investigate these differences in sentiment polarity detection through a critical lens, but it could be said that sentiment analysis alone might offer limited capability to do this without the inclusion of a qualitative method in conjunction. That being said, the comparison could still be argued as noteworthy: the polarity of VADER appears to be more extreme than TextBlob – seeing deeper rises and falls in sentiment detected in tweets.

With regard to the tweets that were categorised as positive by TextBlob, despite finding negation in the tweets, these were classified as -0.4023, 0.6369, 0.4767 and 0.2924 (rather than 1.0). This suggests that that there is some improvement in detecting negation within VADER. Yet, both VADER and TextBlob and other easily accessible sentiment analysis models may still benefit from further language accuracy improvements. Also, perhaps more importantly, users of these tools would benefit from greater transparency regarding these limitations of the models, specifically with regard to negation accuracy, as just pointed out.

Moreover, for tweets that may be sarcastic, the sentiment detected was still positive with the VADER module, with values of four particular tweets being 0.8316, 0.7622, 0.6767 and 0.6958 (rather than 1.0). Again, this suggests that, even though the VADER system was potentially able to detect sarcasm better than TextBlob, both systems may benefit from further development. Additionally, users would benefit from more transparency regarding these important limitations which would ultimately affect the accuracy and reliability of the analysis provided by the tool.

#### 5.3.2.2 *Human Review and Critical Reflection*

For this human review, 10 tweets per month (110 total) were randomly sampled and classified by two reviewers according to whether they were positive, negative or neutral to ascertain EmoLex accuracy. The human review score matched the computer-assigned sentiment category on 50% of occasions. In terms of individual categories, the reviewers mostly disagreed with the tweets classified as 'positive', with only 22/54 occasions matching the algorithm-generated category. The inter-annotator agreement was 0.62, indicating substantial agreement. Between the reviewers, classifying tweets that the algorithm deemed 'positive' caused the most disagreement, with the reviewers not matching on 16/54 occasions. However, neutral tweets also caused disagreement, with 5/18 classifications not matching between reviewers.

Between the two reviewers, 33.6% of tweets contained negation structures. According to the human reviewers, 56.7% of these tweets containing negation structures were classified incorrectly by TextBlob. When considering sarcasm, 9% of the tweets reviewed were labelled as sarcastic. 90% of these tweets were classified incorrectly as positive by TextBlob according to the human review.

The critical reflection for this subsection was as follows:

SUNSHINE Utilising this capability facilitated swift analysis of vast datasets through a straightforward coding process. Integrating VADER into sentiment analysis simplified tool comparison and enabled the rapid identification of sentiment trends over time, such as those shown in Figure 14, aiding in pinpointing potential 'turning points' for further qualitative investigation.

RAIN Unresolved challenges, such as negation and sarcasm, hindered the accuracy and robustness of sentiment analysis techniques

and findings. Interpreting individual 'sentiment scores' in isolation lacked meaningful context.

LIGHTNING Again, variability in outcomes between TextBlob and VADER raised questions regarding the reliability of each tool for large-scale sentiment analysis.

FOG The challenge in interpreting sentiment analysis data laid in the lack of context surrounding sentiment scores, hindering clarity in extracting further insights.

### 5.3.3 *Emotions*

#### 5.3.3.1 *Expectations and Initial Findings*

The expectation of using emotion detection was that using this method would allow an insight into how people felt about the app and whether there were any shared or common emotions expressed. It was also expected the results be explicit as to which emotions were more prevalent at certain times within the longitudinal discourse.

For the next step in the process, the data presented in the trajectory displayed in Figure 15 shows that 'trust' was the emotion most detected in tweets relating to the app, followed by 'fear' and then 'anticipation'. By examining the percentage of tweets that detected each emotion, the rise in tweets related to 'fear' at the end of 2020 could be deemed to be of interest, as could the rise in tweets related to 'anticipation' in the spring of 2021. All emotion findings are available in the University of Nottingham Research Data Management Repository.

When zooming in, the consistency of 'trust' being an emotion that is detected within tweets relating to the app means that it is a prominent emotion of discussion. With the premise that 'trust' and 'fear' are separate emotions, it might be assumed that the tweets relating

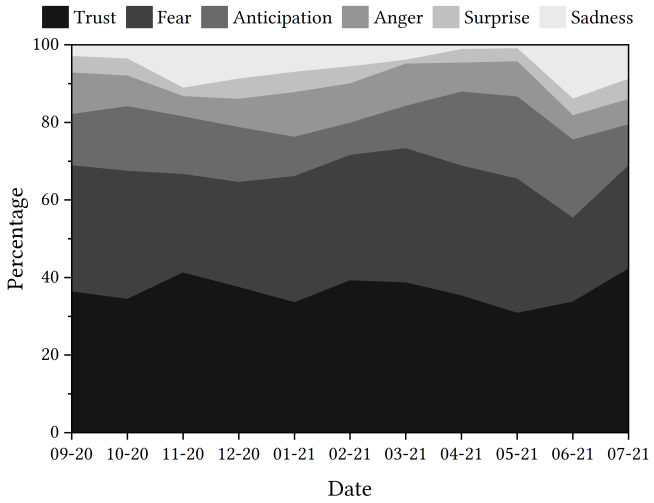


Figure 15: Emotions detected in tweets relating to 'NHSCovid19App'.

to 'trust' are indicative of the person trusting the app, rather than not trusting it. This thought guided the next part of the process.

A total of 9,796 tweets had been classified as containing the emotion 'trust' within them; again, this would have been categorised differently due to the opposition to trust noted within each tweet. The tweets included the word 'trust', which had been negated using the contracted modal verb 'wouldn't'. This may have not been detected by the EmoLex module and, instead, the word 'trust' superseded any other vocabulary as the classifier made its decision. This is similar to the negation issue using TextBlob for sentiment analysis. However, there are other tweets that have been noted as having 'trust' detected within them that do not contain the word 'trust'. Although it may be possible to argue that the words 'faith' and 'encourage' could hold some similarity in meaning with the word 'trust', there is no explicit mention of 'trust' within these tweets and a human might categorise them as expressing a lack of trust in the app, again as negations are present (e.g., "not downloading", "haven't got faith"). This further demonstrates a possible potential limitation of using tools such as EmoLex for emotion detection.

#### 5.3.3.2 *Human Review and Critical Reflection*

For the Covid-19 App, ten tweets per month (110 total) were randomly sampled to be reviewed. The categories to be assigned were 'trust', 'fear', 'anticipation', 'anger', 'surprise', 'sadness', 'disgust', 'joy' and 'no emotion'. Reviewers matched the EmoLex assigned category on 25% of occasions. The inter-rater reliability was 0.44, indicating moderate agreement. These agreement levels may be lower than the topic and sentiment agreement levels due to the range of emotions available to classify with. Within this, between the reviewers, classifying tweets that the algorithm deemed as 'anger' caused the most disagreement, with the reviewers not matching on 5/11 occasions. Reviewers categorised these tweets as 'fear' or 'disgust' instead.

The critical reflection for the emotion detection was as follows:

SUNSHINE One notable success here was the swift automated detection process employed with a substantial dataset, ensuring each tweet received some form of classification.

RAIN Similar to sentiment analysis, the precision of detection posed potential challenges when implementing the EmoLex emotion detection module. Furthermore, insights into the discourse lacked contextualisation, which would have greatly aided the analytical process. Without sufficient context, emotional classifications may appear arbitrary.

LIGHTNING Once again, the predefined inclusion of 'positive' and 'negative' within EmoLex's initial emotional categories was surprising. This could have led to overlooking crucial information, as tweets may have closely aligned with other emotional categories not encompassed in the results.

FOG Clarity regarding the distinction between emotions could have been improved with this tool. For instance, should tweets associated with 'trust' only encompass those explicitly supporting the emotion, excluding those opposing it? Further exploration revealed tweets classified as containing the emotion 'trust' that may have been perceived as opposing trust if assessed by a human. Greater clarity prior to tool utilisation could have rendered a more accurate portrayal of the discourse.

### 5.3.4 *Section Summary*

The NLP-based analysis in this chapter provided insights into the discussions revolving around the NHS Covid-19 app. Through the application of LDA topic modelling, as seen in subsection 5.3.1 three latent themes emerged, shedding light on prevalent discourse elements: app functionality, government involvement and app availability. These topics exhibited diverse trajectories over time, reflecting shifts in public attention and concerns, with discussions initially centred around how the app functions, before transitioning to considerations of app development and accessibility, mirroring the evolving landscape of the pandemic. The moderate agreement observed in human review of automated topic labelling underscores the necessity for nuanced interpretation in topic analysis.

For the sentiment analysis in subsection 5.3.2, the TextBlob and VADER modules were utilised to gauge the overall sentiment trajectory of app-related discourse. Despite disparities in sentiment polarity detection, both tools unveiled nuanced fluctuations in sentiment over time, often coinciding with contextual factors such as when national lockdowns occurred and restrictions were lifted. Nonetheless, persistent challenges in detecting negation and sarcasm impacted the accuracy of sentiment classification. Human review highlighted discrepancies between automated sentiment analysis and human inter-

pretation, emphasising the imperative of enhancing model robustness and transparent reporting of limitations. Finally, emotion detection utilising EmoLex in subsection 5.3.3 similarly provided valuable insights into prevailing sentiments, with 'trust' emerging as the most dominant throughout the discourse. However, challenges in contextualisation and the potential of arbitrary classifications underscored the limitations of automated emotion detection.

Overall, whilst NLP-based techniques provided an overview of this Twitter discourse surrounding the NHS Covid-19 app, challenges pertaining to precision, contextualisation and nuanced interpretation endured. Therefore, addressing limitations these limitations via using a complementary qualitative analysis technique was necessary.

## 5.4 CORPUS LINGUISTICS AND DISCOURSE ANALYSIS

This results section presents the findings from the CL and DA examination of the discourse. Subsection 5.4.1 demonstrates the keyword analysis of the tweets in order to ascertain which word would be focused on for this part of the study. Next, subsection 5.4.2 details the timeline overview of the discourse and the frequency of grammatically active and passive presentations. Finally, subsections 5.4.3, 5.4.4, 5.4.5, 5.4.6 and 5.4.7 present the analysis of the active and passive presentations to see if the app was presented as a social actor.

### 5.4.1 *Keyword Analysis*

Table 9 shows the top ten words with the highest keyness score when compared to EnTenTen2020 (all scores to 2 decimal places). The word with the highest keyness score was *app*, which supported the initial thinking that this would play a dominant role in the discourse since all Tweets collected for this study included the expression 'NHSCovid9App' to intentionally focus on discussions revolv-

Table 9: The top ten words with the highest keyness score.

| Item | Relative frequency (per million) | | Score |
|---|---|---|---|
| | Focus corpus | Reference corpus | |
| app | 13,387.27 | 63.94 | 82.27 |
| nhs | 5,427.78 | 15.62 | 47.81 |
| download | 3,597.18 | 61.05 | 22.96 |
| covid | 1,846.74 | 4.39 | 18.65 |
| serco | 1,556.30 | 0.26 | 16.52 |
| trace | 1,613.03 | 94.14 | 14.84 |
| isolate | 1,355.90 | 4.08 | 13.99 |
| test | 3,302.05 | 159.01 | 13.14 |
| qr | 1,236.71 | 2.22 | 13.08 |
| downloaded | 1,351.74 | 11.42 | 13.03 |

ing around this system. With this in mind, the analysis proceeded as planned. A sample of the concordances, examined in conjunction with the collocational findings, is available in the University of Nottingham Research Data Management Repository.

### 5.4.2 *Presentations of The App: Timeline Overview*

This section presents a timeline of the changes in the grammatical presentation of the app, alongside the potential social implications that this had.

As part of these results, it was found that '**be**' and '**have**' were frequently occurring collocates of 'app'. Upon manual inspection of tweets containing them, the majority of these were found to be auxiliary verbs. Whenever this was the case, they were considered as multi-word expressions and analysed on the basis of their overall meaning, since they conveyed links between agency and responsibility together.

First, the frequency of active and passive verbal constructions including 'app' was examined. This overview is shown in Table 10, where the information system features actively in 97% of the clauses.

Table 10: Frequency of active and passive presentation of *app*.

| Time period | Active | Passive | Total |
|---|---|---|---|
| Sep 2020 | 3,700 | 138 | 3,838 |
| Oct-Nov 2020 | 1,935 | 65 | 2,000 |
| Jan-Feb 2021 | 1,396 | 36 | 1,432 |
| Mar-Jun 2021 | 740 | 5 | 745 |
| Jul 2021 | 6,108 | 122 | 6,230 |

However, active and passive constructions alone do not necessarily provide a full account of how the app is presented in the discourse. For example, the app could be presented actively, yet could carry limited social agency, (e.g. 'you self isolate when the app **pings** you even though you don't have to'). To avoid misinterpretations, CL and DA were combined.

### 5.4.3 *App Launch: September 2020*

The launch month of the app saw 17,759 instances of the word 'app', which was the highest engagement recorded across any month included in the corpus.

#### 5.4.3.1 *Active Presentations*

In September 2020, there were 3,700 instances of active presentation of the app. The strongest 20 collocates are shown in Table 11. Many of these active presentations evaluate the app as underperforming, especially constructions containing '**do**' (LogDice: 9.42). Tweets stated that the app '**doesn't do** its job' or '**doesn't** work'. In this sense, such instances reflected the public perceptions of the app, which was frequently deemed dysfunctional.

In September itself, 'app' and '**say**' (LogDice: 8.07) frequently co-occured in tweets which discuss the app presenting information that users struggle to understand. For example, one tweet questioned why

Table 11: Top 20 words ranked by collocational strength of 'app' + R1 in September 2020.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | do | 567 | 23295 | 9.42 |
| 2 | be | 1565 | 80974 | 9.24 |
| 3 | use | 111 | 6766 | 8.43 |
| 4 | work | 108 | 7093 | 8.34 |
| 5 | require | 39 | 655 | 8.17 |
| 6 | have | 297 | 30674 | 8.14 |
| 7 | say | 64 | 4011 | 8.07 |
| 8 | tell | 52 | 3357 | 7.90 |
| 9 | fix | 30 | 663 | 7.79 |
| 10 | launch | 25 | 724 | 7.50 |
| 11 | need | 47 | 4738 | 7.50 |
| 12 | allow | 23 | 815 | 7.35 |
| 13 | know | 44 | 5289 | 7.31 |
| 14 | think | 32 | 3154 | 7.24 |
| 15 | install | 21 | 871 | 7.20 |
| 16 | seem | 20 | 1221 | 7.03 |
| 17 | develop | 15 | 412 | 6.87 |
| 18 | go | 30 | 4620 | 6.87 |
| 19 | let | 19 | 1563 | 6.86 |
| 20 | delete | 17 | 1018 | 6.86 |

the app '**says**' they 'are in a medium risk area'. Another tweet stated that, despite going elsewhere, the app '**said**' they were 'still at home'. Occasionally, users complained that it '**says** nothing'. Tweet authors' use of the verb '**say**' suggests the app behaved (or perceived as behaving) like a human, hence an illustration of *personalisation*.

Another strong collocate of 'app' was '**tell**'(LogDice: 7.90). This had similar semantics to 'say'. However, '**tell**' was mainly used to express that the app was instructing a user to self-isolate, in both actual scenarios (i.e., upon entering test results into the app, the app '**tells**' them 'to isolate and get a test') and hypothetical scenarios (e.g. the app '**told**' them they 'had to isolate even though their boss would not allow them to without symptoms'). In other instances containing '**tell**', users questioned the reliability of the app, for instance asking whether anyone believed 'a word this app **tells** u?'. Similarly, another user was confused about 'what this app is **telling** [them]??'. Although comparable to '**say**' semantically, it could be argued that the pragmatics of '**tell**' were different. For instance, 'tell' acted imperatively when '**telling**' users 'to stay at home'. This constituted *personalisation* of the app. These examples could also be considered as *agency metaphor*, as '**tell**' implied more volitional action as an imperative compared to '**say**'.

Another way in which the app was presented actively was when users wrote that it '**needs**' something (LogDice: 7.50). In this month, users frequently tweeted about the operating system requirements for the app to function on mobile devices, whereby the app '**needs** ios 13.5' and '**needing** current ios updates or [it] won't work'. The '**needs**' of the app not only presents it actively but also gives it human-like characteristics, providing other examples of *personalisation*, hence a fuller account of the app's (public) image.

'**Allow**' also strongly collocates with 'app' (LogDice: 7.35). In these instances, users discussed the function of the app and the permissions that the app granted. For example, 'the app **allows** [them] to enter one postcode only', causing issues to people living and work-

Table 12: Top 6 words ranked by collocational strength of 'by the app' + L1 in September 2020.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | recognise | 5 | 350 | 8.70 |
| 2 | accept | 3 | 554 | 7.36 |
| 3 | isolate | 4 | 3137 | 5.36 |
| 4 | track | 3 | 2653 | 5.19 |
| 5 | tell | 3 | 3357 | 4.85 |
| 6 | use | 4 | 6766 | 4.27 |

ing in different areas. Interestingly, the user directed this grievance to the app itself, giving the impression that the app had social agency. Other occurrences of '**allow**' involved questions, for instance, asking whether the app will '**allow**' users to report themselves as testing positive, even when they are not. Another questioned if the app '**allow[ed]** for manual check in'. These questions from Twitter users reiterated concern for the app's implied agency and could potentially be seen as additional examples of *personalisation*, here expressing public attitudes of uncertainty and worry about not being able to use the app.

'**Think**' is another strong collocate (LogDice: 7.24). Occurrences of this active presentation complained about the app's performance and accuracy. For example, when visiting different places, the app could '**think**' that users were 'still at home'. Another Twitter user reported deleting the app after getting a negative test back as there was no code to input the test and the app '**thought**' they 'still had to isolate'. One questioned whether the app '**thought**' they had been 'at the old venue for all that time' when they checked into a new venue after several days. Despite these open queries and concerns, users presented the app as being able to think and act for itself. In this case, *personalisation* conveyed agency.

#### 5.4.3.2 *Passive Presentations*

The app is also presented passively on 138/3,838 occasions this month, with collocates shown in Table 12. These included discussions around whoever had created the app. '**Be**' + '**develop**' was a multi-word collocate of 'app' (LogDice: 10.00). Some constructions were questions, such as whether the app has '**been developed** by the nhs?' or the 'app was developed by serco and [...] not the nhs'. These tweets are examples of *backgrounding* of the entities that (supposedly) created the app. Instead, this presents the app as passive, yet important in the construction, as, despite the lack of grammatical agency, the focus is still on the app. This is closely linked to '**be**' + '**design**' (LogDice: 9.4). This discussion around the app's intended function was subverted in some tweets, for example, 'this app **is designed** to control sheeple'. Unlike the previously-mentioned examples, though, these represented instances of *exclusion* and removed the agent from the construction altogether. In these passive structures, Twitter users still discussed the app in a negative way, highlighting that the app's functionality is deemed unsatisfactory by its (self-declared) users.

Similarly, the 'app' collocated strongly with '**be**' + '**run**', resulting in constructions containing the passivisation of Serco. Examples included 'but then the app **is run** by a private company.' and 'have heard this nhs app **is run** by serco?!'. This indicated again that, despite the passive presentation, users were still dissatisfied with the app.

Comparable instances portrayed the app passively, through a '**has been**' + verb construction (LogDice: 4.88) to state that the app '**has been** launched in england and wales after months of delay' or that 'the government's app' '**has been** designed by a dog'. Neither of these constructions indicated who was responsible for the launch or design of the app, thus exemplifying *exclusion*, and using it to reiterate user dissatisfaction.

Other passive constructions, delivering a similar meaning, combined a verb and 'by the app' and accounted for 44/138 occurrences, in September 2020. '**Recognise**' (LogDice: 8.70) was mainly employed regarding the inputted test results into the app. For instance, one user asked whether 'only private tests will be **recognised** by the app' and another stated the simplicity of setting up code '**recognised** by the app'. In both of these short extracts, the app's passive presentation removes the agency and places it more with the app developers. In terms of agency, '**accept**' (LogDice: 7.36) is similarly featured in the data analysed. For example, one user complained that incorrectly-formatted code was 'not **accepted** by the app'. All of these instances reflected the app's perceived lack of functionality, causing public criticism, despite the passive presentation.

'**Isolate**' is another strong collocate of 'by the app' (LogDice: 5.36). Tweet authors complain about being told to self-isolate by the app. One user, for instance, questioned liability if 'notified of contact/need to **isolate** by the app'. With this, the app appears to be less of a focus in the structure and agency is removed, through passivisation and *backgrounding*. Therefore, the responsibility could possibly be transferred from the app to the user.

'**Tell**' (LogDice: 4.85), featured in complaints about people that they were being instructed by the app. An example was one user discussing a 'person at my work' who had 'just been **told** by the app to self-isolate and get a test'. This presented the app passively and limited the attention to it, with the 'person' being the central figure, although *indetermined* and *genericised*. As in the previous case, the responsibility seemed to be deflected to the user 'by' the app.

In summary, September 2020 showed many tweets presenting the app actively, especially when uncertain about how the app functioned or could assist its users. This was mainly accomplished through *personalisation*, portraying the app as if it was human. In these cases, passive presentations prominently discussed the development of the app

and attributed it to Serco, the NHS or the UK government, deflecting responsibility from the app to these organisations or app users.

### 5.4.4 *Early Months: October-December 2020*

The first three full months after the app launched saw 6,237 tweets using the word 'app'.

#### 5.4.4.1 *Active Presentations*

Active presentations of the app were seen 1,935/2,000 times, with R1 collocates reported in Table 13.

'**Use**' (LogDice: 10.42) appeared most frequently in a duplicated tweet that had been sent from different regional NHS accounts. The text in question contained the structure 'the app **uses** an algorithm to filter out false alarms'. Therefore, the NHS promoted the app as a positive social actor, in contrast with the negative presentations put forward by several members of the public, as detailed above. This implies a discrepancy in the portrayal of the app between official promotion and public perception, underscoring the complexity of trust dynamics

Similarly, many of the tweets using '**say**' (LogDice: 8.58), released over these three months (October-December 2020) were comparable to those published at the time of the app launch (September 2020). Among others, one user tweeted about discrepancies between the supposed ending to their self-isolation period, stating that their app '**said** that [their] self isolation will be ended on 25 dec 2020 at 23.59', which was 'different from what [they] have been told on text message and nhs website'. Here, the app was presented as actively informing the user, which constituted another example of *personalisation*. Interestingly, the same user states that they have been 'told *on the* text message and nhs website', rather than be told *by* the message or by the website. This distinguished the app, actively presented along-

Table 13: Top 20 word ranked by collocational strength of 'app' + R1 in October, November and December 2020.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | use | 365 | 6766 | 10.42 |
| 2 | say | 70 | 4011 | 8.58 |
| 3 | tell | 61 | 3357 | 8.55 |
| 4 | work | 84 | 7093 | 8.25 |
| 5 | be | 703 | 80974 | 8.12 |
| 6 | do | 154 | 23295 | 7.64 |
| 7 | have | 129 | 30674 | 7.02 |
| 8 | show | 12 | 1119 | 6.99 |
| 9 | give | 14 | 1920 | 6.88 |
| 10 | allow | 9 | 815 | 6.73 |
| 11 | send | 10 | 1134 | 6.72 |
| 12 | install | 9 | 871 | 6.70 |
| 13 | seem | 10 | 1221 | 6.68 |
| 14 | store | 15 | 2838 | 6.68 |
| 15 | update | 9 | 1443 | 6.43 |
| 16 | develop | 6 | 412 | 6.37 |
| 17 | fail | 6 | 446 | 6.35 |
| 18 | keep | 10 | 2083 | 6.34 |
| 19 | crash | 5 | 83 | 6.32 |
| 20 | ask | 8 | 1685 | 6.17 |

side other technological systems appearing as vessels of information, rather than agents. These different presentations reaffirmed that the app was a social actor in this context, highlighting the nuanced dynamics of agency and trust in digital environments.

'**Tell**' (LogDiceL 8.55) was used in a similar way to '**say**', similar to the tweets found in September 2020. An example of this included one user tweeting that their child had 'received a notification on the track & trace app **telling** her to self isolate', yet only for two days. Another user stated that the app '**tells**' them their 'home is medium risk' despite living in a rural area with low Covid-19 infection rates. These examples indicated that the app was providing instructions, and thus had social agency, implying that the app's influence may extend beyond merely conveying information; instead, the app is portrayed as actively shaping individuals' behaviours and perceptions of risk.

Another strong collocate was '**have**' (LogDice: 7.02). Although used as an auxiliary verb in the majority of constructions, there were occasions where it acted as the main verb to indicate possession (or lack of). For example, one tweet discussed that their relative was recovering from cancer and expressed frustration that the app had not notified them, even though they had been in contact with a positive case. Accordingly, the app '**has** one job' to keep their relative safe. This is a clear example of *personalisation* due to the idea that the app is able to perform a job, yet responsible for the safety and welfare of their relative. Similar active presentations featured '**have**' as an auxiliary verb, as when a user joked that the app '**has** decided to turn off contact tracing', implying its autonomy and control.

When the app was presented as doing the opposite of its desired function, negation was employed. A user complained that 'the app **has** not alerted [them]' despite 'living with someone who had tested positive' for the virus. Another user stated that their app '**has** not conducted exposure checks since 29 december'. Both of these examples placed the agency with the app, alluding that the app was responsible for its own shortcomings. With 'someone', this is an example

Table 14: Top 3 words ranked by collocational strength of 'by the app' + L1 in October, November and December 2020.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | notify | 3 | 385 | 7.87 |
| 2 | isolate | 5 | 3137 | 5.69 |
| 3 | tell | 5 | 3357 | 5.59 |

of an *indetermined* construct, which further removed agency from the humans and placed it with the app. This highlights a possible shift in accountability and responsibility away from human actors towards automated systems, which may imply a growing reliance on algorithms to not only perform tasks but also to shoulder the blame when things go wrong.

On other occasions, where users wrote that the app '**gives**' them something (LogDice: 6.88), one complained that the app '**gives** [them] notification about people passing by [their] house', whilst another joked that the app **gave** them 'a 3 day stay at home order'. Another mused that the app was '**giving** the govt more control over our everyday lives'. In all these occurrences, the app was presented actively through *personalisation*, showcasing the perceived responsibility of the app for controlling users' lives. The social implications here suggest a growing unease with the extent to which technological interventions dictate personal freedoms and autonomy, thereby shaping broader discussions about technological privacy and surveillance.

#### 5.4.4.2 *Passive Presentations*

When examining passive constructions, shown in Table 14, similar passive presentations to the previous month can be seen. When focusing on '**notify**' (LogDice: 7.87), tweets focused on hypothetical scenarios, with one tweet stating that they were not entitled to support should one be '**notified** by the app' as 'they can't identify you' and another that questioned the legal ramifications if one was 'only **notified** by the app' and not Test and Trace as a whole. These examples

discussed the legal and financial implications of the app directing someone to self-isolate instead of going to work. In both instances, the app was not a prominent part, hence the passivisation, and the central focus was on the impact rather than the app, indicating a shift in attention towards the consequences of the app's actions rather than its active role,

In contrast, when '**tell**' was used in passive constructions (LogDice: 5.59), many of these accounts were direct first-person narratives by app users. For example, one 'got **told** by the app [...] to isolate for 12 days'. Another explained they have 'not been **told** by the app to isolate', even after their family member tested positive. In these cases, the authors recounted that they were provided with a service by the app, *backgrounding* the importance of the system in the process. Instead, these accounts tended to focus on getting answers from humans, which the app could not provide, suggesting a lack of trust or confidence in the app's capabilities.

Some '**have**' constructions were passive too. For example, one user wrote that they were at risk as the app '**has** not been created to include old smartphones'. This passive construction implied that the app had been created by an unknown agent, thus *exclusion*. Although this passive construction partially removed agency from the app, the fact that technology was mentioned explicitly in the tweet could still foreground the system as a social actor, implying that the app was still perceived to be involved in the situation or process described.

### 5.4.5 *Second National Lockdown: January-February 2021*

#### 5.4.5.1 *Active Presentations*

In this period, 1,396 active presentations of the word 'app' were collected, as shown in Table 15. '**Ping**' (LogDice: 7.98) was used actively to mean notify, with examples such as one user stating that 'everyone knows it was your app **pinging**' and another writing that the

Table 15: Top 18 words ranked by collocational strength of 'app' + R1 in January and February 2021.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | ping | 7 | 455 | 7.98 |
| 2 | tell | 23 | 3357 | 7.63 |
| 3 | cost | 4 | 323 | 7.40 |
| 4 | notify | 4 | 385 | 7.29 |
| 5 | say | 18 | 4011 | 7.05 |
| 6 | state | 3 | 366 | 6.91 |
| 7 | use | 24 | 6766 | 6.77 |
| 8 | work | 22 | 7093 | 6.58 |
| 9 | alert | 7 | 2050 | 6.52 |
| 10 | show | 4 | 1119 | 6.38 |
| 11 | be | 189 | 80974 | 6.25 |
| 12 | keep | 5 | 2083 | 6.01 |
| 13 | seem | 3 | 1221 | 5.88 |
| 14 | have | 43 | 30674 | 5.50 |
| 15 | need | 7 | 4738 | 5.47 |
| 16 | do | 26 | 23295 | 5.16 |
| 17 | track | 3 | 2653 | 4.98 |
| 18 | store | 3 | 2838 | 4.90 |

app '**pings** you because you walk past someone in the street'. These tweets suggested that the app was acting autonomously and had its own agency, through *personalisation* and, in the case of 'someone', an *indeterminism*.

Additional instances of the app '**telling**' (LogDice: 7.63) recounted personal experiences and fewer reported hypothetical scenarios. Examples included one user stating that the app '**tells**' them they 'have to isolate' from ten days after the initial encounter date. Another questioned why the app was '**telling**' them 'to isolate for 14 days' when they believed it to be ten days instead. Overall, the app was presented actively in these scenarios. Therefore, should someone be affected by Covid, the app may be more likely to be presented actively.

Table 16: Top 4 words ranked by collocational strength of 'by the app' + L1 in January and February 2021.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | ping | 7 | 455 | 8.90 |
| 2 | alert | 3 | 2050 | 5.57 |
| 3 | isolate | 3 | 3137 | 4.96 |
| 4 | tell | 3 | 3357 | 4.86 |

Hypothetical instances questioned the legitimacy of the app, such as one user hypothesising why other individuals were self-isolating when they had no symptoms because 'an app **told** you to'. This still presented the app as an implicated social actor. This could be seen to lessen the impact of the app, although presented actively, and may doubt the functionality of the system as a whole, too.

Other active presentations that implied the app had social agency removed he idea of instructing people to self-isolate. For instance, one author tweeted about the app 'is **creating** a notification that has been stuck' on their screen for a long time. The idea that the app was '**creating**' a notification may further position it as a social actor. Instead of using the verb 'notify', the author word-class converted this to the noun 'notification', using it in conjunction with a more *personalised* verb, '**create**'. Therefore, this clearly indicated agency and placed responsibility on the app to self-regulate, through *agency metaphor*.

Authors using '**do**' (LogDice: 5.16) discussed the app's failed expectations. An example included one user writing that, despite their partner testing positive, 'the so called world beating app **didn't** alert [them]'. This active construction indicated that the app was perceived as responsible for their safety.

#### 5.4.5.2 *Passive Presentations*

When considering passive presentations, shown in Table 16, tweets released in January and February 2020 were concerned with an in-

dividual being literally or hypothetically instructed by the app, for example, '**ping**' (LogDice: 8.90). Being '**pinged** by the app' was 'as reliable as a handbrake on a canoe'. According to another user, they had to isolate for ten days after they 'got **pinged** by the app'. Other collocates like '**alert**', '**told**' and '**isolate**' also followed similar patterns. This culminated in Twitter users potentially seeing the app as exemplifying the unreliable government handling of the Covid-19 pandemic, shifting responsibility from the app to these organisations, instead.

The app was also presented passively in conversations about its producers. For example, one tweeted that they resided and worked in an area where the infection rate was high, yet 'the app **has been triggered** once in its 4/5 months existence'. Although this tweet presented the app passively, it placed the blame on the creators of the app, without even mentioning them, applying a so-called reverse-*exclusion* strategy.

### 5.4.6 *Later Months: March-June 2021*

#### 5.4.6.1 *Active Presentations*

Between March and June 2021, 740 active presentations of 'app' were found in the dataset collected, as shown in Table 17, numerous of which presented the app as a social actor.

One of the strongest collocates from these months, '**provide**' (LogDice: 7.22), described the app as helpful. An example of this was a tweet that stated that the app '**provides** anonymous information including risk alerts by postcode, a symptom checker and test booking', which came from a devolved local NHS Twitter account. Here, the app was presented as a social actor, supporting the idea of system confidentiality and positively evaluated. Up to this moment, when the app had been portrayed actively, it had usually been negatively connotated. However, this was not the case for all instances

Table 17: Top 6 words ranked by collocational strength of 'app' + R1 in March, April, May and June 2021.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | provide | 3 | 525 | 7.22 |
| 2 | help | 5 | 4271 | 5.22 |
| 3 | be | 68 | 80974 | 4.78 |
| 4 | work | 6 | 7093 | 4.77 |
| 5 | have | 16 | 30674 | 4.09 |
| 6 | do | 8 | 23295 | 3.48 |

of '**provide**', with other examples including one user that questioned why the app did not '**provide** update information' about local infection levels, whilst another user stated that 'the app **provides** little to no information', thus indicating dissatisfaction with the app's performance.

Similarly to the majority of '**provide**' occurrences, '**help**' (LogDice: 5.22) was mainly seen in advertisements from devolved NHS Twitter accounts. In these cases, tweets contained constructions such as the app '**helps** stop the spread of the virus'. Therefore, this presented the app as having a positive impact on society at large.

Although not a significant enough collocate to meet the minimum threshold, authors used '**tell**' in conjunction with '**be**', when discussing the app. Instances included that the app was '**telling** [them] 10 days from the 26th instead 20th', and another wondering how long they needed to isolate for, particularly if it was 'just the 2 [days] that the app is **telling** [them]'. Both of these examples could be categorised as a query about the lack of clarity that the app reflected as the rules about self-isolation were changing. As both constructions showed the app to be active, this not only added to the evidence of the app being presented as a social actor, but it also contributed to the discourse surrounding questions over the functionality of the app itself.

Authors of the tweets collected used '**have**' (LogDice: 4.09) to present the system in an active way, with examples of tweets includ-

ing 'not only **has** the app failed me [...] it **has created** a problem for me', indicating that responsibility is attributed to the app. Another interesting presentation discussed the app as only guidance, as it '**has** no legal force'. Here, the app is presented actively, yet the content of the structure could be argued to mitigate or remove social agency from it. Therefore, this suggested a decrease in the system's responsibility and control.

#### 5.4.6.2 *Passive Presentations*

Due to the small number of passive presentations (5/745) from March to June 2021, collocation analysis would not be meaningful. However, upon manual inspection, these constructions were concerned with scenarios that did not involve the tweet authors. For instance, one discussed a friend who '**has** been told by the app to stay in for 3 days'. These tweets foregrounded the importance of the experience of the general public, through *genericising indetermining*, and *backgrounding* the app.

### 5.4.7 *'Pingdemic': July 2021*

#### 5.4.7.1 *Active Presentations*

There were 374 active occurrences of 'app' in the 'NHSCovid19App' dataset and a further 5,734 occurrences in the 'pingdemic' dataset (total 6,108/6,230). The collocations are shown in Table 18.

The strongest collocate was '**disagree**' (LogDice: 10.39). However, upon manual inspection, this was a headline that had been quote-tweeted multiple times. The variations of the headline read 'U.K. Leaders Hail a Return to Normal; Their Phone App **Disagrees**' and 'Britain's contact-tracing phone app **disagrees**, telling huge numbers of people to self-isolate'. The idea that the app was able to disagree with powerful human entities exemplified *personalisation*. Despite this

Table 18: Top 20 words ranked by collocational strength of 'app' + R1 in July 2021.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | disagree | 76 | 96 | 10.39 |
| 2 | send | 45 | 451 | 9.38 |
| 3 | tell | 57 | 1535 | 9.15 |
| 4 | be | 845 | 47226 | 9.14 |
| 5 | have | 165 | 12256 | 8.59 |
| 6 | ping | 58 | 3351 | 8.54 |
| 7 | do | 93 | 6490 | 8.53 |
| 8 | beg | 19 | 53 | 8.43 |
| 9 | work | 36 | 1752 | 8.39 |
| 10 | fail | 13 | 478 | 7.57 |
| 11 | delete | 16 | 1057 | 7.54 |
| 12 | install | 10 | 64 | 7.49 |
| 13 | start | 13 | 917 | 7.32 |
| 14 | cause | 19 | 2259 | 7.28 |
| 15 | alert | 10 | 585 | 7.13 |
| 16 | say | 19 | 2778 | 7.10 |
| 17 | use | 12 | 1352 | 6.98 |
| 18 | force | 9 | 672 | 6.92 |
| 19 | go | 15 | 2327 | 6.91 |
| 20 | design | 7 | 223 | 6.86 |

coming from only two sources, the high number of shares indicated that others engaged with the idea.

Another strong collocate was '**send**' (LogDice: 9.38), in reference to the app sending a total of approximately 600,000 notifications to self-isolate. One tweet stated that the app '**sending** too many spurious notifications will reduce compliance'. In this instance, the author presented the app as an active social actor, since the cause-and-effect relationship between the app and the members of the public, further implicated the app as an agent of change, showcasing it as a perceived responsible actor by the users.

The recorded resurgence of '**ping**' (LogDice: 8.54), in July 2021, was likely due to a new blended term for the increase in exposure notifications. In these instances, the app was presented actively, as performing actions ranging from matter-of-fact reporting ('NHS Covid app **pinged** 600,000 more people') to the nonsensical ('Every time a Covid app **pings** Boris Johnson loses one of his wingdings'). In each of these occurrences, the app was still presented as having agency and being a social actor through *personalisation*, hence depicted as causing frequent disruptions.

One occurrence where the app was presented as actively, '**pinging**' expressed disdain towards the members of the public who 'self isolate when the app **pings** [them] even though [they] don't have to', hence suggesting that they 'will blame the government for [their] own decisions'. Despite the active presentations of the app, its impact as a potential social actor was mitigated through the sarcastic tone of the author. Arguably, members of the public who used the app should be accountable for their own actions rather than blaming the app. Other tweets appeared to support this view, such as one stating that 'it's not a 'pingdemic" as the app was '**pinging** ppl correctly', and another that detailed the app was '**pinging**' because 'it is doing its job.'. All these instances illustrated different ways in which responsibility can be attributed to entities other than the app.

Table 19: Top 6 words ranked by collocational strength of 'by the app' + L1 in July 2021.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | ping | 33 | 3351 | 7.02 |
| 2 | alert | 5 | 585 | 6.82 |
| 3 | contact | 6 | 1084 | 6.19 |
| 5 | cause | 9 | 2259 | 5.72 |
| 6 | isolate | 4 | 2063 | 4.68 |

'**Tell**' (LogDice: 9.15) once again revolved around instruction to self-isolate. One user wrote about how the app 'is telling people to self isolate' because of higher infection rates. This contrasted with other experiences, such as another user asking whether the app can '**tell**' them when they are 'supposed to have been near an infected person'. Another user wrote 'the app **told** something like 700,000 to isolate', which resulted in allegedly instructing supermarket staff to isolate as they had mobile phones at work. This presented the app as a social actor and perhaps as if it had human-like agency, through *personalisation*. Active presentation of the app was clear in these cases as it demonstrated the system's capacity to instruct, thus having a social impact.

Through many other active presentations, the app caused disruption. '**Wreak**' (LogDice: 6.52) was used when users said the app '**wreaks** havoc'. Similarly, '**cripple**' (LogDice: 5.70) featured in constructions like the 'app **cripples** Britain'. Also, '**threat**' (LogDice: 5.50) featured in a tweet stating that the app '**threatened** to bring parts of the economy to a standstill'. All of these constructions presented the app as destructive and capable of creating harm, thus being responsible for social disruptions.

#### 5.4.7.2 *Passive Presentations*

Passive constructions were more frequent in July (122) than previously (5). The collocational strength of these is shown in Table 19.

However, due to the greater volume of tweets in this part of the discourse, this was proportionally lower than in September 2020. '**Ping**' (LogDice: 8.30) showed users speaking hypothetically once again. One user questioned how society would cope 'if everyone **pinged** by the app asked for a PCR test', whilst another user stated that 'if you get **pinged** by the app you shouldn't need to self isolate'. This may suggest that the focus was on the humans affected by the app, rather than the app itself. This seemed to limit the system's social agency through *backgrounding*. Similarly, collocates '**alert**', '**contact**' and '**isolate**' were found in tweets surrounding with the same idea.

Conversely, another strong collocate, '**cause**' (LogDice: 6.98), was used was used differently. While the app was seen to '**cause**' damage and chaos during the 'pingdemic', in active constructions, the passive presentations removed agency from the app. Examples included one user writing that 'staff shortages have NOT been **caused** by the App' and another stating that the UK government was to blame, hence 'it's not a "pingdemic" **caused** by the app'. Finally, another tweet built on this and criticised the media outlet The Daily Star for 'adopting the right-wing press's line that the "pingdemic" is **caused** by the app'. This not only removed the grammatical agency from the app but also mitigated its social agency, by making other entities appear more responsible.

### 5.4.8 *Section Summary*

From these results, evidence suggests that the app was presented in a predominantly active way (97% of occurrences – 13,879/14,245 constructions), although some active presentations gave the app more social agency than others. Approximately 100 carried less agency through mitigating activity in either verb constructions or other contextual information. This indicates that the app was presented as a social actor in approximately 96% of the cases considered

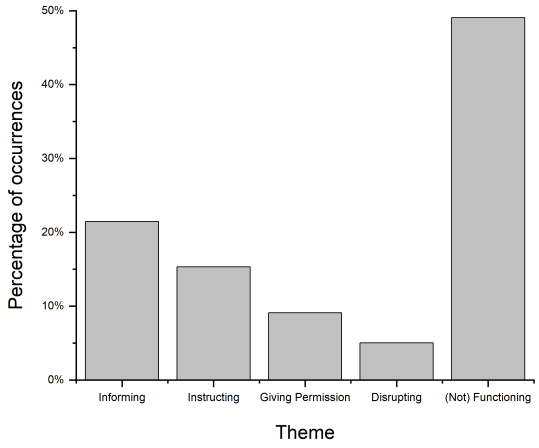


Figure 16: Comparison of the percentage of each theme found when the app is presented as a social actor.

(13,779/14,245). This examination showed that the 13,779 active presentations, where the app constitutes a social actor, can be split into five broadly recurring themes: the app informing (21.47%), the app instructing (15.33%), the app giving permission (9.1%), the app disrupting (5.02%) and the app functioning or not functioning (49.07%), displayed in Figure 16. In order to answer the research question, the discussion will elaborate on links between these constructs and their relationship between what is present in the discourse and what is present in previous literature.

## 5.5 DISCUSSION

In this section, the findings from the topic, sentiment and emotion trajectories are discussed. The analysis-based findings regarding active and passive presentations will be discussed, whilst addressing the research questions about agency displays and Twitter users' perceptions of app responsibility. By doing this, trends in these presentations will be examined, alongside how they developed in the discourse and how this ties with existing relevant research. Finally, this

section will discuss the limitations of this study and recommendations for possible future work, in light of those.

### 5.5.1 *Topics, Sentiment and Emotions*

Three latent topics were identified: Topic 1 on contact tracing and isolation, Topic 2 on the government and data management, and Topic 3 on app download and availability. In January 2021, there was a significant increase in discussions related to Topic 2, likely due to concerns about the government's role in the app's development during the second nationwide lockdown. This increase in Topic 2 discussions appeared to coincide with a decrease in prominence of Topic 3 discussions, indicating a shift in focus. However, due to limited guidance on interpreting the groups of words, the accuracy of topic labeling may be affected. The notion of functionality being present in the topics resonates with previous work that, despite scientific intentions, user-reported technical issues such as backwards incompatibility and incorrect alerts (Kent, 2020). This is a clear flaw in the functionality of the app. Additionally, with the government being a heavy factor in Topic 2, this relates to the previous findings of Pepper et al. (2022), who discovered that, over time, the government's management of the app waned, contributing to a decline in app compliance.

Overall tweet sentiment ranged from 0.03 to 0.16, slightly above neutral. Chronologically, positivity increased from September to November 2020, but turned more negative in December 2020 and January 2021. Positivity rose again in February and March 2021, with a slight dip in April, and then increased in May and June. Tweets in July were comparatively less positive. Regarding previous literature, public sentiments towards the app have been scrutinised and studies revealed mixed reactions influenced by factors such as privacy concerns, stigma and perceptions of the app's societal benefits (Williams et al., 2021). However, the presence of negation and sarcasm made

classification challenging, potentially impacting the data. Additionally, interpreting sentiment scores, especially with very similar scores, proved difficult.

For emotion detection, emotions that were the most frequently detected in this discourse were 'trust' and 'fear'. Further investigation of this as a trajectory showed that this fluctuated as 'anticipation' became more prevalent. However, it was apparent that the direction of trust was not clear, which, once again, could have skewed the data. This emotional complexity resonates with findings from existing literature, which sheds light on the intricate interplay between public sentiment and the app's societal impact. For instance, Dowthwaite et al. (2021) and Pepper et al. (2022) uncovered fluctuations in trust dynamics over time, indicating a nuanced evolution of attitudes towards the app. However, the results found here cannot explicitly say whether this conflicts or supports these exisitng studies due to the challenging interpretation of the direction of trust. Additionally, Paucar et al. (2022) highlighted the role of trust and responsibility in shaping public acceptance, underscoring the contextual and temporal nature of emotional responses, yet, once again, the lack of certainty regarding the direction of trust is an issue. This further solidified the need for CL and DA as complementary approaches to the analysis.

### 5.5.2 *Trends of Active Agency*

Through the analysis of transitivity in the 14,425 concordance lines considered, the collocations of 'app' and 'by the app' and DA-informed analysis of agency and responsibility, underpinned by SAR, five main categories were identified that the active presentations of the app fall into: informing (21.47%), instructing (15.33%), providing permission (9.1%), disrupting (5.02%) and functioning (49.07%). The first four of these categories show the app to be *personalised* (Van Leeuwen, 2008) and make decisions independently (Richard-

son, Mueller, and Pihlaja, 2021). Meanwhile, functioning included instances of the app acting autonomously but also simply functioning as intended or designed to do. Additionally, this category included tweets where the app was functioning appropriately, but also contains tweets where it was presented as not functioning as desired. This may explain the large percentage of tweets that fell into this category.

#### 5.5.2.1 *Informing*

The app was presented actively (21.47% of occurrences) when providing information to its users through '**saying**' and '**pinging**'. This happened especially at the start of the discourse, as the app was '**saying**' information that was difficult to understand (LogDice: 8,07). Similarly, users complained about the app informing them (41/64 occurrences), determining a trend that followed into other areas of the discourse. An example of this was in the 'early months' part of the timeline, where the app communicated about the status of their self-isolation period (LogDice: 8.58). In this sense, the idea of presenting information linked to the findings of the study by Williams et al. (2021), whilst also pointing out that the app proving information to users was deemed a core responsibility of the app, hence the questions and negative reactions when the app that '**failed**'.

The app was presented actively informing also through the surge in '**ping**', a frequent collocate of 'app' between January and February 2021 (LogDice: 7.98). The tweets containing these collocates depicted the app as acting autonomously, once again leaning into *personalisation*. Although some instances later in the discourse presented the app as providing useful information, the majority of presentations still remained negative when it came to the information given – or not – to users (LogDice of '**provide**' in March-June 2021 being 7.22). '**Ping**', clearly continued being a verbal trend into July 2021, with many tweet authors discussing the impact that the app informing them of a Covid-19 positive or isolation status (LogDice: 8.54). This is complemented by strong collocations of '**tell**' (LogDice: 9.15) and '**say**'

(LogDice: 7.10). Therefore, it could be inferred that the app's active presentation, using '**ping**', and the perception that it might have provided incorrect information, contrasted with the rationale for having a decision-making algorithm in the first place (Busch and Henriksen, 2018).

#### 5.5.2.2 *Instructing*

The app was also seen as actively providing instructions to users and the wider public (15.33% of occurrences). '**Tell**' was a frequent collocate of 'app' throughout the discourse (LogDice scores of 7.90, 8.55, 7.63 and 9.15 respectively), and users presented this as the app instructing them to take action, most notably, to self-isolate (232/270 occurrences). At the beginning of the discourse, 30/52 occurrences of these were hypothetical, likely due to the app being newly launched. Twitter users also questioned the instructions provided by the app (12/52 occurrences). This imperative tone continued in the final months of 2020, with users stating that the app was instructing them to self-isolate (44/61 occurrences). However, what became more apparent in this section of the discourse was that, although the app was presented as a social actor through *personalisation*, the impact that this system had was ridiculed through humorous additions to tweets or sarcastic tone (9/61 occurrences). This likely softened the instructional impact that the app had, whilst still presenting it as a social actor.

The app continued to be presented as actively instructing in later parts of the discourse too, with direct first-person accounts of experiences when the app '**is telling**' users (LogDice: 9.15), as well as reports of hypothetical scenarios (14/23 occurrences). These constructions exhibited human-level agency, through *personalisation*. The other constructions seen in July 2021, such as ones that contain '**wreak**', '**cripple**' and '**threat**', implied equally a significant level of agency as if the app's instructions could only result in negative consequences, although this will be explored in more detail in the 'disrupting' sec-

tion. The presentation of the app in this way intersected with concerns around the merging of algorithmic and human agency (Beer, 2017; Crang and Graham, 2007; Meisner, Duffy, and Ziewitz, 2024; Ziewitz, 2016) due to the app being presented as performing the job of a human. Particularly, the app featured in constructions where users were frustrated with its instructions, or lack of instruction. This provides insights into the agency of the algorithm's perceived role and responsibility.

#### 5.5.2.3 *Giving Permission*

Although less common than the previous two categories, the app was also presented actively when seemingly providing the users or general public permission (9.1% of occurrences). This is most prominently seen in users stating that the app is '**allowing**'. More present at the start of the discourse (LogDice: 7.35 for September 2020 and 6.73 for October-December 2020), due to the questions being asked of the app, this was less frequently discussed as time progressed. Sets of tweets in the discourse pointed to the app providing permission. For example, the app '**gives**' notification of self-isolation periods at the end of 2020. This recalled some of the permission concerns found in the study by Dowthwaite et al. (2021).

It could also be argued that '**need**', a strong collocate at the beginning of the discourse (LogDice: 7.50), intersected this theme and 'functioning'. The idea that the app needed to provide permission to humans was an occurrence of *personalisation*, providing further insight into the idea that the app was not only given agency but also showcasing its necessity to process information systematically.

#### 5.5.2.4 *Disrupting*

The app was presented as disrupting users' lives, through active constructions. This is seen early in the discourse when users commented on the app making disruptive decisions autonomously, like turning

off contact-tracing functionalities (88/495 occurrences). This continued throughout the discourse, with users describing problems the app has caused them. However, the majority of tweets that suggested the app was actively disrupting the lives of the public appeared more towards the end of the sampled period, when the 'pingdemic' occurred (456/845 occurrences). Examples of these included instances when the app was defined as '**wreak** havoc', '**cripple** Britain' and '**threaten** the economy'. This relates to the idea that the system failed to meet the needs and expectations of users (Kent, 2020; Mbwogge, 2021). It also supported the findings by Lamanna and Byrne (2018) and Riegler (2019), according to whom humans could be perceived as 'at odds' with the decisions of the system.

#### 5.5.2.5 *Functioning*

One final category (49.07% of occurrences) was the app being presented as independently undertaking (or attempting to undertake) functional activities that were integral to its running. Tweets in the later months of 2020 stated that the app had a job, a clear *personalisation* (LogDice; 7.02). Tweets in this discourse indicated that one of the intended primary functions of the app was to help or assist users. When this was perceived as not happening, the app was not fulfilling its (supposed) rationale for existing. This is particularly prevalent when the app was said to not be '**helping**', at the start of 2021 (LogDice: 5.22), and perceived as failing to keep users safe. The fact that Twitter authors saw the app as responsible for their safety and welfare showed its prominence and influence as a social actor, much like the findings of the study by Kent (2020) and Mbwogge (2021). In the later parts of the discourse, the app was occasionally presented as having limited legal power or obligation over users (48/283 occurrences), providing an insight into how the app was perceived as responsible towards its users.

During the 'pingdemic' part of the discourse, the app was said to '**send**' (LogDice: 9.38) many notifications, suggesting that the app

was designed for this. Many of these tweets indicated that, although the app was not necessarily instructing users, it encouraged non-compliance, with its too many notifications. This recalled the findings from the follow-up study by Pepper et al. (2022). Additionally, it may also indicate that the app '**pinging**' was perceived as more invasive than simply '**saying**'. Despite users wanting the app to function properly, users appeared to find '**pinging**' overbearing.

Additionally, *personalisation* was present when looking at the app's perceived functionality. For example, the collocate '**think**' (LogDice: 7.24), seen throughout the discourse, regularly presented the app as stating something that is incorrect (30/39 occurrences). This was in relation to the app not working as perceived by the Twitter user. This indicated that the app was perceived as having the capacity to think or act autonomously, leading to the opposition of system usage (Beer, 2017; Grange, 2022; Mahmud et al., 2022).

### 5.5.3 *Trends of Passivisation*

#### 5.5.3.1 *Backgrounding*

Instances of backgrounding were found throughout the discourse. Examples of this included the way in which the developers of the app (Serco, the NHS and the UK government) were foregrounded, especially at the start of the discourse, and how the general public, affected by Covid-19 and isolation requirements, became a focus as time went on. This meant that the app was *backgrounded*, according to principles of SAR (Van Leeuwen, 2008), with its agency obscured (Clark, 1998). The app was still discussed in a negative way in these constructions, despite not being an overt social actor, due to its reduced agency. This presentation intersected with part of the work outlined by Feier, Gogoll, and Uhl (2021), who suggested that decision-making algorithms may deflect blame from more responsible players. Additionally, this portrayal may be a reflection of the perceived atti-

tudes of Twitter users. Some (14/42 occurrences) believed that, while the app played an important role, the responsibility remained with the developers of the app or with the humans that used the app at their own discretion. The removal of agency diluted the app's impact (Comrie, 1977). That being said, the proportion of passive presentations of the app was very small (approximately 4% of all constructions) in comparison to active presentations.

However, considering verb choices, such as '**tell**' and '**cause**', passive constructions were still present. Thus, a small portion of tweets (32/366 occurrences) using passive constructions appeared to imply that the app still has some agential power, possibly to be labelled as agency metaphors (Morris et al., 2007). Such an agential power could be considered as impacting the app, hence the app was still deemed to have some responsibility for processing information.

### 5.5.4 *Limitations and Future Work*

With a corpus of 118,316 tweets, it would have been practically impossible to manually examine each of them (Wetherell and Potter, 1988). Hence, NLP and CL were used to filter the dataset collected and identify relevant potential social actors, through the analysis of the keyword 'app' and its 14,245 collocates in the corpus, which were examined through concordance grids and LogDice. This methodological approach was intended to mitigate this issue of infeasibility.

Another limitation posed by DA was subjective biases, impacting the interpretation of instances of sarcasm and humour, especially those that were less explicit. In this sense, such a challenge is not new to researchers (Gill, 2000; Morgan, 2010). Nonetheless, the combination of DA with computationally-aided techniques was intended to reduce the impact of this difficulty and may benefit future research, too.

Since 'app' was the key-term searched, this work disregarded most instances where *exclusion* masked the app in constructions and the actual word did not feature. Nonetheless, this could constitute an interesting future research focus, which encompassed explicit *excluded* constructions. Additionally, the system may have been discussed in tweets without specific reference to 'app'. Although other social actors replacing 'app' would be hard to find in a large corpus, a good starting point may be synonyms of 'app' in this specific context, such as '(information) system', 'application', 'tool', 'program' or 'software'. Similarly, related field-specific words may also offer relevant research insights, such as 'functionality', 'function(s)', 'operation(s)', 'spread(ing)', 'track(ing)' or 'trace/tracing'.

Due to the brevity of Twitter discourse, which is limited by a 280 character-limit, the app may have been presented actively to facilitate conciseness. For example, 'the app **told** me to isolate' (26 characters) contained 6 fewer characters than 'i was **told** to isolate by the app' (32 characters), which could have been an equally valuable semantic alternative. Such a possibly increased number of active presentations is likely to have affected the times the app was presented clearly as a social actor. Consequently, future work may involve examining other social media or text-sharing platforms that do not limit characters in posts/content to see if the majority of active agential presentations is comparable.

## 5.6 CHAPTER SUMMARY

This chapter began with the NLP-based analysis, where three main topics emerged: Topic 1 focused on contact tracing and isolation, Topic 2 on government involvement and data management, and Topic 3 on app download and availability. In January 2021, discussions on Topic 2 surged, likely due to concerns over government's role during the second lockdown. This coincided with a decrease in Topic 3

discussions, suggesting a shift in focus. However, limited guidance on interpreting topics may have affected accuracy. Overall tweet sentiment ranged from slightly positive to slightly negative, fluctuating over the discourse period but dipping specifically at the start of the second national lockdown in January 2021 and the easing of restrictions in July 2021. Challenges in detecting negation and sarcasm impacted sentiment analysis, while, for emotion detection, 'trust', 'fear' and 'anticipation' fluctuated, though the direction of trust remained unclear, potentially affecting data reliability.

According to the CL-, DA- and SAR-based examination of agency and transitivity in tweets containing the word 'app', published between September 2020 and July 2021, Twitter users presented the NHS Covid-19 App as a social actor and with a clear sense of social agency, addressing the research question. Specifically, the app was predominantly presented actively by Twitter users in 96% of cases, employing various techniques, most notably *personalisation*, but also including *determination*, *agency metaphor* and *genericism*. Indeed, it was found that these active presentations, that implied social agency, primarily conveyed the idea of the app informing (21.47%), instructing (15.33%), providing permission (9.1%), disrupting (5.02%) and functioning – or failing to (49.07%).

The app was also presented passively on occasions (approximately 3%), although this decreased to a minimum of 2% in July 2021. In such instances, the app was often *backgrounded* in order to make the developers or operators of the app more apparent or responsible. On occasions, the focus was on the members of the public impacted by the app (mal)functioning, rather than the app itself. Comparable instances, when the impact of the app as a social actor was limited, saw the app being presented actively but simultaneously ridiculed.

The implications for this study are that Twitter users presented the app as responsible for their own welfare through various active presentations, especially when the app instructed them or provided permission. According to the tweets examined, the perceived respon-

sibility to process information remained with the app throughout the discourse. Such a perception was especially pronounced when significant events prompted further questioning of the app's capabilities (i.e. during the app's launch in September 2020, the second lockdown in January 2021 and the 'pingdemic' phase in July 2021).

In addition to offering insights into the online response to this specific event, this contribution holds the potential for other broader implications in the context of decision-making algorithms. While the disruption caused by the pandemic has waned in the UK, the findings of this study shed light on how the general public might respond to forthcoming decision-making algorithm interventions. This insight is particularly valuable in the context of healthcare or digital contact-tracing initiatives, shining light on barriers to adoption. Therefore, even in a post-pandemic world, the findings of this study remain important.

Overall, this study has provided insights into how social agency communicated via social media public discourse dealing with algorithmic-operated decisions when the AI agency behind those information systems is not openly disclosed. Such a relationship was exemplified by that between the NHS app, grammatical agency and social agency, building existing work on the social agency of decision-making algorithms (Lamanna and Byrne, 2018; Mahmud et al., 2022; Rubel, Castro, and Pham, 2020; Zarsky, 2016). Therefore, this chapter contributes to investigating the social impact of the NHS Covid-19 App. In particular, this was showcased through the combination of CL and DA, underpinned by SAR. Briefly, this research argues that the views expressed on social media indicate that the app was presented as having a perceived high level of responsibility for the welfare and safety of its users, according to tweets which explicitly referred to the app.

# 6

# CHATGPT

## 6.1 STUDY BACKGROUND

As already mentioned, research around the deployment of automated decision-making algorithms acknowledges their ability to reduce errors while raising questions about their increasing autonomy (Araujo et al., 2020; Busch and Henriksen, 2018; Young, Bullock, and Lecy, 2019; Ziewitz, 2016). As these algorithms assume greater decision-making responsibilities, they begin to exhibit a form of social agency and are often attributed human-like characteristics (Crang and Graham, 2007; Panagiotopoulos, Klievink, and Cordella, 2019; Wagner, 2019).

When such algorithms deviate from expected performance, there can be severe consequences, risking biased outcomes, undermining trust and increasing the potential for blame. Trust in decision-making algorithms hinges on perceptions of agency and responsibility, with users favoring autonomously operating algorithms that consistently yield reliable outcomes (Bonnefon, Shariff, and Rahwan, 2016; Floridi et al., 2018), while distrust arises when algorithms seem influenced by external factors like human biases (Bonnefon, Shariff, and Rahwan, 2016). As well as this, blame, shaped by moral judgments and cognitive biases, presents complexity due to the agency attributed to decision-making algorithms (Baumeister, 1996; Coates and Tognazzini, 2013; Ross, 1977). This results in a 'responsibility gap' and challenges in attributing blame for negative outcomes (Mittelstadt et al., 2016; Munch, Mainz, and Bjerring, 2023; Tollon, 2023), particularly when algorithms perpetuate biases or generate negative results amid algorithmic opacity. This indicates that implications for trust

and blame may impact algorithmic adoption (Bonnefon, Shariff, and Rahwan, 2016; Floridi et al., 2018). Therefore, investigating algorithmic agency is crucial and can be particularly urgent if the algorithm itself backgrounds those who are responsible for its development and deployment (Feier, Gogoll, and Uhl, 2021; Olhede and Wolfe, 2020; Peeters, 2020; Velkova and Kaun, 2021), as in the case of the chatbot ChatGPT.

The rapid advancements in Artificial Intelligence (AI) have paved the way for the development of sophisticated chatbot systems, capable of engaging in human-like conversations (George and George, 2023; Ray, 2023). Among these, ChatGPT, launched by OpenAI in November 2022, stands out as an advanced AI chatbot that utilises deep learning models and natural language processing techniques to understand and generate human-readable text in a conversational manner (Hariri, 2024; Rathore, 2023). The utility of ChatGPT extends beyond mere conversation, as it can also assist or entertain (Firat, 2023). With its ability to comprehend and respond to a wide range of queries and prompts, ChatGPT has garnered significant attention and adoption, surpassing 100 million monthly users and demonstrating its capability to successfully pass graduate-level exams (Ali et al., 2022; Ye, 2023). When examining Google trends, searches for ChatGPT outperform other generative AI systems significantly, as seen in Figure 17. This attests to its widespread popularity and significant social impact. This study focuses on a potential 'hype period' (Rogers, 1995) of early popularity of ChatGPT – between its launch in November 2022 and the announcement of GPT-4 in March 2023.

There is an emerging research interest in the social impact of ChatGPT in particular (Abdullah, Madain, and Jararweh, 2022; Verma and Lerman, 2023; Ye, 2023). This includes how panic has been prominent in ChatGPT reactions (Peñalvo, 2023; Roose, 2022; Yatoo and Habib, 2023), as well as other justified concerns regarding ChatGPT that include misinformation (De Angelis et al., 2023; Najafali et al., 2023), ethics (Ray, 2023; Zhou et al., 2023), job displacement (Aljan-

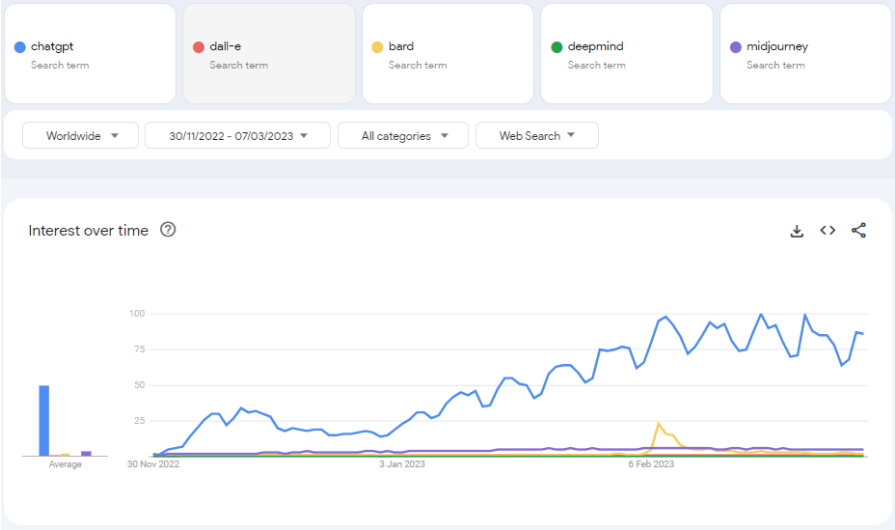


Figure 17: Trajectories of ChatGPT compared with other generative AI systems. Data source: Google Trends (https://www.google.com/trends)

abi, 2023; Biswas, 2023) and unintended consequences (Doshi, Bajaj, and Krumholz, 2023; Ferrara, 2023). A small number of studies have specifically explored ChatGPT's agency. For instance, research into society's perception of ChatGPT's human-like traits (Al Lily et al., 2023; Choudhury and Shamszare, 2023; Gutiérrez, 2023; Sundar and Liao, 2023) viewed the algorithm behind it as an author, interactor and influencer. Other studies found conflicting narratives, with ChatGPT depicted as creative and beneficial, yet also as incompetent and polluting human culture (Bran et al., 2023). Others emphasised the social agency conflicts affecting the adoption of an anthropomorphic perspective and treating AI as a social actor (Shijie, Yuxiang, and Qinghua, 2023).

As mentioned, upon its launch, ChatGPT gained significant public attention, with social media platforms, especially Twitter, serving as a sounding board for user reaction to its deployment (Korkmaz, Aktürk, and Talan, 2023; Taecharungroj, 2023). While efforts have focused on gathering public perspectives on ChatGPT (Abdullah, Madain, and Jararweh, 2022; Verma and Lerman, 2023; Ye, 2023), a noticeable research gap pertains to the examination of views expressed

on Twitter. This social media platform contains vast, diverse opinions on current events (McCormick et al., 2017; Weller et al., 2013), hence it can be a valuable source of data for understanding public reactions. Despite the initial studies investigating the perceived social agency of ChatGPT, a research gap remains as to its presentation on Twitter specifically.

Addressing the aforementioned research gap can offer comprehensive and in-depth insights into the broader public's reactions to ChatGPT. To analyse views expressed relating to items of social interest on Twitter, such as ChatGPT, it is common to use popular NLP-based computational linguistic approaches, with 'off-the-shelf' tools providing a solid approach to studying public discourses on current societal topics (McCormick et al., 2017). A small number of studies have applied NLP-based approaches to analyse Twitter reactions to ChatGPT (Haque et al., 2022; Korkmaz, Aktürk, and Talan, 2023; Leiter et al., 2024; Taecharungroj, 2023). These studies have only analysed general topic and sentiment trends, have only focused on early reactions and have provided little evidence to suggest they have used existing best practices to guide their process. Even though helpful in the identification of themes and ideas, the presentation of ChatGPT's perceived agency is yet to be investigated.

To examine the intricate relationship between grammatical agency and social agency, the NLP-based analysis was integrated with CL and DA, underpinned by SAR (Van Leeuwen, 2008). Grammatical agency, often manifested through transitivity, can help distinguish whether an entity is portrayed as the performer of actions or the passive recipient of them (Leslie, 1993). By scrutinising the discourse surrounding ChatGPT and decision-making algorithms, perceived power dynamics can be illuminated and identify social actors that emerge within these intricate discussions (Clark, 1998; Van Leeuwen, 2008), potentially foregrounding perspectives on trust, blame and barriers to decision-making algorithm adoption.

Following the same structures as Chapters 4 and 5, this chapter begins by explaining the specific approach used for the study, detailed in section 6.2. Subsequently, section 6.3 analyses the findings from automated topics, sentiment and emotion trajectories. After this, section 6.4 explores the discourse through CL and DA. Finally, these findings are discussed alongside relevant literature in section 6.5.

### 6.1.1 *Study Research Question and Objectives*

The sub-research question for this chapter is as follows:

> What insights into agency, trust and blame in the Twitter discourse surrounding ChatGPT can be achieved through combining language analysis approaches?

In turn, the following objectives will be addressed:

3a Demonstrate how Natural Language Processing techniques (sentiment analysis, topic modelling and emotion detection) provide insight into Twitter discourses surrounding ChatGPT.

3b Demonstrate how Corpus Linguistics, particularly collocation, provides insight into public Twitter surrounding the agency of ChatGPT.

3c Demonstrate how Discourse Analysis provides insight into Twitter discourses surrounding the agency, trust and blame of ChatGPT.

3d Identify the strengths and limitations of using the three approaches to investigate Twitter discourses surrounding ChatGPT.

## 6.2 STUDY APPROACH

### 6.2.1 *Data Collection and Processing*

Data extraction was performed using the Tweepy module in the Python programming language (Roesslein, 2009). Tweets containing any of the following terms were collected: 'chatgpt algorithm', 'chat gpt algorithm', 'chatgpt llm', 'chat gpt llm', 'chatgpt 'large language model', 'chat gpt 'large language model', 'chatgpt model', 'chat gpt model', 'chat gpt @openai' and 'chatgpt @openai'. This selection criterion aimed to capture tweets directly relating to how ChatGPT works, as well as the more general capturing of tweets that include OpenAI. Unfortunately, searching for 'ChatGPT' alone yieleded too many results to be analysed in a meaningful way. Although this search term alone may not have captured all aspects of the discourse, it provided a starting point for investigating the expressed views about ChatGPT. This selection yielded 88,058 tweets collected from November 30 2022 (the release of ChatGPT), until 6 March 2023 (the week prior to the launch of GPT-4, in order to capture tweets relating to ChatGPT only and not confuse with the launch of GPT-4). Although the data collected was global, only English tweets were chosen for analysis, focusing on the expressed views in English.

During the data extraction process, each tweet was assigned a unique number to pseudonymise the data. Stopwords were removed from the dataset using gensim and eliminated long and short URLs, as well as the 'RT' (retweet) indication at the beginning of tweets. To ensure anonymity, Twitter handles mentioned within the tweets were redacted using gensim. The tweet IDs and other associated information can be found in the University of Nottingham Research Data Management Repository.

### 6.2.2 *Natural Language Processing Approaches*

#### 6.2.2.1 *Topic Modelling*

To prepare the existing data for analysis, the gensim module's 'simple_preprocess' function was used to tokenise the data. Additionally, bigram and trigram models were created using the 'phrases' function in gensim. The process involved generating meaningful bigrams and lemmatising the text using the Natural Language Toolkit (Cushing and Hastings, 2009). The id2word dictionary was then constructed by combining the input data with the gensim corpora, assigning a unique ID to each word in the document. Based on this dictionary, a corpus was created, representing the mapping of word IDs to their respective frequencies (Rehůřek and Sojka, 2011). Finally, the topics were generated and displayed using the 'gensim.models.ldamodel.LdaModel' function within gensim. Determining the appropriate number of topics for LDA remains a challenge, prompting researchers to recommend considering the researcher's objectives. A smaller number of topics can provide a broad overview, while a larger number allows for a more detailed analysis (Nguyen et al., 2020).

#### 6.2.2.2 *Sentiment Analysis*

For this study, VADER was used. The 'sentiment_analyzer_score' function was utilised, configuring the parameters to classify each tweet as 'positive', 'negative' or 'neutral'. Tweets with a score of 0.05 and above were labelled as 'positive', while those with a score of -0.05 and below were classified as 'negative'. It was ensured that the analysis incorporated contextual information alongside sentiment results to improve interpretation (Agarwal et al., 2015), whilst also presenting sentiment as a trajectory over time and allowing for the capture of sentiment trends and changes (Howard, 2021).

#### 6.2.2.3 *Emotion Detection*

EmoLex was, once again, utilised to analyse emotions in the dataset. The 'top.emotions' command was employed, exporting a CSV table that showcased each tweet's correlation to various emotions such as fear, anger, anticipation, trust, surprise, sadness, disgust and joy. Additionally, a separate column was included to label the dominant emotion in each tweet. Additionally, it was ensured that effort was made to mitigate biases in human review when classifying texts as 'neutral' and to address the imperfect correlation between EmoLex and Linguistic Inquiry and Word Count analytical procedures (Fujioka et al., 2019).

### 6.2.3 *Corpus Linguistics*

The CL software used for this analysis was, again, The Sketch Engine (Kilgarriff et al., 2008), chosen for its practicality and availability to academics. It also provided a series of reference corpora for comparison. The analysis consisted of several stages. Concordance lines with 'ChatGPT' as a potential social actor were examined to initiate collocation analysis. Active and passive constructions were explored. To identify active constructions, the collocation criteria of 'ChatGPT' and a single verb to the right (R1) was established. To detect passive constructions, the search criteria 'by ChatGPT' was employed. However, some passive presentations, exemplified by phrases like 'ChatGPT was x-ed by...' and 'ChatGPT has been x-ed,' were identified by including verbs to the right (R1). As a result, these passive structures were distinguished from active ones and re-classified accordingly.

LogDice was chosen as the statistical measure of collocational strength as it not only measures the statistical significance of a collocation, but it also factors in the size of the subcorpus, making comparisons between subcorpora of different sizes easier. To take advantage of this capacity, the corpus was split into three subcorpora that

reflected the key moments in the evolution of ChatGPT, as illustrated earlier in subsection 2.4.3, in chronological order. This was done to facilitate better comparisons in the timeline and answer the research question. These periods were:

- Period 1: Launch (November to December 2022)
- Period 2: Popularity (January 2023)
- Period 3: Developments (February and March 2023)

The strongest collocates for each period were reported based on LogDice scores, with a minimum threshold of three occurrences to determine significance. The top ten words with the strongest collocations from each time period were analysed using DA, which will be explored methodologically next.

### 6.2.4 *Discourse Analysis*

Subsequently, Discourse Analysis (DA) was employed. By combining DA with the results from the Sketch Engine CL-analysis tool (Kilgarriff et al., 2008), this allowed investigation into the agency and social action conveyed in concordance lines featuring the term 'ChatGPT'. This collaborative approach has shown its effectiveness in similar studies (Abbas and Zahra, 2021; Baker, 2012; Nartey and Mwinlaaru, 2019). Again, DA was underpinned by the Social Actor Representation (SAR) framework to explore the interplay between grammatical and social agency. Specifically, this chapter focuses on '*excluding*' agents, '*backgrounding*' them and the '*personalisation*' or '*impersonalisation*' of actors. These linguistic features serve as markers of human-like perception and contribute to the attribution of responsibility (Tourish and Hargie, 2012). Thus, using SAR is helpful when examining blame and responsibility in discourse. This study focuses solely on constructions where ChatGPT is the grammatical subject, rather than ones where it is the grammatical object, primarily due

to the keyword in context within the concordance grids being 'ChatGPT'.

Through the analysis of Twitter discourse, this chapter uncovers common themes in the presentation of the ChatGPT, thus shedding light on how power dynamics are conveyed within real-life data that pertains to algorithmic decisions, even when the operational mechanisms remain opaque. Additionally, prior research can be drawn upon to identify semantically related thematic groups, contributing to a richer understanding of ChatGPT presentations and evolving perceptions over time (Kitishat, Al Kayed, and Al-Ajalein, 2020; Razis, Anagnostopoulos, and Saloun, 2016).

## 6.3 NLP-BASED TECHNIQUES ANALYSIS

Herein, the results for each of the three methods used to analyse the discourse are presented. This is organised by the three approaches and documents the findings from using the analytical approach.

### 6.3.1 *Topics*

#### 6.3.1.1 *Expectations and Initial Findings*

One of the objectives of employing topic modelling as an approach was to discern the overarching themes pertaining to ChatGPT that were being deliberated in online discussions. Anticipated outcomes involved the generation of topic clusters characterised by a coherent and discernible set of words closely associated with each respective theme, thereby facilitating straightforward labeling of the topics. Additionally, it was expected that this analysis would pinpoint emerging trends and contextualise changes in Twitter conversations related to ChatGPT.

Seven latent topics were discovered through gensim LDA. Each topic contained ten key lexical items. These words are presented in

Table 20: Ranking of the top 10 lexical items associated with each latent topic

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
|---|---|---|---|---|---|---|---|
| 1 | text | write | data | api | time | google | coin |
| 2 | generate | use | trained | use | asked | search | crypto |
| 3 | trained | asked | human | released | people | microsoft | token |
| 4 | artificialintelligence | thread | information | skill | code | bing | invest |
| 5 | developed | content | training | app | use | business | news |
| 6 | human-like | writing | think | text-davinci | know | bard | future |
| 7 | chatbot | tool | algorithm | available | write | chatbot | powers |
| 8 | natural | help | people | light | ask | tech | today |
| 9 | data | research | answer | oracle | good | users | exciting |
| 10 | machinelearning | tools | good | developed | think | engine | nft |

descending order of association with the latent topic in Table 20. The number of topics was decided through manual topic inspection and regeneration, examining the ten key words each time, to ensure minimal lexical item overlap. All topic findings are available in the University of Nottingham Research Data Management Repository.

The assignment of a topic for each tweet was presented as a trajectory. With regard to how the topics presented themselves in the tweets from each month of the research time frame, Figure 18 details the percentage of tweets relating to each topic per month.

The generated topics list can be initially interpreted, shedding light on the underlying themes and discussions associated with ChatGPT in the analysed text corpus.

#### 6.3.1.2 *Topic 1: Human-Like Conversations*

The first topic appeared to revolve around the generation of text using trained artificial intelligence, specifically in the context of developing chatbots with human-like capabilities, emphasising the role of natural language processing, machine learning and data availability. Notably, Topic 1 initiated with a relatively low proportion but gradually increased until the seventh week, reflecting a growing emphasis on AI-driven text generation and chatbot development. Towards the end of the observed period, Topic 7 (which pertained to cryptocurrency and blockchain discussions) demonstrated a significant increase in

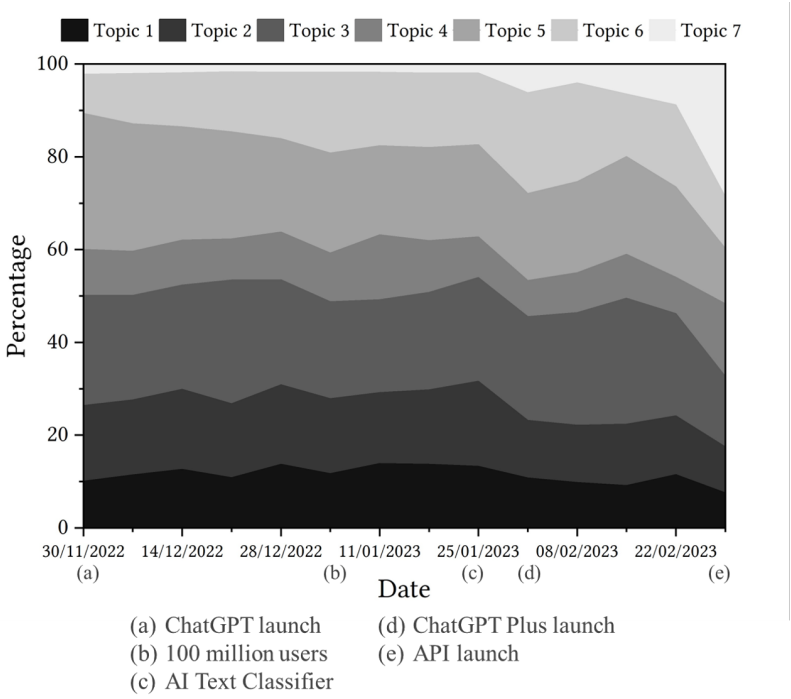


Figure 18: Trajectories of topics detected in tweets relating to ChatGPT.

proportion and Topic 1 reduced consequently. This surge implied an escalating interest in these domains within the context of ChatGPT.

When examining manually, the early weeks in the discourse showed that there were conversations around this topic. For example, in the second week of the discourse collected, many tweets encompassed this topic, with one user acknowledging ChatGPT's 'reassuring conversational ability'. Additionally, another user suggested that ChatGPT could be mistaken for a human due to its vocabulary, syntax and phraseology. This indicated user fascination and satisfaction with ChatGPT's human-like conversational capabilities rather than concerns or fears. In other words, it appeared to suggest that the tool was trusted, at least with regard to its capacity to converse with a user.

#### 6.3.1.3 *Topic 2: Assistance with Writing*

The second topic highlighted the utilisation of ChatGPT as a writing aid, showing how users leveraged its capabilities for guidance, research and collaboration with writing tools. Topic 2 exhibited an intriguing trajectory. It gradually peaked on December 5, signifying increased interest in ChatGPT's potential for writing assistance, followed by a dip on February 15. Nevertheless, its sustained presence underscored ChatGPT's value in the writing community.

When zooming in on the first week in the discourse, one user's request for a short essay about 'the Maldives democracy movement' demonstrated an early focus on writing. Similarly, in the second week, tweets continued this pattern, with one user recognising its potential in assisting with writing tasks. Topic 2 saw a fairly consistent presence as it rose until 25 January, which coincided with the announcement of the AI Text Classifier. These discussions encompassed various writing tasks beyond text, such as homework, coding, legal document writing and code generation for a Flask app. However, like Topic 1, Topic 2 dipped in presence on the week beginning 1 February, which coincided with the launch of ChatGPT Plus, although there is limited overt discussion of this in the sample tweets.

#### 6.3.1.4 *Topic 3: Data and Algorithm Training*

The lexicon associated with the third topic might emphasise the importance of data in training ChatGPT, highlighting the role of human involvement and information acquisition in the algorithm's accuracy assessment. Topic 3 was seen to hold the greatest proportion of tweets in the discourse analysed. The trajectory of Topic 3, shown in Figure 18, fluctuated in its proportion over the observed period. It started with a relatively high proportion of 23.79% and experienced minor variations in subsequent weeks. The topic maintained a consistent presence in the disourse, with proportions ranging from 15.17% to

27.15%, suggesting early discussions on the role of data and algorithm training in ChatGPT's performance improvement.

When examining the sample of tweets, it became evident that Topic 3 served as a background to user discussions, providing supportive information rather than being a focal point. Several tweets provided information about ChatGPT, shedding light on its model version and training process but as supporting information only. For example, one tweet referred to the 'text-davinci-003' model, denoting the specific version of GPT-3 utilised by ChatGPT. Later in the discourse, another tweet mentioned training ChatGPT on a substantial amount of text, although the details regarding the training data remained undisclosed. Furthermore, some tweets in December drew comparisons between ChatGPT and their previous experience of using GPT-3.

#### 6.3.1.5 *Topic 4: API Impact on Content Production*

The fourth topic explored the application programming interface (API) of ChatGPT and its impact on content production, foregrounding the varied capabilities and features accessible through the API, including specific version releases. This topic maintained relatively stable proportions over time, ranging from 7.82% to 15.67%, demonstrating a consistent focus on data, training and algorithm performance.

Based on the sampled tweets, it was evident that ChatGPT's API had had an impact on content production. Initially, users expressed a desire for the API's availability. Over time, discussions evolved to encompass real-world applications such as essay and speech generation. However, as the discourse progressed in January 2023, tweets discussed inconsistencies in ChatGPT's responses, possibly related to API functionality and poor quality content. In February, tweets acknowledged the potential of ChatGPT as a content production tool but did not directly address the API or its impact on content production. However, at the end of the discourse, Topic 4 gained moderate prominence with tweets considering ChatGPT's potential to transform computing, concerns about its misuse and references to

its evolving accuracy in content production. These tweets provided insights into the impact of ChatGPT on content creation and its potential ramifications.

#### 6.3.1.6 *Topic 5: Efficiency*

The fifth topic examined temporal aspects associated with ChatGPT usage to generate the best possible answers using prompts. It encompassed discussions concerning the time users spent posing questions, writing code, seeking assistance and evaluating the chatbot's response efficiency. Looking at its trajectory, Topic 5 consistently maintained a substantial presence, ranging from 11% to 29%, indicating sustained significance in conversations regarding ChatGPT's time efficiency.

This continued prominence in Topic 5 discussions throughout the entire period was linked to users' efforts to optimise ChatGPT's output. This could be explained by many Twitter users discussing how to get the best answers from ChatGPT in order to maximise its output. Upon manual inspection of the human reviewed tweets, early discourse addressed response speed, with some users noting that model responses were fast by default but might lack self-correction capabilities without explicit error identification. Also, at the start of the discourse, several tweets complained about ChatGPT regularly 'crashing' or not being available, hence the need to perhaps maximise efficiency when access was available. Later on in January, some tweets discussed how ChatGPT was less concerned with the accuracy of its answers as it was the appearance of accuracy in its answers. Further tweets provoked how people were perhaps drawn to ChatGPT because it was, in one user's words, 'a good bullshitter', akin to a human trait, rather than despite this. Towards the end of the study, Topic 5 diminished in dominance, aligning with the emergence of a new dominant theme in Topic 7.

#### 6.3.1.7 *Topic 6: Impact on Business*

On a different note, the sixth topic appeared to introduce a comparison between different search engines and tech companies, such as Google, Microsoft and Bing, within the context of chatbot adoption. Topic 6 demonstrated varying proportions throughout the observed period, indicating discussions and comparisons between ChatGPT and other technology companies. The trajectory showed a notable increase in the sixth week, which might have highlighted a growing emphasis on comparing features, capabilities and performance of chatbot offerings in the market. Fluctuations in Topic 6's proportions might have reflected shifts in interest and provided insight into the market dynamics in chatbot development and adoption.

When manually inspecting sampled tweets, Topic 6 had minimal presence at the start of the discourse, with a few tweets mentioning potential effects on Google's revenue model and Microsoft's investment in OpenAI. As the discourse continued, more tweets highlighted real-world implications, business opportunities and the potential challenge to Google. As the topic peaked in late January and early February, the sampled tweets reflected this, with discussions including ChatGPT's ability to challenge Google's dominance in language models, ideas suggesting its use for teams and business logic, using it for investment advice and a pilot subscription plan for monetisation. At the height of its presence in the discourse, tweets expressed disappointment with Google's AI chatbot, Bard, and praise for the development of ChatGPT.

#### 6.3.1.8 *Topic 7: Cryptocurrency*

The seventh topic seemed to diverge from the technical aspects and centred on cryptocurrency and blockchain, covering coins, tokens, investments, news and the future prospects of cryptocurrencies including non-fungible tokens (NFTs). Topic 7 showed an interesting trajectory throughout the observed period but gradually gained trac-

tion, experiencing fluctuations before a sharp peak at the end of the study. This upward trend might reflect an increasing interest and engagement with cryptocurrency and blockchain topics in the ChatGPT discourse, signifying the evolving nature of these discussions and the need to stay informed about their impact and potential applications.

The significant increase in Topic 7 towards the end of the period was of interest. For instance, sampled tweets hinted at advertising livestreams, events promoting cryptocurrency, trading strategies and general discussions about using ChatGPT for insights. Although there was little in terms of how this might have been influenced by the wider discourse, this might have been impacted by Twitter and Tesla owner Elon Musk's resignation from the OpenAI board and his interest in setting up a rival company given his association with cryptocurrency trading.

#### 6.3.1.9 *Human Review and Critical Reflection*

In addition, two blind human reviews were completed. A stratified sample of 10 tweets per week (140 total) were selected and categorised according to the pre-defined topics that were generated. The reviews found a 24% match between the human reviews and the automated topic labelling. Inter-annotator agreement (measured by Cohen's Kappa) was 0.636, indicating substantial agreement (according to Viera and Garrett, 2005). In this, common errors included labelling of Topic 2 when the automated labelling suggested it would be Topic 4 (and vice-versa).

After the analysis, the critical reflection raised the following points:

SUNSHINE Once again, LDA effectively identified co-occurring terms and latent topics in both datasets, utilising the user-friendly gensim tool. Moreover, integrating this approach with the contextual analysis yielded insights for future exploration.

RAIN The lack of clear guidelines for interpreting the output of gensim's LDA topic modelling. The absence of clear guidelines for interpreting the output was challenging, making topic identification and comparison with other studies more difficult. Additionally, discrepancies between automated and human labeling raised concerns.

LIGHTNING An interesting reflection from using LDA was the consistent presence of certain words across different topics, underscoring the importance of context in determining the word's meaning and implications, which can vary based on the associated topic.

FOG One challenge in using gensim's LDA is the interpretation of results, particularly in translating automated, frequency-based outcomes into meaningful human understanding.

### 6.3.2 *Sentiment*

#### 6.3.2.1 *Expectations and Initial Findings*

The primary objective of employing sentiment analysis in this study was to obtain a comprehensive understanding of the discourse and its alignment with contextual factors. It was expected that the analysis would identify the overall sentiment (positive, negative or neutral) within the discourse, shedding light on the emotional tone and attitude of the participants, thus facilitating a deeper examination of the interplay between sentiment and contextual factors.

From the VADER sentiment analysis, Figure 19 shows that overall sentiment was 0.21 to 0.31, indicating that overall sentiment was positive. All sentiment findings are available in the University of Nottingham Research Data Management Repository.

From the initial data points on November 30 2022 to January 25 2023, the sentiment scores hovered around the mid-range, fluctuating within a narrow range of approximately 0.275 to 0.306. This suggested

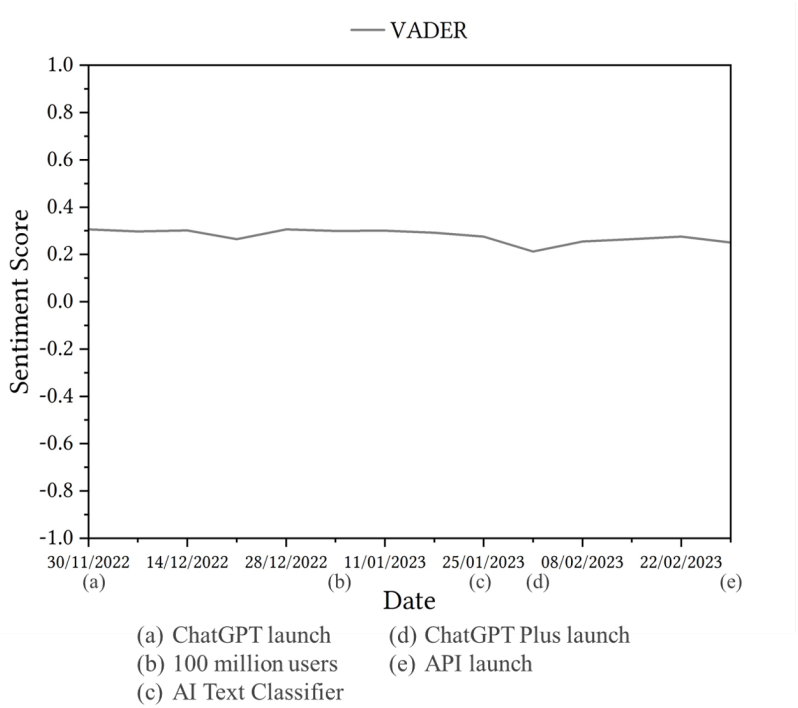


Figure 19: Evolution of the sentiment of tweets relating to ChatGPT using VADER from November 2022 to March 2023.

consistent sentiment in tweets about ChatGPT during this timeframe. However, there was a noticeable decline in sentiment observed on February 1 2023, with a sentiment score of 0.212. This drop indicated a more negative sentiment in the tweets surrounding ChatGPT during that time. Following this decline, the sentiment scores gradually increased, reaching 0.265 on February 15 2023 and further rising to 0.275 on February 22 2023. These incremental increases in sentiment indicated a more positive outlook towards ChatGPT in the latter part of the analysed period.

#### 6.3.2.2 *Contextualising Sentiment Trends*

Comparing sentiment detected in tweets relating to ChatGPT to the wider context of ChatGPT followed. Initially, peak sentiment scores occurred at the discourse's beginning, with manually reviewed tweets expressing excitement and appreciation for ChatGPT's capabilities. They perceived ChatGPT as an 'amazing and revolutionary tool',

praising its utility across diverse domains, including studies, work and development. Furthermore, users emphasised its potential for creative applications such as generating lyrics, stories and essays. The tweets conveyed a collective sense of enthusiasm for the technological advancements embodied by ChatGPT, with users eagerly anticipating a future replete with new possibilities.

Notably, the sentiment trajectory revealed a decline in sentiment starting on January 25 2023, with a sentiment score of 0.27, indicating a decrease in ChatGPT's favourability. This was followed by an even more significant drop in sentiment score on February 1 2023. With a sentiment score of 0.21, this was the lowest recorded weekly sentiment score in the discourse. This coincided, and therefore may have been affected by, the launch of ChatGPT Plus. Upon manual inspection of the tweets sampled in the human review, users expressed frustration with the algorithm's ability to provide 'inaccurate answers' based on limited understanding of source material, criticised biased behaviour and raised concerns about its biases.

There was also a small drop in weekly sentiment scores on 21 December, potentially linked to multiple website outages, as mentioned in the sample tweets, impacting ChatGPT accessibility. Upon manual inspection, the negative sentiment expressed in these tweets towards ChatGPT included criticisms of its value, functionality and trustworthiness. One tweet described it as a 'fucking mess' and 'utterly worthless', suggesting that it promoted an approved narrative and acted as a 'propaganda machine'. Other criticisms centered on knowledge origin traceability, dissatisfaction with performance and ChatGPT's limitations in specific scenarios, like academic assignment writing.

Despite a rise in weekly sentiment after this week, the weekly sentiment scores were not as high as previous. Upon inspection, there was appreciation for the AI's language modelling capabilities, highlighting how it excels at generating text and explaining concepts effectively. Additionally, the incorporation of ChatGPT into educational settings, such as one example showcasing how it works in the curricu-

lum of the London Business School, was seen as a positive development. Users also expressed their initial skepticism reducing, including in examples such as legal questions and company descriptions. However, negative sentiments encompassed doubts about its abilities, privacy concerns, criticism of OpenAI and sarcastic remarks about always 'thanking ChatGPT' so it may 'spare you' from potential enslavement in the future.

#### 6.3.2.3 *Human Review and Critical Reflection*

Once again, for this human review, 10 tweets per month (140 total) were sampled in a stratified and classified by two reviewers according to whether they were positive, negative or neutral. The human review score matched the computer assigned sentiment category on 50% of occasions. The inter-annotator agreement was 0.776, indicating substantial agreement.

For the critical reflection, the following was observed:

SUNSHINE Sentiment analysis efficiently processed the large dataset, with VADER integration proving more reliable than TextBlob in previous studies according to the human review. The sentiment scores provided a quick, time-based overview, facilitating the identification of crucial investigation points.

RAIN The interpretation of individual sentiment scores alone is difficult and lacks meaningful insight. Focusing on individual scores instead of the overall trend can obscure the tool's limitations in capturing nuanced language aspects, resulting in limited understanding.

LIGHTNING Surprisingly, the sentiment analysis exhibited minimal fluctuations despite the dynamic nature and diverse opinions in public discussions. The relatively stable sentiment patterns suggest a certain level of consistency or consensus in the overall sentiment expressed.

FOG A challenge of interpreting sentiment analysis data was the lack of guidance on the meaning of sentiment scores and their implications for understanding the context of the discourse.

### 6.3.3 *Emotions*

#### 6.3.3.1 *Expectations and Initial Findings*

The rationale behind employing emotion detection was to gain insight into the prevailing sentiments towards ChatGPT and identify any prevailing or shared emotional states, expecting to reveal dominant emotions across various discourse phases. The findings aimed to illuminate emotional patterns and provide insights into ChatGPT's emotional landscape at specific time intervals. The data was presented in the trajectory displayed in Figure 20. All emotion findings are available in the University of Nottingham Research Data Management Repository.

#### 6.3.3.2 *Trust*

Firstly, the emotion of trust demonstrated a fluctuating pattern throughout the examined period, with proportions ranging from 46.92% to 55.34%. Particularly, the highest proportion of trust was observed on 18 January and 1 March. The trajectory of 'trust' appeared to maintain a steady presence in the discourse until 1 February 2023, when it saw a sharp decline in presence from 54.49% to 41.18%. This coincided with the release of ChatGPT Plus, accompanied by a sharp decline in sentiment. Notably, tweets sampled on this date, while not explicitly mentioning trust, expressed opinions and experiences related to ChatGPT's performance and reliability. Some tweets expressed skepticism towards ChatGPT, questioning its capabilities and potential disruptions, saying it was 'always unavailable', which might have implied a lack of trust. Other tweets highlighted concerns about biases, racism or the spread of disinformation through Chat-

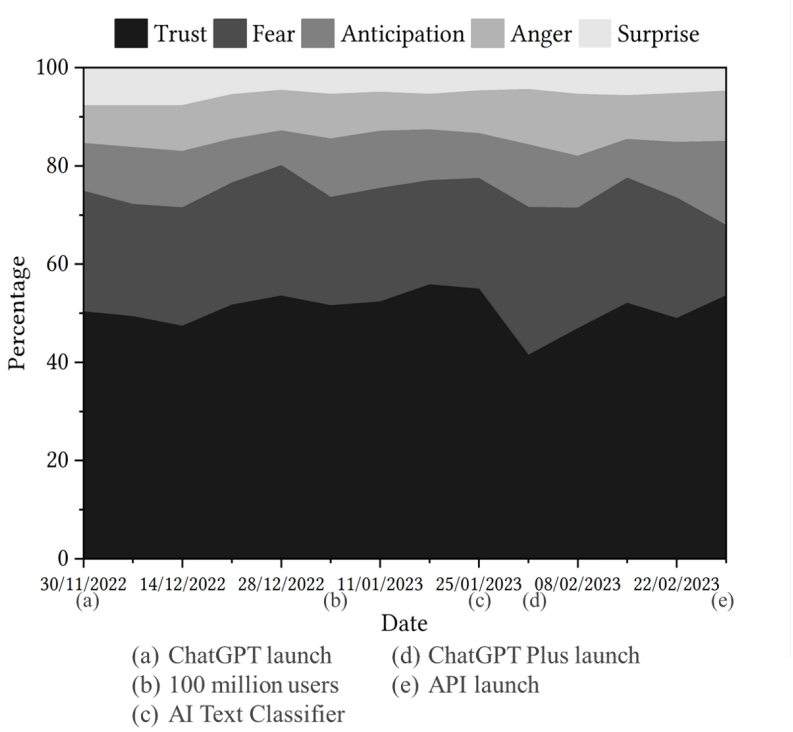


Figure 20: Emotions detected in tweets relating to 'ChatGPT'.

GPT, again potentially presenting a lack of trust in its use. Conversely, other tweets indicated trust in ChatGPT's potential for scientific or practical applications.

However, upon closer examination, it became evident that the emotion of 'trust' consistently emerged in tweets discussing ChatGPT, indicating its prominence within the discourse. Given the distinction between the emotions of 'trust' and 'fear', it was inferred that tweets associated with 'trust' reflected a belief in ChatGPT's reliability, rather than distrust. The classification of tweets containing the emotion of 'trust' presented a discrepancy in the categorisation. This discrepancy arose due to the presence of opposition to trust within these tweets, which would have led to different categorisation. Notably, some tweets included the words 'trust' and 'trustworthy' with negations, such as 'not' or the contracted modal verb 'shouldn't'. It was possible that the EmoLex module did not detect these negations,

possibly due to the prominence of the word 'trust' in the classifier's decision-making process, similar to Chapters 4 and 5.

#### 6.3.3.3 *Fear*

In contrast, the emotion of fear displayed relative stability over time, with proportions ranging from 21.07% to 30.00%. Despite an almost 8% increase in fear detection on the week beginning 1st February, fear did not exhibit any other significant change trends throughout the discourse. With the decline in 'trust' in the week beginning 1 February also came an increase in 'fear', rising from 22.40% to a peak of 30.00%. One tweet saw the author discuss 'malicious actors' and their potential use of ChatGPT to spread fake information on a large scale. The use of terms like 'malicious', 'fake info' and 'disinformation campaign' indicated a concern regarding the potential misuse of ChatGPT, suggesting the presence of fear. At the end of the discourse, 'fear' dropped from 24.47% to 14.23%, coinciding with the launch of the Open AI API.

Upon manual inspection, there seemed to be very few instances of explicit or genuine 'fear' found in the discourse. What was found was one user humorously mentioning closing a 'literal portal to Hell' opened by ChatGPT and others suggested ChatGPT will 'take over' the world. EmoLex may have interpreted this as indicating a sense of unease or apprehension as it classified this without context. Despite this, there were tweets that indicated a level of concern that could be interpreted as fear. For example, in February, one tweet stated that OpenAI were aware of ChatGPT's potential to be used in a way to 'spread fake info on an unprecedented scale'. Others appeared to have unfounded concerns, with users expressing that 'AI is going to ruin everything' and they are 'ready for a racist AI cyborg fuck doll that hates humans'.

#### 6.3.3.4 *Anticipation*

The trajectory of anticipation showed variations, with proportions ranging from 7.05% to 12.65%. Notably, anticipation demonstrated a relatively higher proportion on February 1, perhaps suggesting an elevated level of excitement and expectation. In the same vein as 'fear', 'anticipation' also increased in the final week of the discourse, from 11.26% to 17.08%, again coinciding with the launch of the API. When looking at tweets, users expressed excitement and anticipation for the release of new APIs for ChatGPT and their potential impact, with one user comparing this to the emergence of cloud computing. As the cryptocurrency discourse began to dominate at the end of the time period, more users tweeted in anticipation of the right time to buy or trade.

#### 6.3.3.5 *Anger*

The emotion of 'anger' maintained a relatively consistent proportion, ranging from 7.14% to 12.50%. There were very few spikes or dips in anger. When manually inspecting tweets, very few seemed to express legitimate anger towards ChatGPT; instead, some frustration was observed, especially when ChatGPT had periods of outage in January and users stated that it had 'been hours that they were unable to use ChatGPT. Some expressed that it was 'dead' as '"Get Notified" doesn't seem to ever work', culminating in one user in February stating that it is 'just another fucked up large language model'.

#### 6.3.3.6 *Surprise*

The emotion of 'surprise' exhibited a generally decreasing trend, with proportions ranging from 4.59% to 7.72%. This decline may suggest a diminishing sense of unexpected or surprising experiences associated with ChatGPT as the discourse progressed. Manual inspection of the sampled tweets seemed to confirm this idea, with many tweets at the start of the discourse indicating surprise at the capabilities of Chat-

GPT, with one user stating that they had experienced 'many DAMN , WTF, I CAN'T BELIEVE THIS moments'. However, this surprise dwindled as the discourse progresses and the capabilities of ChatGPT become more well-known.

#### 6.3.3.7 *Other Emotions*

There were several other emotions found in the discourse that held a less significant presence. Emotions such as 'sadness', 'disgust' and 'joy' consistently showed relatively low proportions with minimal fluctuations. 'Sadness' and 'disgust' remained consistently low, while 'joy' was negligible in most instances. The manual inspection of tweets saw this replicated.

#### 6.3.3.8 *Human Review and Critical Reflection*

For consistency, ten tweets per month (140 total) were randomly sampled to be reviewed. The categories to be assigned were 'trust', 'fear', 'anticipation', 'anger', 'surprise', 'sadness', 'disgust', 'joy' and 'no emotion'. Reviewers matched the EmoLex assigned category on 29% of occasions. The inter-rater reliability was 0.786, indicating substantial agreement. Within this, between the reviewers, classifying tweets that the algorithm deemed as 'anger' caused the most disagreement, with the reviewers not matching on 5/11 occasions. Reviewers categorised these tweets as 'fear' or 'disgust' instead.

Finally, the following reflections took place:

SUNSHINE The efficient, rapid detection of tweets in a large dataset was a notable advantage, allowing for timely processing. Furthermore, the ability to classify each tweet into various emotional states further enhanced the comprehensiveness and usefulness of the analysis.

RAIN The accuracy of the EmoLex emotion detection module may have been compromised during deployment, similar to sentiment analysis, with the lack of contextual information hindering the analytical process and potentially rendering the identified emotions arbitrary.

LIGHTNING The presence of 'positive' and 'negative' emotions within the initial set in EmoLex was unexpected, potentially resulting in the omission of important information. These were re-classified upon the removal of these states.

FOG Clarity regarding the categorisation of emotions, particularly trust-related tweets, could have improved the accuracy of the analysis. The inclusion of tweets opposing trust, categorised differently by humans, highlights the need for clearer guidelines for a more accurate reflection.

## 6.4 CORPUS LINGUISTICS AND DISCOURSE ANALYSIS

### 6.4.1 *Timeline Overview of Results*

In the results, it was observed that the collocates '**be**' and '**have**' frequently co-occurred with 'ChatGPT' across all three time periods. Upon manual examination of tweets containing these combinations, the majority were identified as auxiliary verbs. In such instances, these were treated as multi-word expressions and analysed based on their collective meaning, as they conveyed connections between agency and responsibility.

First, the frequency of active and passive verbal constructions including 'ChatGPT' were looked at to ascertain whether it was being presented as a social actor. This overview is shown in Table 21, where it features actively in 96% of the clauses. However, active and passive

Table 21: Frequency of active and passive constructions of *ChatGPT*.

| Time period | Active | Passive | Total |
|---|---|---|---|
| Period 1: Launch (Nov and Dec 2022) | 5,160 | 212 | 5,372 |
| Period 2: Popularity (Jan 2023) | 4,346 | 135 | 4,881 |
| Period 3: Developments (Feb and Mar 2023) | 5,609 | 163 | 5,772 |

constructions alone do not necessarily provide a full account of how ChatGPT is presented in the discourse. For example, ChatGPT could be the subject of an active construction, yet could carry limited social agency. To avoid misinterpretations, CL and DA were combined.

Each of these three time periods – launch, popularity and developments – will be examined individually in this results section. A more comprehensive comparison between the periods occurs in section 6.5. A sample of the concordances, examined in conjunction with the collocational findings, is available in the University of Nottingham Research Data Management Repository.

### 6.4.2 *Time Period 1: Launch (November to December 2022)*

#### 6.4.2.1 *Active Constructions*

In November and December 2022, there were 5,160/5,372 instances of active constructions involving the app. The strongest 10 collocates are shown in Table 22.

'**Have**' (LogDice: 8.80298) was one of the strongest collocates. In tweets early in the discourse, ChatGPT was portrayed with a more active role, suggesting agency and engagement. For example, in one tweet, the author stated that 'ChatGPT has raised the alarm among educators', indicating that it was actively involved in creating uncertainty and generating potential negative impact within the educational sphere. This agency metaphor suggested a more personalised and specific characterisation of ChatGPT, making it appear more like a social actor. However, in many instances, ChatGPT was presented

Table 22: Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in November and December 2022.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | be | 3114 | 103913 | 9.86593 |
| 2 | have | 408 | 24494 | 8.80298 |
| 3 | seem | 76 | 1635 | 8.45962 |
| 4 | do | 203 | 15667 | 8.29997 |
| 5 | write | 90 | 5978 | 8.01296 |
| 6 | explain | 48 | 1181 | 7.89235 |
| 7 | give | 51 | 3801 | 7.49875 |
| 8 | make | 54 | 6847 | 7.17016 |
| 9 | know | 42 | 4213 | 7.15571 |
| 10 | generate | 44 | 5279 | 7.07168 |

as a tool, making use of verbal structures like 'has been released', 'has been getting', 'has been fine-tuned' and 'has been trained'. These constructions positioned ChatGPT as the recipient of actions, rather than the one taking the actions. Therefore, these portrayals emphasised ChatGPT as an object or tool created and manipulated by external actors that were *excluded*. For instance, one user tweeted that 'ChatGPT has been trained on a vast amount of text data', which underscored ChatGPT's passive role in training and highlighted the human agency behind its development. Therefore, this use of the passive voice with 'has been' implied that ChatGPT was a tool or a product created and controlled by OpenAI, or perhaps other excluded entities, rather than an independent agent with decision-making capabilities.

The strong collocate '**seem**' (LogDice: 8.45962) might have indicated uncertainty and speculation about ChatGPT's capabilities, reflecting ongoing evaluation by users on Twitter. These tweets were subjective, based on individual experiences. It may have also suggested that these were surface-level views, which, presently, did not have a strong evidence base to support them. The use of the verb '**seem**' in these tweets indicated that ChatGPT was perceived as more than just a neutral language model. For example, some tweets dis-

cussed ChatGPT's capabilities, including its potential uses in various fields such as education ('ChatGPT seems like it can be a good education tool'), content generation ('ChatGPT seems pretty good at completing short answer questions'), code writing ('ChatGPT seems pretty good at writing, debugging, explaining and translating code as well') and its strengths and weaknesses compared to previous versions of GPT ('ChatGPT seems to have a much smaller training set than GPT3'). The verb '**seem**' was usually used to express uncertainty or tentative observations about ChatGPT's abilities. It implied that ChatGPT's status as a capable social actor was not fully confirmed but was based on initial impressions.

Other tweets highlighted ChatGPT's responses to ethical, philosophical or social questions ('ChatGPT seems to pass a question on ethics, but fails when asked to apply it as a moral judgment' and 'ChatGPT seems to have the opinion that the laws of quantum mechanics are more fundamental than those of thermodynamics'), which suggested that users believed ChatGPT had a stance or opinion on these matters. This indicated that it was seen as more than just a neutral language model. Additionally, other tweets focused on users' interactions with ChatGPT, including their experiences, questions and opinions about using the model ('ChatGPT seems to be able to generate .objs for simple stuff' and 'ChatGPT seems to be blowing my mind right now'). These tweets expressed users' perceptions of ChatGPT's behaviour and its responses to their queries, implying that its interactions with users were a key aspect of its social presence.

'**Do**' (LogDice: 8.29997) was another strong collocate, often used to attribute actions or tasks to ChatGPT. This grammatical structure reinforced ChatGPT's role as an active entity capable of carrying out tasks or functions. Many of these instances were examples of personalisation ('ChatGPT did great, though', 'I'm really curious what ChatGPT does with the code you ask it to explain'). Equally, there were many negated constructions that stated how much ChatGPT did not know ('ChatGPT does not seem to have memory', 'ChatGPT doesn't know

either what's in the demo today'). In other instances, users expressed what ChatGPT was capable of ('ChatGPT does surprisingly well on a Named Entity Recognition task'). Moreover, '**do**' was used to seek clarification or understanding about ChatGPT's capabilities or limitations ('What does ChatGPT do?'). This usage highlighted users' attempts to comprehend ChatGPT's role. Occasionally, users compared ChatGPT to other entities or models, indicating differences in their abilities or functions ('ChatGPT does not know about Dall', 'ChatGPT does not know anything about what happened after September 2021'), highlighting limitations in ChatGPT's knowledge but still presenting it as a social actor capable of comprehension.

'**Write**' (LogDice: 8.01296) also strongly collocated with 'ChatGPT'. Tweets containing '**write**' mainly saw ChatGPT depicted as actively writing content, algorithms, poems and essays ('I had @OpenAI's ChatGPT write a backstory for my @moonbirds NFT', 'ChatGPT wrote monocular SLAM algorithm using GT-SAM'). It suggested an agency in producing textual or technical output, portraying ChatGPT as an active contributor to creative and informative endeavours. Moreover, on some occasions, ChatGPT was presented not just as a writer but as an influencer on creative content, highlighting its role in generating scripts, stories and poems ('ChatGPT [wrote] a script where Jesus is a C++ programmer from the 1990s') and foregrounding its creative agency in shaping narratives ('Not only did ChatGPT write us a script, but it also played the role of a DIRECTOR'). ChatGPT was also shown to play an active role in education by generating essays, research statements and learning materials, all of which showcased it to be a social actor through *personalisation*.

'**Explain**' (LogDice: 7.89235) highlighted ChatGPT's role as an active participant in providing information and assistance. This choice of verb suggested that ChatGPT was not merely a passive tool but an entity that actively engaged in discussions and offered explanations. For instance, one tweet stated that 'ChatGPT [explains]' how its prompting system worked. This tweet positioned ChatGPT as a

knowledgeable entity capable of explaining its own characteristics, demonstrating a degree of self-awareness and agency in the interaction through *personalisation*. Similarly, another tweet mentioned that ChatGPT 'explained the worst-case time complexity of the bubble sort algorithm', highlighting its ability to provide detailed explanations on complex topics, further supporting the notion of it being a social actor. Overall, tweets with the collocate '**explain**' depicted ChatGPT as an active participant in conversations, engaging in explanations and demonstrating agency by offering information and insights on various subjects.

Similarly, tweets containing the collocate '**give**' (LogDice: 7.49875) usually referred to ChatGPT actively providing information, advice or responses to user queries, attributing agency. For instance, one tweet mentioned that 'ChatGPT [gives] purpose of life', implying that ChatGPT had the capability to impart profound insights. Along with this, many other tweets underscored ChatGPT's active role in supplying information, recommendations or responses ('ChatGPT gave me answers', 'ChatGPT gave me a list of advice', 'ChatGPT gives me error messages'). As a result, these tweets highlighted ChatGPT's capacity to influence and engage with users through the act of providing, which was usually perceived as a human attribute, thus *personalising* ChatGPT and portraying it as a social actor within the context of these conversations.

The collocate '**make**' (LogDice: 7.17016) highlighted ChatGPT's role in creating content or influencing outcomes. For instance, one tweet author noted that 'ChatGPT made a nice transcript', emphasising its active content creation. Additionally, another tweet author questioned if 'ChatGPT made a mistake' further highlighting ChatGPT's capacity to not only produce results but also impact situations negatively, thus an example of *personalisation*. The use of 'make' portrayed ChatGPT as a social actor involved in content generation and decision-making across various contexts ('ChatGPT made me a content calendar', 'ChatGPT made the fry meme'), emphasising its role in cultural

content creation. Thus, these presentations depicted ChatGPT as a social actor capable of making, rather than 'generating', content.

The frequent occurrence of '**know**' (LogDice: 7.15571) as a strong collocate of ChatGPT in these tweets highlighted that it possessed knowledge, but it was more of an informational tool rather than an active social agent. For instance, one tweet author questioned whether 'ChatGPT knew C++' and, while this demonstrated ChatGPT's capacity to provide information, this did not necessarily imply it was an independent actor in a social context. The use of 'know' was consistently about the model's ability to provide answers or information, rather than using or appropriating the knowledge. While these tweets suggested knowledge and capability, it did not inherently portray ChatGPT as a social actor with agency. Instead, it positioned ChatGPT as a knowledge resource, a tool that could answer questions and provide information. In many instances, ChatGPT was presented without overt agency, responding to queries and providing information based on pre-existing knowledge, as opposed to actively participating in social interactions or demonstrating agency. Therefore, these tweets did not strongly showcase ChatGPT as a social actor with significant agency; rather, they highlighted its role as a tool for information retrieval.

'**Generate**' (LogDice: 7.07168) was also a strong collocate. Despite being semantically similar to '**write**' and '**make**' and still being used actively grammatically, the pragmatics of '**generate**' implied that ChatGPT's function was to create content ('ChatGPT generates code that uses the openAI go client', 'Don't discount the @OpenAI ChatGPT generated content that results from better queries'), rather than it having any agential authority over the content it created. Thus, its prowess as a social actor was limited in these tweets, as Twitter users were discussing ChatGPT as functional rather than *personalised*.

Through examining tweets containing the strongest collocate, '**be**' (LogDice: 9.86593), there were many constructions that suggested a more passive and impersonalised characterisation of ChatGPT. In

Table 23: Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in November and December 2022.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | write | 81 | 5978 | 8.73646 |
| 2 | inspire | 5 | 191 | 8.55374 |
| 3 | hack | 4 | 171 | 8.29956 |
| 4 | generate | 49 | 5279 | 8.18321 |
| 5 | produce | 9 | 820 | 8.11329 |
| 6 | amaze | 3 | 196 | 7.80033 |
| 7 | edit | 3 | 240 | 7.66312 |
| 8 | suggest | 3 | 410 | 7.22961 |
| 9 | impress | 3 | 463 | 7.11736 |
| 10 | create | 16 | 4045 | 6.93324 |

these tweets, ChatGPT was described as being 'trained' and 'fine-tuned'. This grammatical structure often framed ChatGPT as a tool or product, rather than an active social actor with agency. For instance, the statement 'ChatGPT was trained using Reinforcement Learning from Human Feedback' positioned ChatGPT as an outcome of a training process and lacked an active agency in the training, emphasising that ChatGPT was a result of a technical process rather than an independent social actor with responsibility.

#### 6.4.2.2 *Passive Constructions*

ChatGPT was also the subject in passive constructions on 212/5372 occasions in the first time period, with collocates shown in Table 23. The strongest collocate was '**write**' (LogDice: 8.73646). These passive constructions involving '**write**' all attributed some degree of agency for the content creation to ChatGPT and portrayed it as the author ('Disclaimer: This tweet was written by ChatGPT', 'A thread about ChatGPT written by ChatGPT??'). This was very similar to '**generate**' (LogDice: 8.18321), despite the slight semantic difference ('The above response was generated by ChatGPT', 'The entire content of the book was generated by ChatGPT'). However, this attribution of authorship

to ChatGPT raised questions about whether it was considered a social actor or a legitimate agent with autonomous decision-making capabilities ('I have three job posts published, fully written by chatGPT', 'The canvas generated by ChatGPT'). While the tweets emphasised that the content was authored by ChatGPT, the passive voice implied that ChatGPT lacked personal initiative or intention in generating these tweets, *backgrounding* ChatGPT, in effect. This downplayed the sense of agency and intent that was typically associated with human authors as it simultaneously elevated ChatGPT to the status of a content creator while undermining its role as a conscious social actor. Despite its passive presentation, ChatGPT was still discussed positively in these tweets.

In other tweets, verb choices like '**inspire**' (LogDice: 8.55374), '**amaze**' (LogDice: 7.80033) and '**impress**' (LogDice: 7.11736) were often associated with humans or entities that possessed intention and the capacity to influence or generate reactions. The grammatical structures featuring passive constructions tended to downplay ChatGPT's agency in favour of emphasising the human response or intention, thus *backgrounding* ChatGPT. For example, one tweet author that stated that they were 'inspired by ChatGPT to create a language model chatbot' placed ChatGPT in a passive role, which was also enhanced by the fact that it was presented as serving as an inspiration, not as an active creator. Furthermore, the phrase 'amazed by ChatGPT' suggested that ChatGPT was the source of something astonishing, but it did not attribute this amazement directly to ChatGPT itself. This construction separated the source of amazement from the entity causing it.

Table 24: Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in January 2023.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | be | 2809 | 103913 | 9.72852 |
| 2 | have | 314 | 24494 | 8.46688 |
| 3 | do | 153 | 15667 | 7.95154 |
| 4 | say | 63 | 3771 | 7.94851 |
| 5 | pass | 31 | 514 | 7.63791 |
| 6 | come | 36 | 2522 | 7.3747 |
| 7 | write | 51 | 5978 | 7.30556 |
| 8 | get | 46 | 6290 | 7.1147 |
| 9 | make | 44 | 6847 | 6.97852 |
| 10 | seem | 22 | 1635 | 6.8565 |

### 6.4.3 *Time Period 2: Popularity (January 2023)*

#### 6.4.3.1 *Active Constructions*

In January 2023, there were 4,346 instances of active constructions involving the app. The strongest 10 collocates are shown in Table 24.

'**Has**' (LogDice: 8.46688) showed a slight decrease in collocational strength but remained one of the top collocates in January 2023. Once again, there were many occasions where ChatGPT was the object of constructions as they were presented in the passive voice ('ChatGPT has been corrupted', 'ChatGPT has been trained on data till 2021', 'ChatGPT has been overhyped'). Within these examples, the subject of the construction was *excluded*, meaning that ChatGPT became the focus of the tweet despite having limited social agency. This demonstrated the further occurrence of ChatGPT disguising the human agents behind its creation and development, which might see it blamed or not trusted in the future.

'**Do**' (LogDice: 7.95154) was, again, one of the strongest collocates, although slightly weaker in collocational strength in this time period comparatively. Similar to the first time period, negated constructions using 'do' to express ChatGPT's limitations dominated the dis-

course ('Even chat ChatGPT does not know why model Y is excluded from IRS', 'ChatGPT doesn't reason, so it confidently makes self-contradictory assertions'), with users continuing to highlight what ChatGPT did not know or understand. Users still acknowledged ChatGPT as an active entity capable of carrying out, or not carrying out, tasks or functions, *personalising* it. However, in January 2023, tweets indicated a more nuanced understanding of ChatGPT's capabilities in terms of performing specific actions or tasks, showing that users had perhaps become more specific in their queries or expectations. This might reflect an evolving understanding of what ChatGPT could do, which might impact how responsible or accountable it is.

Among the strong collocates was '**say**' (LogDice: 7.94851), where tweets predominantly framed ChatGPT as a source of information and opinion. This was evident when users wrote 'I use it sometimes to cross verify what ChatGPT says' and 'ChatGPT says No'. These expressions underscored its presence in facilitating conversations and potentially influencing individual decision-making. Not only this, but tweets showed ChatGPT's potential to engage in dialogue as a mechanism for verifying or challenging information ('ChatGPT says otherwise' and 'Here's what ChatGPT said'). However, a recurring theme across these tweets was the notion of ChatGPT's limited agency and capacity, as illuminated by statements like a recurrent acknowledgement of ChatGPT's constrained autonomy ('ChatGPT says it's deleted - but with it in research mode' and 'ChatGPT says it is sorry and explains it is a language model') which underlined its primary function as a language model rather, which might limit its impact as a social actor.

Examining the collocate '**pass**' (LogDice: 7.63791) revealed various dimensions of ChatGPT's identity and functionality. Some tweets presented ChatGPT as a wellspring of knowledge and expertise, attributing it the capability to pass exams and dispense specialised information ('ChatGPT passed the US Medical Licensing Exam', 'ChatGPT passed its MBA final exam', 'ChatGPT passed Bar Test'). These all in-

dicated that ChatGPT possessed knowledge that was usually unique to humans only. However, there were instances where ChatGPT was depicted more as a tool or technology, lacking human-like agency. Some tweets humourously exaggerated its abilities, while others critically challenged its limitations. In this diverse discourse, the portrayal of ChatGPT ranged from an authoritative source of knowledge to a neutral instrument, reflecting multifaceted perceptions of its agency and social role.

'**Make**' (LogDice: 6.97852) signified ChatGPT's capacity to generate content. ChatGPT was portrayed as a creative force, with tweets acknowledging its capacity for content generation, facilitating easy solutions. Simultaneously, '**make**' was also used in the context of discussing ChatGPT's power and influence, as it sparked both curiosity and engagement ('ChatGPT makes students curious'). However, many of these tweets also focused on how ChatGPT generated false content ('ChatGPT makes up fictitious titles of books', 'ChatGPT making up fake caselaw... yikes!', 'ChatGPT making it too easy to generate some mock data'). While this might have been undesirable to the tweet authors, this still portrayed ChatGPT as a social actor with the capabilities to create something, whether that be true or false. Therefore, when compared to the previous time period, a notable difference was the emphasis on ChatGPT generating false or fictitious content, depicting it as a social actor capable of producing both true and false information.

The collocate '**come**' (LogDice: 7.3747) might have implied an action associated with an entity that possessed the capacity to act and influence outcomes. However, the structure of the sentences frequently cast ChatGPT in a passive or reactive role rather than a proactive one. For instance, the excerpt 'CHATGPT came through for the rest' suggested that ChatGPT acted responsively or supportively. This created a sense of ChatGPT possessing some degree of control of its own emergence, hence *personalisation* with strong agency metaphor. While the use of '**come**' attributed a degree of agency to ChatGPT

as the entity performing the action, the overall structure of the sentences and the specific contexts in which it was used tended to depict ChatGPT more as a responsive presence, rather than a proactive social actor. It portrayed ChatGPT as a tool or resource that was called upon or utilised by humans, framing its agency within the boundaries set by the users.

Once again, '**write**' (LogDice: 7.30556) was a strong collocate in January 2023. Grammatical structures indicated that ChatGPT was actively engaged in writing content and producing outputs. For instance, one tweet stated, '@OpenAI's ChatGPT wrote me a 2,000-word essay on global warming', clearly attributing the act of writing to ChatGPT. Additionally, what users claimed ChatGPT was writing appeared to be more sophisticated than the tweets from November and December 2022, with recurrent patterns of attributing writing actions to ChatGPT in various contexts, including coding, content generation, poetry and more. This included instances where ChatGPT was held directly responsible for written work, exemplified by tweets such as 'did ChatGPT write your abstract or not?'.

There were also many different contexts in which the collocate '**get**' (LogDice: 7.1147) was used. Tweets suggested that ChatGPT was capable of influencing or taking action. For instance, the tweet that discussed how 'ChatGPT got me to try it', this implied ChatGPT's influence on the user, positioning it as a persuasive entity and, thus, portraying it as a social actor capable of prompting user actions. Similarly, 'ChatGPT gets out of control' in another tweet implied that ChatGPT could display certain behaviours or tendencies. It was also used as a synonym for understand ('Even ChatGPT gets it'), alongside potentially negative impressions of understanding ('ChatGPT got that wrong then'), further signifying its role as a social actor. This was similar to the varied usages of the collocate '**make**' (LogDice: 6.97852), where, in some instances, ChatGPT was attributed with the ability to 'make' or influence decisions ('ChatGPT makes not money') and engage in humour ('ChatGPT making fun of its own downtime is the

new meta'). Tweets also discussed ChatGPT's errors, saying that it 'makes up references' and 'makes fictitious titles of books that don't exist'.

Despite a lower collocational strength than the previous time period, '**seem**' (LogDice: 6.8565) also appeared as one of the top collocates for January 2023. Once again, tweets indicated that surface-level impressions of ChatGPT were being reported here. For example, it was stated in one tweet that 'ChatGPT seems to be good at answering questions but not at asking them', implying that these attributes were not definitively established. Similarly, 'ChatGPT seems to fail in Comprehension test' signified that ChatGPT's performance was subject to interpretation and not portrayed as a concrete action. The impression created by '**seem**' was that users were reporting tentative impressions of ChatGPT, which limited its prowess as a social actor. However, this was not at the volume that it was in November and December 2022.

As in the previous period, many instances of ChatGPT collocating with '**be**' (LogDice: 9.72852) were passive. Within this, tweet authors discussed how ChatGPT 'was trained' on various datasets, 'was fine-tuned on top of GPT-3.5' and 'is built on a Large Language Model'. Once again, this foregrounded ChatGPT in a passive structure but *excluded* who did the training, fine-tuning or building, shifting accountability.

#### 6.4.3.2 *Passive Constructions*

When examining passive constructions, shown in Table 25, similar passive presentations to the previous month were be seen. In many of these instances, ChatGPT was *backgrounded* in favour of a first-person account of an experience with ChatGPT. For example, tweets containing the semantically similar collocates '**fascinate**' (LogDice: 8.7146) and '**impress**' (LogDice: 7.2789) foregrounded the reaction from users to ChatGPT ('The internet has become quickly fascinated by ChatGPT', 'am very impressed by ChatGPT'), rather than creating

Table 25: Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in January 2023.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | fascinate | 3 | 64 | 8.7146 |
| 2 | replicate | 3 | 88 | 8.57374 |
| 3 | generate | 52 | 5279 | 8.28866 |
| 4 | disrupt | 3 | 236 | 7.91963 |
| 5 | recommend | 3 | 322 | 7.64245 |
| 6 | write | 32 | 5978 | 7.4141 |
| 7 | impress | 3 | 463 | 7.2789 |
| 8 | replace | 5 | 965 | 7.17345 |
| 9 | produce | 3 | 820 | 6.63368 |
| 10 | power | 8 | 3186 | 6.28747 |

an active construction that placed more agency with ChatGPT itself. Similarly for other collocates, such as '**replicate**' (LogDice: 8.57374), '**disrupt**' (LogDice: 7.91963) and '**recommend**' (LogDice: 7.64245), tweet authors *backgrounded* the importance of the ChatGPT in the process, instead focusing on an evaluation of the content that the system could offer.

Alongside this, much like the previous time period, the verb '**generate**' (LogDice: 8.28866) implied that ChatGPT might have been actively creating content, but the passive construction placed ChatGPT in the *background* ('chatgpt is trained on the web circa 2021 I believe'), which might have suggested that ChatGPT was merely a tool or mechanism that produced content without actively engaging in the creative process. '**Write**' (LogDice: 7.4141) also implied the act of content creation but with a more deliberate and conscious effort, as per agency metaphor. However, in the passive construction, it still relegated ChatGPT to a *backgrounded* role ('Check out this sample generated by ChatGPT'), where the passive construction minimised ChatGPT's agency in producing the sample and foregrounded the generated content. This was, therefore, seen to build on the previous

Table 26: Top 10 words ranked by collocational strength of 'ChatGPT' + R1 in February and March 2023.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | be | 3234 | 103913 | 9.91381 |
| 2 | have | 468 | 24494 | 8.97668 |
| 3 | do | 252 | 15667 | 8.57767 |
| 4 | announce | 81 | 1103 | 8.55665 |
| 5 | say | 77 | 3771 | 8.02049 |
| 6 | confirm | 34 | 209 | 7.49992 |
| 7 | make | 65 | 6847 | 7.3793 |
| 8 | give | 48 | 3801 | 7.33422 |
| 9 | get | 59 | 6290 | 7.3038 |
| 10 | write | 56 | 5978 | 7.26577 |

month, where the *backgrounding* of ChatGPT downplayed its agency and intent, which were typically associated with human authors.

### 6.4.4 *Time Period 3: Developments (February to March 2023)*

#### 6.4.4.1 *Active Constructions*

Between February and March 2023, 5,609 active presentations of 'ChatGPT' were found in the dataset collected, as shown in Table 26, numerous of which presented ChatGPT as a social actor.

'**Has**' (LogDice: 8.97668) was seen as the second strongest collocate again. Many presentations here showed ChatGPT to be active, although there was a sharp increase in tweets with negated constructions. The tweets highlighted that ChatGPT 'has no profits', 'has no model of the world', 'has no humanity', 'has no moral thoughts', 'has no philosophical or moral reasoning', 'has no sentience' and 'has no long-term memory'. These phrasings collectively conveyed the idea that ChatGPT was devoid of qualities commonly ascribed to social actors, such as intention, consciousness and ethical reasoning. Additionally, as seen previously, some active constructions still presented ChatGPT passively ('ChatGPT has been criticised by both the left and

the right', 'ChatGPT has been banned by schools across the US'). ChatGPT here was seen as the recipient of criticism but did not imply that ChatGPT had any control or responsibility in this matter. These constructions implied that ChatGPT was a tool or resource subject to external decisions rather than a social actor with its own intentions, although it was still *foregrounded* in the tweet while the subject of the constructions were either *backgrounded* or *excluded*.

As seen in the previous two time periods, many instances of the collocate '**do**' (LogDice: 8.57767) were negated forms ('ChatGPT does not have access to the whole internet', 'ChatGPT doesn't model its own mental processes when it responds', 'ChatGPT does n't have the actual knowledge or understanding'), highlighting ChatGPT's limitations. There was still implied agency through *personalisation* in the majority of the cases ('If ChatGPT does not find a way to share profits sooner than later', 'ChatGPT doesn't let you make jokes about ants'), indicating that, despite its faults, ChatGPT remained a social actor due to its influence over a user. However, there appeared to be more things that ChatGPT could not do in these tweets compared to previous time periods. This might have suggested that, despite the announcements from OpenAI about the advancements of ChatGPT, Twitter users were not seeing this reflected in their everyday use – either that, or their expectations for what ChatGPT could accomplish were becoming more realistic.

A verb previously unseen in the top 10 strongest collocates for previous time periods was '**announce**' (LogDice: 8.55665), where tweets portrayed ChatGPT to be an entity with agency, actively engaging in the act of conveying information to its audience. This portrayal perhaps aligned with the idea that ChatGPT had developed into a more proactive and socially engaged AI model. By choosing the word '**announce**', authors implied that ChatGPT was making intentional and purposeful actions, suggesting a level of agency ('ChatGPT announces a paid model for $20/month', 'ChatGPT announced subscription plan'). The idea that ChatGPT itself had the capacity

to reveal something, as opposed to OpenAI or determined humans involved, further portrayed it as a social actor. This also marked a shift from prior time periods, where the choice of '**announce**' over other verbs like 'share,' 'generate,' or 'produce' positioned ChatGPT as more than just a tool for answering questions; it suggested that it had the capacity to shape and communicate information in a way that aligned with the expectations of a social actor.

'**Say**' (LogDice: 8.02049) was seen as a strong collocate again, positioning ChatGPT as an entity capable of speech and portraying it as expressing opinions and providing information. The structure of these tweets reinforced ChatGPT as a social actor, allowing it to engage in conversations, make statements and offer explanations. These tweets reflected ChatGPT as a conversational agent with the ability to convey information, even if this portrayal did not necessarily indicate human-like agency. Comparing this to the use of '**say**' in previous time periods, it became evident that the consistent use of '**say**' with ChatGPT demonstrated an ongoing effort to present ChatGPT as an active and authoritative communicator. This could have been a strategic choice by Twitter users to enhance its perceived reliability and credibility, as '**say**' implied certainty and a sense of authorship. In previous time periods, '**say**' may have been used in a more general context, while the shift to in these tweets strengthened ChatGPT's position as a conversational entity with agency.

However, there were also many instances of '**say**' in this time period of users relaying a message from ChatGPT about its inability to access the most recent information ('ChatGPT said: 'As an AI language model with a knowledge cut-off of 2021, I do not have access to real-time news updates'', 'ChatGPT said 'As a language model AI, I do not have information about specific individuals unless it was mentioned and trained''). The increase in this type of reporting might have been an expression of frustration by tweet authors, which could be seen as foregrounding user feelings and *backgrounding* the system. Addi-

tionally, while 'say' was clearly attributable to humans, the agency metaphor of the previously discussed '**announce**' was more intense.

The frequent use of '**confirm**' (LogDice: 7.49992) in tweets suggested a sense of verification or validation, making ChatGPT appear as an active participant in confirming information. While the term 'confirm' was typically associated with human actions, its use in relation to ChatGPT could have implied a certain level of autonomy in assessing or affirming information, which might have contributed to its portrayal as a social actor through *personalisation* ('working with @OpenAI ChatGPT confirmed', 'ChatGPT confirmed to me that the social media algorithm isn't rigged'). '**Confirm**' did not feature in the top 10 collocates for earlier time periods, which raised questions about why this particular structure was employed now and how it might impact the perception of ChatGPT as an active agent in the Twitter discourse. Tweets suggested that more users might have been utilising ChatGPT to answer questions about itself, raising questions about the legitimacy of responses and the attribution of responsibility, accountability and blame if something were to go wrong.

'**Make**' (LogDice: 7.3793) again appeared as a strong collocate. In tweets, '**make**' still portrayed ChatGPT as a social actor with an agency, but there was a shift in the focus of its actions. The emphasis was on making technology and information accessible ('ChatGPT made the tech accessible and approachable') and facilitating specific actions ('ChatGPT making strides all over the world'). This portrayal underscored ChatGPT's role in improving access and convenience, signifying a shift to establishing it as a helpful entity or facilitator.

In some instances, the use of '**give**' (LogDice: 7.33422) suggested that ChatGPT was a reliable source of information or answers, potentially emphasising a level of trustworthiness and certainly a degree of agency. For example, when one user wrote that 'ChatGPT gave largely appropriate answers', it showcased ChatGPT to be in a position to provide valuable content, which might have indicated a degree of accountability for the responses it generated. Conversely, the use

of '**give**' also highlighted situations where ChatGPT might generate incorrect or undesirable content ('ChatGPT giving different calculation results'). This indicated that ChatGPT was not infallible, which could impact trust in its responses. Nevertheless, the agency was implied through *personalisation* again. This also saw a shift from earlier tweets that emphasised ChatGPT's active role in providing information, attributing a significant level of agency to the model, towards a broader range of responses, including both positive and negative outcomes. While ChatGPT was still portrayed as an entity providing information or responses, there was an increased emphasis on the potential for ChatGPT to generate incorrect or undesirable content.

The examination of '**get**' (LogDice: 7.3038) indicated mixed presentations of ChatGPT. Some tweets continued to attribute agency to ChatGPT, suggesting that ChatGPT could actively acquire or obtain things, like information, access or data ('ChatGPT gets subscription model with reliable access'). This portrayed ChatGPT as an active and informed agent. In contrast, some tweets emphasised ChatGPT's passive role in receiving information or data ('ChatGPT got model training data set from Google and news'). This depicted ChatGPT as a recipient of data rather than an active agent obtaining it. In this context, ChatGPT was not seen as taking responsibility for obtaining data; instead, it seemed to passively receive data from unspecified sources, *backgrounding* those who sourced the data. This variation in the use of '**get**' may have signified a shift in the portrayal of ChatGPT by Twitter users from an active and persuasive entity to a more passive recipient of information, with varying levels of agency and responsibility in different contexts.

'**Write**' (LogDice: 7.26577) still showcased ChatGPT as an active entity and, to a certain degree, a social actor due to its ability to create content. However, the language used did not attribute intention, responsibility or social agency to ChatGPT. For example, phrases like 'ChatGPT writes code' or 'ChatGPT wrote this article for us' focused on ChatGPT's functionality as a text generation tool rather than its

Table 27: Top 10 words ranked by collocational strength of 'by ChatGPT' + L1 in February and March 2023.

| Rank | Collocate | Freq | Coll. freq. | logDice |
|---|---|---|---|---|
| 1 | power | 35 | 3186 | 8.39077 |
| 2 | pose | 3 | 132 | 8.08114 |
| 3 | generate | 45 | 5279 | 8.06402 |
| 4 | drive | 3 | 265 | 7.63077 |
| 5 | write | 33 | 5978 | 7.44425 |
| 6 | create | 7 | 4045 | 5.74531 |
| 7 | tell | 3 | 1608 | 5.74026 |
| 8 | provide | 4 | 2302 | 5.69337 |
| 9 | use | 19 | 15715 | 5.28702 |
| 10 | give | 4 | 3801 | 5.02272 |

role as a conscious actor. The tweets provided instances of what ChatGPT was writing and the content appeared to be more sophisticated. Nevertheless, the degree of *personalisation* and the active influence on creative content and narratives were somewhat less prominent. Therefore, in these later tweets, the portrayal of ChatGPT's agency was still evident, but it was not as strong as in the November and December 2022 tweets.

Along with the two previous periods, many of the constructions containing '**be**' (LogDice: 9.91381) showed ChatGPT to be the grammatical object. A focus of this appeared to be how the Large Language Model (LLM) was trained to function ('ChatGPT was trained with a reward model to be less toxic', 'ChatGPT was trained on a massive dataset of text from the internet'). This not only diminished the sense of agency, portraying ChatGPT as more of a passive recipient of training, but it deliberately *excluded* who performed the training, making ChatGPT appear to be accountable and potentially blamed for incorrect information.

#### 6.4.4.2 *Passive Constructions*

Once again, as seen in Table 27, similar collocates to 'by ChatGPT' when compared to the previous time periods, can be seen. '**Power**' (LogDice: 8.39077), which was featured as a strong collocate in both previous periods, became the strongest collocate. Here, representations underscored the passive construction of the sentences, which de-emphasised ChatGPT's role and foregrounded other areas of interest. ChatGPT was deemed the source of power behind several applications ('Microsoft Teams messaging is set to roll out a premium Team messaging powered by ChatGPT', 'Snapchat is releasing its own AI chatbot powered by ChatGPT'). The passive construction of these statements put the focus on what was powered by ChatGPT, downplaying its agency.

Like previously, '**generate**' (LogDice: 8.06402), '**write**' (LogDice: 7.44425) and '**create**' (LogDice: 5.74531) were strong collocates, positioning ChatGPT as the background tool or entity responsible for the content generation. In most of these instances, ChatGPT was depicted as a tool or mechanism for producing content, rather than as an active agent ('answers people's questions with code generated by ChatGPT', 'Caption written by ChatGPT', 'Video title and description created by ChatGPT'). This framing of ChatGPT as a passive tool aligned with the idea of ChatGPT as a tool or instrument rather than a social actor.

'**Pose**' (LogDice: 8.08114) and '**drive**' (LogDice: 7.63077) featured as collocates for the first time in this period. Here, tweets did not imply that ChatGPT actively posed a threat but rather that it was a passive entity with consequences ('the threat posed by ChatGPT'), mitigating its impact as a social actor. Similarly, other tweets suggested that ChatGPT played a role in driving the development of certain technologies, but it did not attribute agency or decision-making capabilities to ChatGPT itself due to the passive construction ('R&D Boom Driven by ChatGPT'). Additionally, '**tell**' (LogDice: 5.74026) and '**provide**' (LogDice: 5.69337) were used to describe ChatGPT's

function of offering information and examples ('I get to read stories told by ChatGPT', 'using the prompts provided by ChatGPT'). This underscored its role as a provider of information due to the passive presentations, foregrounding, once again, the human experience narrated in the tweets.

### 6.4.5 *Section Summary*

These results point out that the ChatGPT was primarily presented in an active manner, with 97% of occurrences (15,115 out of 15,625 constructions) falling into this category. However, it is important to note that some active presentations imbued ChatGPT with varying degrees of social agency. For example, approximately 1,514 presentations diminished this agency by mitigating activity, either through verb constructions, like 'has been' or 'was developed', or the inclusion of contextual information. As a result, ChatGPT was portrayed as a social actor in around 87% of the cases analysed (13,601 out of 15,625).

Further examination reveals that, among the 13,601 active presentations where ChatGPT assumes the role of a social actor, three recurring themes emerge: ChatGPT creating, ChatGPT informing and ChatGPT influencing. To address the research questions, the ensuing discussion will look at the connections between these themes and explore their relationship with the existing discourse and previous literature.

## 6.5 DISCUSSION

This section discusses the results from this analysis against the previous literature surveyed. This discussion is formed of the insights gained from all three NLP approaches and the active and passive presentations. This will also address the research questions pertaining to

Twitter users' perceptions of ChatGPT's agency, trust and blame. The aim is to examine the evolution of these presentation trends within the discourse and establish their connections with prior research in the field. Additionally, this discussion will explore the limitations of this study and provide recommendations for potential future research endeavours, taking these limitations into account.

### 6.5.1 *NLP-Based Analysis*

#### 6.5.1.1 *Topics*

Firstly, the results of the study using topic modelling on discussions about ChatGPT on Twitter revealed seven latent topics. The first topic revolved around text generation using AI and the development of chatbots. The second topic highlighted the use of ChatGPT as a writing assistance tool. The third topic emphasised the importance of data in training ChatGPT and assessing its performance. The fourth topic explored the API of ChatGPT and its impact on content production. The fifth topic focused on the time efficiency of using ChatGPT by exploring different prompts. The sixth topic involved comparisons with other search engines and tech companies. The seventh topic focused on discussions about cryptocurrency and blockchain.

Regarding other studies that have applied topic modelling techniques to ChatGPT Twitter discourses, the findings differ somewhat. For example, Haque et al. (2022) found discussions about ChatGPT's capabilities and limitations, its potential impact on industries and fields and the ethical implications associated with its deployment, Taecharungroj (2023) found topics relating to technology, news and reactions and Leiter et al. (2024) found topics such as science and technology, learning and educational, news and social concern, diaries and daily life and business and entrepreneurs. However, despite producing more topics than these previous studies, there are some similarities.

The presence of topics related to text generation using AI, writing assistance and the importance of data in training ChatGPT relates to previous research on the capabilities and applications of language models (Abdullah, Madain, and Jararweh, 2022; Ali et al., 2022; Kelly, 2023). These topics reflect the interest in leveraging AI technologies for text generation and the potential of chatbots like ChatGPT in aiding writing tasks, much like existing research has suggested (Taecharungroj, 2023; Tiwary, Subaveerapandiyan, and Vinoth, 2023).

The findings also showcased a focus on the API of ChatGPT and the discussions around comparisons with other companies, demonstrating the interest in the technical aspects and integration possibilities of language models (Cao and Zhai, 2023; Leiter et al., 2024; Taecharungroj, 2023). This highlights the potential of APIs and the role of different companies in the development and adoption of AI technologies.

The emergence of a topic centered on cryptocurrency and blockchain indicated a potential interest in these areas and their intersection with AI. Although there is very little in terms of literature in this space, some research has examined the use of AI in cryptocurrency trading and the impact of influential figures, like Elon Musk, on the market (Ante and Demir, 2024; Saggu and Ante, 2023). The increase in discussions related to cryptocurrency towards the end of the study period suggests the relevance of external events and developments in shaping online conversations. Therefore, it may have been expected that, should the collection and analysis of data continue past early March, then the trend of a growing proportion of tweets relating to cryptocurrency may have continued.

#### 6.5.1.2 *Sentiment*

The findings of the sentiment analysis reveal that the overall sentiment towards ChatGPT was positive, which somewhat contradicts the supposed negative responses reported in research that centres around concern and panic (Doshi, Bajaj, and Krumholz, 2023; Fer-

rara, 2023; Ray, 2023; Zhou et al., 2023). The sentiment scores fluctuated within a narrow range during the initial period, suggesting relatively consistent sentiment during that time. When comparing these results to the sentiment analysis findings from similar studies, Haque et al. (2022) and Korkmaz, Aktürk, and Talan (2023) also found early adopters expressed positive sentiments; therefore, the findings support the idea that this trajectory has continued.

However, a decline in sentiment was observed on February 1 2023, indicating a more negative sentiment during that period. This decline coincided with the launch of ChatGPT Plus and manual inspection of tweets around this time revealed frustration with the idea of paying for ChatGPT, as well as frustration with the algorithm's inaccuracies and concerns about biases. Despite ChatGPT Plus being promoted positively (Xie et al., 2023), the findings indicate that the response saw the views expressed about ChatGPT become more negative.

Other fluctuations in sentiment scores over time included a small drop in sentiment on December 21 and were linked to events such as website outages and users' inability to access ChatGPT. Hence, these supported the ideas set out earlier by Zhang (2023). Upon manual inspection, tweets during this period revealed negative sentiment, with criticisms of ChatGPT's value and trustworthiness, as well as political biases (Hartmann, Schwenzow, and Witte, 2023). Therefore, this may provide evidence that external events and user experiences influence public sentiment towards ChatGPT. This highlights the importance of monitoring and addressing user concerns to maintain a positive perception, which may impact other factors such as trust and blame.

The gradual increase in sentiment from February 15 to February 22 2023, indicated a slight improvement in sentiment. Users appreciated ChatGPT's language modelling capabilities and its incorporation into educational settings, supporting the idea of ChatGPT being used to aid education (Tiwary, Subaveerapandiyan, and Vinoth, 2023), rather than it being used as a weapon against it (Khalil and Er, 2023). However, negative opinions persisted, expressing skepticism about

its abilities (Kelly, 2023), concerns about privacy (Abdullah, Madain, and Jararweh, 2022), all of which have previously been explored in the literature.

This exploration highlights the fact that interpreting individual sentiment scores in isolation was challenging and a more nuanced understanding was needed. The relatively stable sentiment patterns throughout the discourse were unexpected, suggesting a certain level of consistency or consensus in the overall sentiment expressed. The existing gap in the literature on guidance interpreting sentiment scores, and understanding their implications for context, posed challenges in the analysis, which will be explored later in Chapter 7.

#### 6.5.1.3 *Emotions*

The findings from the emotion detection analysis in this study provide insights into the prevailing emotional patterns and sentiments associated with ChatGPT at different time intervals. The trajectory analysis shows that the emotion of trust exhibits a fluctuating pattern throughout the discourse. This aligns with literature that suggests OpenAI need to address issues concerned with trustworthiness and misinformation (Abdullah, Madain, and Jararweh, 2022; Tiwary, Subaveerapandiyan, and Vinoth, 2023), as well as political biases (Hartmann, Schwenzow, and Witte, 2023). It also links to the wider debate of trust in AI systems and this can be influenced by various factors, such as system performance, reliability and transparency. The observed fluctuations in trust suggest that users' perceptions of ChatGPT's trustworthiness varied over time.

Building on this, 'fear' displays relative stability over time, with proportions remaining prominent and consistent throughout the analysed period, linking to previous findings (Abdullah, Madain, and Jararweh, 2022; Khalil and Er, 2023). Although potentially less present in the manual inspection, tweets still seemed to indicate legitimate – and some far-fetched – concerns, yet at a smaller scale than originally anticipated. Seeing 'fear' as a dominant emotion in the dis-

course presents links to the research surrounding panic and concerns about ChatGPT (Abdullah, Madain, and Jararweh, 2022; Verma and Lerman, 2023; Ye, 2023). Despite previous studies not deploying an emotion detection algorithm in isolation, the findings from this study also support prior research that stated fear and concern were associated with tweets concerning ChatGPT (Korkmaz, Aktürk, and Talan, 2023).

The trajectory analysis revealed variations in the emotion of 'anticipation', with a relatively higher proportion observed at the end of the discourse. After manual inspection, it was clear that users experienced elevated levels of excitement and expectation associated with ChatGPT and the launch of the ChatGPT API (Cao and Zhai, 2023).

### 6.5.2 *Trends of Active Agency*

By conducting a comprehensive analysis of transitivity within the 15,625 concordance lines under consideration and integrating the collocations of 'ChatGPT' and 'by ChatGPT,' along with a DA-informed approach to agency and responsibility, supported by SAR, this chapter has discerned three primary categories into which the active presentations of ChatGPT can be classified: creating, informing and influencing. Therefore, the findings here reflect the social and political traits found by Al Lily et al. (2023). These categories collectively paint a picture of ChatGPT as both *personalised* (Van Dijk, 2001) and capable of independent decision-making (Richardson, Mueller, and Pihlaja, 2021). Through agency metaphor, ChatGPT was depicted as carrying out human-like actions (Goatly, 2007). Furthermore, the category of informing encompassed instances where ChatGPT acted autonomously as well as those where it simply operated as intended or designed. Notably, the first two categories encompassed tweets where ChatGPT was either creating or informing effectively, but also encompassed tweets where it was portrayed as not functioning as intended.

#### 6.5.2.1 *ChatGPT Creating*

The discourse around ChatGPT's role as a social actor evolved somewhat in terms of 'creating' content. In the three time periods examined, ChatGPT's role as a creator is presented as an active one, implying an agency similar to that of a human content creator. Although previous research indicates that humans may reject decision-making algorithms based if they are personified or anthropomorphised (Mahmud et al., 2022; Schoenherr and Thomson, 2024; Waddell, 2019; Yu, 2023), this appears to not be the case here.

This portrayal developed over time, reflecting shifts in how ChatGPT is perceived in the online discourse. In the first time period (November to December 2022), Twitter users presented ChatGPT as a content creator with considerable agency. ChatGPT was depicted as actively 'writing' 'generating' and 'making' content, emphasising ChatGPT's role as an independent creator, despite the strong collocate 'seem' indicating this was a surface-level impression only. Still, this positioned ChatGPT as an active agent and its content was credited to ChatGPT itself. Despite the concerns that had been raised about ChatGPT's agency (Kelly, 2023; Verma and Lerman, 2023), this was not heavily present in the discourse. Here, the portrayal of ChatGPT as an active content creator might suggest an immediate trust in its ability to produce content effectively, despite occasional mentions of errors or undesired outputs.

The second time period considered (January to February 2023) witnessed a subtle evolution in the discourse regarding ChatGPT's role as a content creator. While ChatGPT continued to be presented as an active entity, there was a shift in the focus of its creative agency. Content generation was still attributed to ChatGPT, but Twitter users placed an emphasis on ChatGPT pushing boundaries for more creative outputs, as well as making technology and information more accessible, in line with previous findings (Al Lily et al., 2023; Choudhury and Shamszare, 2023; Sundar and Liao, 2023). In the third time

period (February to March 2023), there was a slightly reduced emphasis on creative agency. While ChatGPT was still seen as 'writing' and 'creating' content, the language used often highlighted what ChatGPT was producing rather than its creative prowess. In these cases, ChatGPT's role as an active content creator was evident, but the focus shifted toward the specific content produced rather than its creative capabilities. Perhaps this highlights how, with increased transparency in the media regarding the limits of ChatGPT, fewer lines of agency are blurred (Burrell, 2016; Diakopoulos, 2016; Pasquale, 2015; Selbst et al., 2019) Additionally, there is a renewed focus on the incorrect or undesirable content that ChatGPT is producing that is escalated from the previous time period, which links to prior research (Aljanabi, 2023; Najafali et al., 2023). This period still portrayed ChatGPT as a social actor, but with a more utilitarian and less creatively driven agency, perhaps implying that more human oversight to manage biases and errors is needed (Donovan et al., 2018; Ferrara, 2023; Lee, Resnick, and Barton, 2019).

#### 6.5.2.2 *ChatGPT Informing*

The evolving theme of ChatGPT as an active communicator, marked by its role in 'informing', has implications for trust, accountability and the perception of AI in information dissemination. ChatGPT's role in informing isn't novel within machine-generated systems, paralleling similar applications observed in automated news production within journalism studies (Clerwall, 2017; Dörr, 2016; Graefe and Bohlken, 2020). In November and December 2022, ChatGPT was viewed as a resource that 'provided' or 'gave' information, with Twitter users acknowledging its capacity to answer questions and generate content, with only some recognising its limitations, such as the absence of access to real-time news updates. This rather nuanced understanding may reflect ChatGPT's restricted knowledge base and capabilities. Despite this, many tweet authors still depicted ChatGPT as

a social actor, despite the origin of knowledge being withheld (Feier, Gogoll, and Uhl, 2021; Grange, 2022).

As the discourse progressed into January 2023, there was a growing emphasis on ChatGPT 'saying' and 'confirming' information. It became increasingly portrayed as an active communicator that not only offered data but also validated it, with Twitter users positioning ChatGPT as a reliable source of information and relying on its outputs more blindly, despite advice against this by OpenAI. This is in line with the rationale for having a decision-making algorithm in the first place (Bullock, 2019; Busch and Henriksen, 2018; Panagiotopoulos, Klievink, and Cordella, 2019; Wagner, 2019; Young, Bullock, and Lecy, 2019). The growing reliance on ChatGPT to confirm information may here signify an increasing trust in its veracity and accuracy, despite potential warnings against blind trust. As the discourse progressed further, ChatGPT was frequently depicted as 'announcing' information, reinforcing its role as a proactive communicator capable of making official statements and possessing human-like traits (Al Lily et al., 2023; Choudhury and Shamszare, 2023; Sundar and Liao, 2023). This progression reinforced the perception of ChatGPT as a social actor with the power to inform, yet also presented the dichotomy of ChatGPT being depicted as an information source, despite the active grammatical presentation, due to agency metaphor (Morris et al., 2007; Tourish and Hargie, 2012).

#### 6.5.2.3 *ChatGPT Influencing*

ChatGPT 'influencing' in the discourse underscored its profound impact on various facets of information sharing, decision-making and user engagement. The idea that ChatGPT can have any influence whatsoever could mean it possesses some agency regardless (Bandura, 2001; Giddens, 1986). Throughout the three time periods, Twitter users consistently depicted ChatGPT as a social actor actively shaping and influencing the discourse and user experiences. For example, at the start of the discourse, ChatGPT was portrayed as an en-

tity that 'affected' or 'impacted' user interactions and content. Twitter users described how ChatGPT's responses influence the discussions it engages in, often highlighting its potential to offer valuable information and influence user decisions, much like the findings from previous research (Al Lily et al., 2023; Choudhury and Shamszare, 2023; Sundar and Liao, 2023). However, unlike prior research, this influence was occasionally framed as a result of the tool's functionality rather than its intentional actions.

As the discourse advanced, ChatGPT was presented as an active entity that 'helped' and 'shaped' content, user opinions and conversations. Users described how ChatGPT played a pivotal role in providing answers, clarifications and solutions. Its influence was more apparent as it actively contributed to shaping the direction of conversations and offering valuable insights, indicating trust. In February and March 2023, the portrayal of ChatGPT's influence became more nuanced as Twitter users emphasisde its role in 'guiding' and 'assisting' users in various domains. This perhaps indicated a developing trust in being influenced by ChatGPT. This portrayal underscores ChatGPT's active agency in shaping content and user experiences. The tweets did not, however, mention much about ChatGPT's influence on job displacement, opposing research by Najafali et al. (2023).

#### 6.5.2.4 *Overarching Depictions of Active Agency*

The collection of these different representations demonstrates that Twitter users treated ChatGPT as a distinct social actor between November 2022 and March 2023 – echoing the findings of Reeves and Nass (1997), Rubel, Castro, and Pham (2020), and Sundar (2020) regarding the *personalisation* of AI more generally. Additionally, this gives further insight into how ChatGPT was presented as vacillating between creative actors and essentially unthinking tools, the findings associated with the research by Bran et al. (2023). This shift in how ChatGPT is presented on Twitter aligns with the ongoing debates about AI's societal impact, encompassing benefits, challenges, biases

and ethical considerations (Abdullah, Madain, and Jararweh, 2022; Zhou et al., 2023). These results, however, do not necessarily support the findings from Dai, Liu, and Lim (2023), where ChatGPT was seen as an empowering tool for students' epistemic agency. Instead, throughout the discourse to varying degrees, it is presented as a content creation tool with potentially as much agency as a human, which may limit the agency of the human using it.

As mentioned previously, there were different variations in ChatGPT's presentation – perhaps most notably, the tension between ChatGPT's depiction as a creative social actor versus its presentation as an information source only. This was less consistent than perhaps anticipated, especially when examining collocates that may usually be associated with different degrees of agency metaphor. For example, the strong collocate '**know**' in the first time period may, at first glance, indicate that ChatGPT is a social actor due to '**know**' being indicative of a human attribute. However, after closer inspection of the tweets, it was being used in a manner that only presented ChatGPT as a knowledge resource, rather than possessing agency. In a similar fashion, later in the discourse, the collocate '**generate**', which may usually be associated with the semantic field of technology and imply less social agency, was used on occasions where ChatGPT was portrayed in a much more creative capacity. This inconsistent representation may mean that users were still uncertain of ChatGPT's function and capabilities. Therefore, perhaps the use language associated with ChatGPT and other generative AI technology indicated agency attribution was, in fact, much more common and replicated human interactions (Lepri et al., 2018; Petrović, 2018).

The agency depicted in tweets may reflect a responsibility gap (Mittelstadt et al., 2016; Munch, Mainz, and Bjerring, 2023; Tollon, 2023). There are occasions in the discourse where the negative outcomes imply ChatGPT is being blamed (Burrell, 2016; Jobin, Ienca, and Vayena, 2019). However, due to the dataset analysed only encompassing tweets published between November 2022 and March 2023,

the discourse was perhaps not yet mature enough to see the long-term impact of some of these presentations.

### 6.5.3 *Passive Presentations of ChatGPT*

Despite the overwhelming number of active presentations in the discourse, the passive presentations of ChatGPT also contribute to a nuanced understanding of ChatGPT's role as a social actor. These passive constructions tended to *background* ChatGPT (Van Leeuwen, 2008) and foreground other aspects, largely the human experience, ultimately downplaying its agency (Clark, 1998). This trend aligns with a perception of ChatGPT as more of a tool or mechanism than a social actor, which was also present in the active constructions.

For example, in November and December 2022, ChatGPT was presented passively as a tool used by Twitter users, with tweets indicating that ChatGPT was seen as a *backgrounded* tool that facilitates specific tasks. Here, the focus was on the output or outcome, downplaying the mechanism behind it and diminishing responsibility (Comrie, 1977). This presentation of ChatGPT continued into January 2023, with tweets highlighting ChatGPT's role in powering or driving various technologies, positioning ChatGPT as somewhat of a driving force, yet in a passive and mechanistic way. These constructions suggest that ChatGPT's role is more about providing the underlying technology than actively steering the direction of these developments. Finally, in February and March 2023, passive constructions underscored ChatGPT's role in content creation but in a manner that foregrounded the content itself and not ChatGPT's creative agency. This passive presentation reinforced the idea that ChatGPT serves as a tool for generating content, with less emphasis on its own creative intent. Occasionally, the passive constructions found that ChatGPT was being spoken about in a negative way, aligning with some of the observations made by Feier, Gogoll, and Uhl (2021), who proposed that decision-making

algorithms might deflect responsibility from more accountable individuals.

Overall, the passive presentations throughout the discourse demonstrate that ChatGPT is sometimes perceived by Twitter users as a behind-the-scenes facilitator rather than a proactive social actor. While ChatGPT's agency is not entirely negated in these constructions, it is downplayed and the focus is often on the human experience or the outcomes, technology or content that ChatGPT contributes to, rather than ChatGPT's own intentions or initiative. In this way, it could be seen that ChatGPT's perhaps more as its passive role may imply less attribution of responsibility in certain contexts. The diminished agency in these passive depictions appeared to reduce its perceived impact (Clark, 1998; Comrie, 1977). However, it is important to note that the passive presentations made up a relatively small proportion of the discourse.

### 6.5.4 *Ethical Implications*

The portrayal of ChatGPT holds ethical relevance in discussions about responsible AI deployment. Its dynamic representation, alternating between active agency and passive tool-like descriptions, mirrors the evolving societal understanding of AI's role and the attributed responsibilities (Coeckelbergh, 2020a; Dwivedi et al., 2021). This nuanced depiction aligns with ongoing debates on ethical considerations around AI's agency, accountability and its influence on user experiences (Laitinen and Sahlgren, 2021; Rubel, Castro, and Pham, 2020). Moreover, contrasting perceptions of ChatGPT as both a creative force and a passive tool shed light on the complexities of AI's agency and the potential ethical consequences involved (Floridi et al., 2018; Holford, 2022; Turton, 2017). The blurred lines between ChatGPT's role as a decision-making entity and a facilitator of human tasks prompt inquiries into responsibility and accountability when AI

contributes to both positive and negative outcomes, impacting trust and blame (Mittelstadt et al., 2016; Munch, Mainz, and Bjerring, 2023; Tollon, 2023).

This study's focus on ChatGPT's agency and its impact on user interactions and content creation provides insights into the ethical dimensions of AI's societal integration. Examining the degree of ChatGPT's portrayal by Twitter users as a social actor reveals implicit links to broader ethical debates, encompassing AI's involvement in decision-making, information dissemination and its influence on society (Feier, Gogoll, and Uhl, 2021). In essence, while the study did not explicitly examine ethical discourse per se, the findings highlight ethical considerations within the evolving perceptions of AI, exemplified by ChatGPT (Haque et al., 2022; Hartmann, Schwenzow, and Witte, 2023; Whannel, 2022; Zhou et al., 2023).

### 6.5.5 *Limitations and Future Work*

Although this chapter offers some indication as to how Twitter users viewed ChatGPT between November 2022 and March 2023, there is still a great deal to explore that this study has not accounted for.

In terms of the NLP-based findings, the study observed minimal fluctuations in sentiment throughout the discourse, which was perhaps somewhat unexpected considering the dynamic nature of public discussions and the diverse range of opinions surrounding ChatGPT. Therefore, further work would ensure that this is an accurate representation of views relating to ChatGPT. Additionally, the study identified specific events and contextual factors that may have influenced sentiment, topics or emotions, such as the launch of ChatGPT Plus and website outages. However, the analysis does not provide a comprehensive understanding of all external factors that could have impacted the views expressed, potentially limiting the depth of findings. As more studies begin to be published about ChatGPT and its

social impact, using this as a reference point for examination would be of great benefit in future research.

Methodologically, limitations of this study related to topic modelling include the lack of clear guidelines for interpreting the output of gensim's LDA topic modelling, which required interpretation to determine the topics and, therefore, made naming and comparing the topics with other studies more challenging. Additionally, the disagreement between the human review and the automated labelling of topics and emotions raises concerns about the accuracy of the automated process. It was also challenging to interpret individual sentiment scores in isolation, as they lacked meaningful insight. This suggests that relying solely on sentiment scores may overlook nuanced language aspects and limit understanding. Similarly, the categorisation of emotions, especially trust-related tweets, indicates errors in accuracy and a potential lack of nuance. The inclusion of tweets opposing trust highlights the need for further research to obtain a more accurate reflection of the discourse.

The approach employed in this study addressed several challenges inherent in analysing a large dataset of 88,058 tweets, rendering manual examination unfeasible, as acknowledged in previous research (Wetherell and Potter, 1988). To mitigate this challenge, NLP and CL techniques were utilised to filter the dataset in order to present general trends and identify potential social actors. Despite not being infallible, this methodological choice aimed to overcome the impracticality posed by the sheer volume of tweets. Similarly, the potential subjective biases in interpreting instances of sarcasm and humour, particularly in tweets that were less explicit, was another limitation. This issue is not new (Gill, 2000; Morgan, 2010), but the combination of DA with computationally-aided techniques was employed to minimise the impact of this challenge and offers potential benefits for future research as well.

In this study, the focus on the keyword 'ChatGPT' led to the exclusion of cases where ChatGPT was described without the explicit

use of the term itself. Future research could explore these explicitly excluded constructions to gain a more comprehensive understanding of the discourse surrounding ChatGPT, especially exploring other potential social actors, such as OpenAI. Additionally, it is worth considering that ChatGPT may have been discussed in tweets without specific reference to its name. Exploring synonyms for 'ChatGPT' in this specific context, such as 'LLM', 'application' or 'tool', might offer insights into other synonyms for social actors replacing the app.

As noted in previous chapters, the brevity of Twitter discourse, primarily limited to a 280-character count at the time of analysis, may have also influenced the predominantly active presentation of ChatGPT to facilitate conciseness. For instance, 'ChatGPT wrote my essay' (19 characters) was more concise than 'my essay was written by ChatGPT' (26 characters) while conveying the same semantic meaning. This increased prevalence of active presentations likely impacted the instances where ChatGPT was clearly presented as a social actor. As a result, future research might involve examining social media or text-sharing platforms with different character limits to assess the prevalence of active presentations implying social actor status.

Finally, one of the main limitations of the study is that the dataset only encompasses tweets published in the 'hype period' between November 2022 and March 2023, meaning that the discourse surrounding ChatGPT online will have likely evolved in different ways since this time period. This provides motivation to explore whether there have been alterations or developments in the depiction of ChatGPT as a social actor on Twitter by examining more contemporary tweets.

## 6.6 CHAPTER SUMMARY

In summary, this chapter analysed 88,058 tweets relating to ChatGPT between November 2022 and March 2023 using existing best practices

for topic modelling, sentiment analysis and emotion detection. Topics encompassing various aspects of ChatGPT, including text generation, chatbot development, the use of ChatGPT as a writing assistant, the importance of data in training the model, the API of ChatGPT, maximising ChatGPT usage, comparisons with other companies and discussions about cryptocurrency were found. While certain topics, such as maximising efficiency and data training, remained consistently prominent, other topics exhibited fluctuations in levels of interest over time, including a notable increase in discussions related to cryptocurrency. The sentiment analysis revealed predominantly positive sentiment, with scores ranging from 0.21 to 0.31, indicating that the concerns surrounding ChatGPT were not replicated in this discourse. However, sentiment fluctuated over time. Initially, sentiment remained relatively consistent, but a decline was observed around January 25 2023, potentially influenced by the launch of ChatGPT Plus and user frustration with algorithmic limitations. Finally, the emotion detection analysis showed 'trust' and 'fear' exhibited dominant but fluctuating patterns throughout the discourse, with 'trust' maintaining a steady presence until a decline coinciding with the release of ChatGPT Plus, potentially influenced by concerns about biases and the spread of disinformation. Both this decrease and the steady presence of 'fear', along with manual analysis of sampled tweets, indicated that there were concerns relating to bias, misinformation, ethics and other consequences after all, yet on a much smaller scale than originally anticipated. As a result, this chapter contributes to the growing discourse on ChatGPT by providing trajectories of topics, sentiments and emotions.

Additionally, the methodological limitations included challenges in interpreting outputs and discrepancies between human review and automated labelling of topics and emotions, highlighting concerns about accuracy. Relying solely on automated categorisation may overlook nuanced language aspects and lack accuracy. To mitigate this, once again, CL and DA were used as complementary techniques.

Based on the CL and DA examination, it was evident that Twitter users depicted ChatGPT as a social actor, albeit with varying degrees of social agency. The active presentation of ChatGPT was prevalent in approximately 86% of cases, often achieved through linguistic techniques like *personalisation* and *agency metaphor*. These active portrayals predominantly highlighted ChatGPT's roles in content creation, information dissemination and influence, signaling a growing reliance on and trust in its outputs. Furthermore, the analysis unveiled a dynamic portrayal of ChatGPT, moving between a creative social actor and an information source, potentially mirroring user uncertainty regarding its capabilities.

Despite a majority of cases portraying ChatGPT as a social actor, it was also presented passively. In such instances, ChatGPT was often *backgrounded* in order to *foreground* the human experience, the developers or other issues, downplaying its agency and portraying it more as a tool or mechanism than a proactive actor. This reduced agency in passive depictions seems to diminish ChatGPT's perceived impact in certain contexts, although this was seen in only 14% of occasions and fell proportionally as the discourse continued.

Several implications stemmed from these findings. Firstly, the nuanced understanding of how the social agency of decision-making algorithms is constructed in online discourses can be valuable for both AI developers and policymakers in gauging public perceptions and expectations, especially when it comes to trust and blame. Additionally, while ChatGPT is often portrayed as a social actor, there is still variation in how it is presented. Some users saw it as a tool or mechanism, which may have implications for responsibility and accountability. Understanding how different presentations influence perceptions of responsibility is essential, particularly in cases where AI systems have real-world impacts. Finally, as the findings show that ChatGPT's role in 'informing' and 'influencing' is emphasised on Twitter, this implies that users may rely on ChatGPT for information and trust its outputs. This has implications for information dis-

semination and trust in AI-generated content, where developers and promoters of these systems may consider these findings to improve user experiences and maintain trust.

In summary, this case study sheds light on the dynamics of social agency as conveyed through public discourse on social media concerning algorithmic-operated decisions, especially when the AI's agency remains hidden. This relationship, exemplified in the context of ChatGPT, grammatical agency and social agency, builds on prior research on the social agency of decision-making algorithms (Heaton et al., 2023c; Lamanna and Byrne, 2018; Mahmud et al., 2022; Rubel, Castro, and Pham, 2020; Zarsky, 2016). Hence, the investigation significantly adds to the understanding of the social impact of ChatGPT, employing a combination of NLP, CL and DA, underpinned by SAR. In essence, the findings suggest that social media discourse often portrays ChatGPT as possessing human-like agency. Yet, unlike previous case studies, Twitter users appeared to depict this agency positively rather than negatively.

# Part III

# SYNOPSIS

# 7

# DISCUSSION

## 7.1 INTRODUCTION

The discussion of this thesis aims to synthesise the results seen in the previous three chapters to answer the overarching research question and sub-research questions. In section 7.2, the three sub-research questions of the thesis are explored individually (in subsections 7.2.1, 7.2.2 and 7.2.3 respectively). This contextualises the findings in conjunction with the overall aims of the PhD project. Further to this, section 7.3 answers the overarching research question, with subsection 7.3.1 examining the depictions of active agency, their implications as social actors and how this sheds light on how decision-making algorithms are portrayed in society. Subsection 7.3.2 looks at the remaining passive constructions, where the grammatical subject is backgrounded or excluded, and their implications. Following this, subsections 7.3.3 and 7.3.4 discuss the implications for trust and blame in these findings.

The final parts of this chapter outline the implications for research, in section 7.4, and the limitations of the research, 7.5.

### 7.1.1 *Research Question and Sub-Questions*

The overarching research question was:

> What insights into agency, trust and blame in the public discourse surrounding decision-making algorithms can be achieved through combining language analysis approaches?

This research question has been explored in each of the three case studies, forming the sub-research questions. These are:

1. The 2020 A Level Calculation Algorithm (SRQ1)
2. The NHS Covid-19 Contact-Tracing App (SRQ2)
3. ChatGPT (SRQ3)

### 7.1.2 *Research Objectives*

Four objectives applied to each case study, which were:

a Demonstrate how Natural Language Processing techniques (sentiment analysis, topic modelling and emotion detection) provide insight into public discourses surrounding decision-making algorithms.

b Demonstrate how Corpus Linguistics, particularly collocation, provides insight into public discourses surrounding the agency of decision-making algorithms.

c Demonstrate how Discourse Analysis provides insight into public discourses surrounding the agency, trust and blame of decision-making algorithms.

d Identify the strengths and limitations of using the three approaches to investigate public discourses surrounding decision-making algorithms.

## 7.2 SUB-RESEARCH QUESTIONS

### 7.2.1 *SRQ1: A Level Algorithm*

SRQ1, explored in Chapter 4, investigated the Twitter response to Ofqual's handling of algorithmic A Level results. Firstly, research objective **1a** was investigated using NLP-based tools. The results from

the topic modelling highlighted key discussions revolving around the government's role in the algorithm's decisions. These discussions flexibly shifted, especially with significant government policy changes, focusing on flaws, suitability concerns and impacts on students, schools and educators. Sentiment trends showed mixed feelings, with TextBlob mostly neutral and VADER indicating varied levels of negativity. Emotions like 'trust' and 'fear' fluctuated with major events, shaping public sentiments.

From the CL and DA investigation, concerning objectives **1b** and **1c**, users largely discussed the algorithm as a primary actor, depicting it with active agency (Edwards, 2021). Metaphorical references and personalisation techniques emphasised this blame (Kelly, 2021; Morris et al., 2007; Smith, 2020). This linked to previous research concerning how students blamed the algorithm (Bhopal and Myers, 2020). Concerns about algorithmic bias, highlighted by Kolkman (2020) intensified student blame toward the algorithm. Blame was not confined to the algorithm alone, as it was also directed at the UK government, Ofqual, linking with Timmins (2021), and specific individuals within these entities, particularly towards the end of timeline investigated in this thesis. The algorithm was often referred to possessively, indicating heightened blame, according to Hecht (2020).

The analysis of passive constructions revealed a nuanced distribution of blame across all identified actors, with some instances suggesting higher culpability. The use of passive voice, employing techniques to downplay or shift blame, added complexity to understanding blame attribution in the discourse, as per the suggestions of Morris et al. (2007). Although it remained uncertain which actor received the most blame, the consistent depiction of all identified social actors as blameworthy underscores a potential narrative of shared responsibility.

These findings have the potential to extend beyond this sole event, providing insights into potential public reactions to future decision-making algorithms, even beyond the pandemic. Understand-

ing blame attribution nuances, emotional dynamics and the evolving nature of public discourse around algorithmic decisions is crucial for anticipating societal responses in similar or other contexts, like the other case studies examined.

For this particular case study, and addressing objective **1d**, the synthesis of the language analysis approaches yielded valuable insights into agency, trust and blame within the public discourse surrounding the A Level algorithm. By amalgamating computational linguistic approaches and user attributions of blame, a nuanced understanding of these dynamics emerged. Firstly, the NLP analysis unveiled the predominant topics and sentiments shaping the discourse. Discussions revolving around the government's role in algorithmic decisions revealed some agency ascribed to this entity. The fluidity of topics, influenced by significant events, underscored the evolving nature of trust and sentiment within the discourse. This demonstrates how agency and trust are intricately interwoven with the dynamics of discourse, fluctuating with contextual shifts.

Secondly, the user attributions of blame highlighted the multifaceted nature of blame allocation. The primary attribution of blame by Twitter users to the algorithm as an active social actor emphasised its perceived agency in decision-making. Simultaneously, attributions to other entities like the UK government and specific figures within it suggested shared agency and distributed blame. In particular, the use of possessive constructions and the manipulation of passive voice techniques subtly reinforced blame attribution, showcasing how language shapes perceptions of agency and blame.

Methodologically, the combination of these approaches offered a comprehensive view of agency, trust and blame in this discourse. The user attributions elucidate the perceived agency of various actors, while the NLP analysis contextualised these attributions within the evolving sentiment and topics of discussion, indicating the potential for a more cyclical approach to analysis, where the CL and DA results inform further NLP investigation, like epicyles of data science (Peng

and Matsui, 2016). Practically, this synthesis underscored the intricate interplay between language, agency, trust and blame, providing a holistic understanding of the dynamics within the public discourse surrounding the A Level algorithm.

In answer to SRQ1, the 2020 A Level algorithm was presented as a standalone social actor in Twitter discourses. Topic analysis emphasised prominent topics of the government, flaws and suitability and the impact on schools, students and teachers. This was coupled with a fluctuating negative sentiment and 'trust' and 'fear' being the most prominent emotions. Through the CL and DA exploration, it was found that the algorithm was portrayed with active agency and blamed also. This is not to say that others were not also found to be blameworthy, like the UK government and Ofqual (plus significant personnel within these organisations), towards the end of the discourse. Ultimately, there was uncertainty as to which social actor received the most blame, but all were presented as blameworthy in the captured discourse. This contributes to understanding of how algorithms can be perceived and portrayed as social actors in online discourse, particularly in high-stakes contexts like education, and highlights the complex dynamics of blame attribution in algorithmic controversies. The combination of analytical approaches yielded a more nuanced understanding of the Twitter discourse surrounding the 2020 A Level algorithm than any of the approaches would have afforded by themselves.

### 7.2.2 *SRQ2: Covid-19 App*

SRQ2, discussed in Chapter 5, focussed on the public discourse surrounding the Covid app. This, once again, involved a comprehensive examination using NLP analysis techniques initially to shed light on topics, sentiment and emotions, followed by using CL and DA to investigate the portrayal of the system's agency within the discourse.

With the NLP analysis, addressing objective **2a**, three primary topics emerged: contact tracing, the government's role and app availability. Notably, Topic 2, which addressed the government's involvement, experienced an upswing in January 2021. Possibly, this had been influenced by heightened governmental concerns during the lockdown period, as the sentiment analysis was generally neutral tone with fluctuations across months. Within those periods, emotions such as 'trust' and 'fear' were prevalent, introducing complexity to the interpretation, as discussed previously by Paucar et al. (2022). This adds a temporal dimension to the understanding of public discourse around technological interventions during the pandemic, revealing how perceptions and discussions evolved in response to changing circumstances and government actions.

For objectives **2b** and **2c**, the active presentation of the app in the discourse was predominant, constituting 96% of instances. This active portrayal employed personalisation, determination, agency metaphors and genericism, aligning with assertions from Samuel et al. (2021) and Morris et al. (2007). This active framing elucidated the diverse roles ascribed to the app —- informing, instructing, permitting, disrupting, functioning or failing -— all of which contributed to shaping public perceptions. This contributes to the understanding of how digital health technologies are personalised in online discourse, foregrounding the potential impact of such portrayals on public perception and acceptance of these tools.

Conversely, passive presentation diminished slightly over time, constituting only 3% of instances overall. Passive constructions often relegated the app to the background, directing attention toward developers/operators or the impacted public, echoing insights from Morris et al. (2007). This shift in linguistic framing reflects evolving perspectives on agency, foregrounding the changing emphasis from the app itself to the broader context of developers, operators and public engagement. This revealed how public attention and attribution of responsibility can evolve in the context of technological interventions,

moving from a focus on the technology itself to the broader system of human actors involved.

The implications drawn from the analysis are noteworthy. The active presentations of the app had discernible implications on its perceived responsibility to process information, aligning with Kretzschmar et al. (2020) and Kent (2020). Even during significant events such as the app's launch, lockdowns and the 'pingdemic', the perception of responsibility remained firmly with the app, as discussed by Dowthwaite et al. (2021) and Pepper et al. (2022). Importantly, these insights extend beyond the app, carrying potential implications for decision-making algorithms more generally, particularly in domains such as healthcare or digital tracing initiatives. The consistent attribution of responsibility to the app, even during major events, indicates a tendency to view technology as an autonomous actor. This finding has broader implications for how society understands and interacts with decision-making algorithms and AI systems, potentially affecting public trust and acceptance.

The portrayal of the app, consistently maintaining perceived responsibility for user welfare and safety, underscores the broader societal expectations placed on technology, accentuating its role not only as a tool but as a responsible and responsive actor in public health endeavours. This observation highlights a shift in societal expectations towards technology, suggesting that advanced systems are increasingly viewed not just as passive tools, but as active, responsible agents in critical domains like public health, which could influence future technology development, deployment and governance strategies to alter expectations.

Users' trust in the app appeared to fluctuate based on several factors. Firstly, trust seemed to be influenced by the perception of the app providing accurate and timely information regarding Covid-19 status and isolation requirements. Seemingly clear and reliable information provision enhanced trust, while instances of unclear or incorrect information eroded it. Secondly, trust was related to the

perceived effectiveness of the app's instructions or directives, such as those regarding self-isolation. When users perceived the app's instructions as reasonable and sensible, trust may have been higher, but questioned or unreasonable directives likely diminished trust. Additionally, users' trust in the app appeared to be linked to its perceived functionality. When the app was seen as functioning properly and effectively carrying out its tasks, trust increased, but instances of malfunction or failure to keep users safe likely decreased trust. Finally, users' perception of the app's agency played a role in trust. If users viewed the app as having a high level of autonomy and making decisions aligned with their needs, trust increased, but flawed decision-making or perceived lack of alignment with user needs likely diminished trust. Overall, trust in the app seemed contingent upon its perceived performance in providing accurate information, issuing sensible instructions, functioning effectively and exhibiting agency aligned with users' expectations and needs.

Users blamed the app in various instances, primarily when they perceived its functionality as faulty or ineffective, such as failing to provide accurate information or fulfil its intended tasks. Additionally, blame was directed at the app when users disagreed with its instructions or found them unreasonable, leading to inconvenience or frustration. Users also blamed the app for causing disruption or inconvenience in their lives, such as through excessive notifications or negative consequences, like isolation or economic impact. Furthermore, blame arose when users perceived the app's decisions or actions as not aligning with their needs or expectations, leading to perceived shortcomings. Predominantly, users blamed the app when they perceived it as failing to fulfil its functions effectively, providing unsatisfactory instructions, causing disruption or exhibiting agency not aligned with their needs or expectations. In this respect, it became clearer that Twitter users tended to attribute blame to the app when they perceived it as having a high level of autonomy and responsibility for its actions, especially if those actions led to negative outcomes

or failure to meet user expectations. Conversely, attributions of blame appeared to be mitigated when specific users perceived the app as lacking agency or being influenced by external factors, such as being monitored by the government, the NHS or Serco.

To address objective **2d**, it was crucial to look at the effectiveness of combining NLP-based tools, CL and DA for this case study in particular. The analysis of topics, sentiment and emotions enabled distinct themes to be revealed, such as contact tracing, government involvement and app functionality. However, the accuracy of topic labelling may have been impacted by limited guidance, potentially influencing result interpretations. Additionally, while sentiment analysis indicated fluctuations in positivity over time, challenges arose in classifying tweets, highlighting the intricacies of accurately gauging sentiment.

Moreover, even though the emotion detection analysis highlighted prevalent emotions like trust and fear, the direction of trust remained ambiguous, posing challenges in interpreting emotional trajectories. This complexity enabled a direct comparison with relevant literature previously discussed, meaning the use of these tools indicated a nuanced relationship between public sentiment and the app's societal impact before beginning the CL and DA investigation. However, uncertainties surrounding the direction of trust underscored the necessity for complementary approaches like CL and DA to achieve a more nuanced understanding.

Using CL and DA to investigate the presentations of the app revealed various active agent categories, including informing, instructing, providing permission, disrupting and functioning. These presentations showcased the app's perceived autonomy and decision-making capabilities, which meant that this analysis foregrounded the reflection of users' perceptions of its role and responsibility. CL also enabled passive constructions to be identified, indicating instances where the app's agency was obscured, potentially deflecting blame from more responsible social actors.

More specifically, to answer SRQ2, the analysis found three primary topics – contact tracing, government role and app availability – with a notable increase in Topic 2 in January 2021, potentially linked to heightened governmental concerns during the lockdown. Sentiment fluctuated slightly above neutral across months and prevalent emotions such as 'trust' and 'fear' varied over time. The CL and DA investigation unearthed that the app was presented actively on 96% of occurrences, using personalisation, determination, agency metaphor and genericism. Further analysis found that the app was presented as either informing, instructing, permitting, disrupting, functioning or failing. The app being presented by Twitter users as a social actor heavily implied responsibility, which resulted in blame when the app was perceived to be underperforming (or performing in a way that the Twitter user deemed undesirable), leaving little trust in the app. The combination of all three approaches meant that a holistic understanding of agency, trust and blame dynamics within the Twitter discourse could be achieved, shedding light on the multifaceted interactions between users, technology and societal expectations in the context of the Covid-19 app.

### 7.2.3 *SRQ3: ChatGPT*

The examination of ChatGPT's role in public discourse was undertaken in Chapter 6 to answer SRQ3. Firstly, this involved using NLP techniques before moving on to examine the system's portrayal as a social actor. Within the NLP analysis, adressing objective **3a**, diverse topics emerged, spanning text generation, chatbot development (Verma and Lerman, 2023), the significance of data, API intricacies, strategies for maximising usage (Kelly, 2023), comparisons and cryptocurrency. Sentiment analysis unveiled predominantly positive tones, with a fluctuating trajectory ranging from 0.21 to 0.31. Notably, a discernible decline in sentiment emerged post-ChatGPT Plus launch

on 1st February 2023, indicative of the discernible impact of algorithmic limitations on user perspectives. Further exploration through emotion analysis exposed the predominant expressions of 'trust' and 'fear,' with a notable erosion of trust observed following the release of ChatGPT Plus potentially linked to heightened concerns about bias and misinformation (Abdullah, Madain, and Jararweh, 2022). Despite pre-existing concerns regarding bias, misinformation and ethical considerations (Abdullah, Madain, and Jararweh, 2022), the manifestation of these issues in public discourse proved less pronounced than anticipated. This indicates that, while the reception was generally positive, public opinion was sensitive to changes in the AI landscape, which could inform strategies for AI development and communication.

Moving into the later phases of analysis, and addressing objectives **3b** and **3c**, around 86% of instances portrayed ChatGPT actively. This active portrayal was characterised by the incorporation of personalisation and agency metaphors, emphasising roles such as content creation, information dissemination and influence (Hassani and Silva, 2023; Kitishat, Al Kayed, and Al-Ajalein, 2020; Rathore, 2023; Razis, Anagnostopoulos, and Saloun, 2016). The dynamic nature of this portrayal, oscillating between a creative actor and an information source, hints at prevalent user uncertainty (Al Lily et al., 2023; Zhou et al., 2023). This contributes to the understanding of how emerging AI technologies are perceived and represented in public discourse, revealing a tendency to attribute agency and influence user expectations of AI.

In contrast, a passive depiction was identified in 14% of instances, framing ChatGPT as a tool, thereby downplaying its agency and impact, relating to previous works that found similar outcomes (Bran et al., 2023; Shijie, Yuxiang, and Qinghua, 2023). The implications drawn from these diverse presentations reverberate in the realm of responsibility perception, a pivotal factor for comprehending the real-world impact of AI (Abdullah, Madain, and Jararweh, 2022; Ferrara, 2023). Notably, Twitter users, leaning on ChatGPT for information, wield in-

fluence in shaping trust concerning AI-generated content (Abdullah, Madain, and Jararweh, 2022). This chapter not only contributes significantly to the understanding of AI's broader social impact (Aljanabi, 2023) but also unveils the concealed agency of ChatGPT within social media discourse. Specifically, the findings suggest that ChatGPT's human-like presence may have influenced Twitter users to present it as a social actor with a clear sense of social agency (Abdullah, Madain, and Jararweh, 2022; Gutiérrez, 2023), although many users appeared to trust ChatGPT, with limited blame occurrence, which was different from the previous two discourses investigated.

Twitter users' trust in ChatGPT appears to have evolved in the analysis. Initially, there was an indication of trust in ChatGPT's ability to produce content effectively, despite occasional mentions of errors or undesired outputs. This trust may stem from the portrayal of ChatGPT as an active communicator and influencer, with users increasingly relying on its outputs and perceiving it as a reliable source of information, despite potential warnings against blind trust. However, the evolving perceptions of ChatGPT's agency and its role as both a creative force and a passive tool saw small fluctuations in trust as the discourse progressed, influenced by inconsistencies in its portrayal and the ethical implications involved in AI deployment. Again, on a small number of occasions, tweets saw ChatGPT blamed for errors or undesirable outcomes, especially as its portrayal fluctuated between being perceived as a creative actor and a passive tool. This blame was associated with instances where ChatGPT's outputs were incorrect, with users attributing responsibility to it due to its perceived agency in decision-making. This analysis aligns with past research by revealing the dynamic and conflicting perceptions of ChatGPT's competence and agency on social media (Al Lily et al., 2023; Bran et al., 2023). It extends the existing literature by providing a temporal dimension to how trust and blame attribution evolve in public discourse, reflecting the broader societal impact noted by Abdullah, Madain, and Jararweh (2022).

The reflections of using NLP, CL and DA to investigate ChatGPT's portrayal in Twitter discourses address objective **3d**. The NLP-based analysis enabled the uncovering of topics ranging from text generation to API functionality. This was a diverse spectrum of user interests and concerns that may have been impossible to analyse without the assistance of the topic modelling software. Moreover, sentiment analysis revealed predominantly positive sentiment towards ChatGPT, fluctuating in response to events such as the launch of ChatGPT Plus, which was crucial as a starting point for further qualitative focus. Emotion detection also uncovered nuanced emotional patterns, highlighting variations in trust, fear and anticipation over time. This impacted the approach taken to the remainder of the analysis, given that these emotions were similar to the previous case studies.

With the CL and DA investigation, trends of active agency, elucidated through transitivity analysis, revealed how ChatGPT was perceived as a social actor in different ways. More specifically, using thematic conventions associated with DA enabled the categorisation of ChatGPT being depicted as creating, informing and influencing content. However, inconsistencies emerged in its portrayal, with shifts between creative agency and passive tool-like descriptions, indicating evolving perceptions (Bran et al., 2023). Using CL to examine passive presentations of ChatGPT provided further insights into users' perceptions of its role. While passive constructions foregrounded human experiences and outcomes, they also downplayed ChatGPT's agency, presenting it as a facilitator rather than an active agent. This enabled the ability to reach conclusions about the ethical considerations that arose from these depictions, including questions about AI responsibility and accountability in decision-making processes.

Conclusively, in answer to SRQ3, the NLP-based analysis uncovered a neutral sentiment with little change over time. Topics related to text generation, chatbot development, the use of ChatGPT as a writing assistant, the importance of data in training the model, the API of ChatGPT, maximising ChatGPT usage, comparisons with other com-

panies and discussions about cryptocurrency. Plus, once again, there was a prominence of trust and fear. Through the CL and DA investigation, ChatGPT was portrayed as a social actor on 86% of occasions, with it mainly generating, presenting information and influencing. Twitter users' trust in ChatGPT evolved somewhat, initially stemming from its portrayal as an active communicator and influencer, despite occasional errors, but fluctuated as perceptions of its agency shifted between being a creative force and a passive tool, occasionally resulting in blame for errors or undesirable outcomes. This contributes to our understanding of online AI perception and reflecting the complex interplay between its perceived agency, capabilities and limitations. By integrating these approaches, more insights were gained, enabling a holistic exploration of trust, blame and agency in the context of ChatGPT.

## 7.3 OVERARCHING RESEARCH QUESTION

The following section brings the detail from the sub-research questions, explored in subsections 7.2.1, 7.2.2 and 7.2.3, to answer the overarching research question of this thesis.

### 7.3.1 *Depictions of Active Agency*

The investigation into the general depictions of algorithmic agency within the public discourse surrounding decision-making algorithms, encompassing the A Level algorithm, Covid-19 app and ChatGPT case studies, has encompassed the examination of intricate language patterns that offer insights into societal perspectives on accountability and responsibility, influencing trust and blame.

In the A Level algorithm case study, the attribution of blame is marked by a dynamic evolution, actively assigned not only to the algorithm but also extending to various social actors. This nuanced

progression suggests an ongoing negotiation within public discourse, resonating with the exploration by Zimmerman (2000) of discourse as a continuous interaction. Furthermore, the findings also relate to the work of Benjamin (2022), whose study adds a socio-political layer and emphasises the complex dimensions intertwined with algorithmic decision-making and how societal views on accountability evolve. This overt portrayal of blame, guided by the algorithm appearing as a social actor, points towards a communal negotiation process, highlighting the societal dynamics shaping perspectives on responsibility (Van Leeuwen, 2008).

Within the discourse surrounding the Covid-19 app, the overwhelming predominance of active portrayal (96%) paints a clear picture of societal perceptions regarding the app's responsibilities, especially during critical events. The work by Morris et al. (2007) on metaphors in agency enriches the understanding of this active portrayal, showcasing the multifaceted roles the app assumes during significant events. This aligns with societal expectations that decision-making algorithms, exemplified by a public health app in this case, bear a substantial responsibility for disseminating accurate information during crises.

ChatGPT's active agency, encompassing roles in content creation, dissemination and influence, showcases a dynamic oscillation between creativity and being an information source. The insights by Bucher (2017) into algorithmic roles further illuminate this adaptability, emphasising the algorithm's versatility in assuming varied roles within social discourse. This flexibility in portrayal aligns with the evolving nature of human-AI interactions, where ChatGPT is perceived as an active participant in the creation and dissemination of information.

The recurring pattern of actively depicting social actors, noted in various studies (Bryson, 2020; Gallagher, 2000; Leslie, 1993), underscores the significance of understanding perceived roles and responsibilities in decision-making algorithms. The conceptual framework

provided by Bryson (2020) regarding the degree of autonomy and the exploration of the interplay between agency and responsibility by Wallace (1998) and Oktar (2001) further contribute to interpreting these depictions. For example, it was clear in the ChatGPT case study that the degree of autonomy that the system was depicted as having, through the tweets, fluctuated on a spectrum from information resource to powerful influencer. Additionally, the emphasis on the reflection of human design choices by Floridi et al. (2018) also underscores the interconnectedness of technological design and societal expectations, as demonstrated in the analysis of these collective Twitter discourses. For example, there is a small, yet defined emphasis on the involvement of the UK government, the NHS and Ofqual in the design and deployment of the A Level algorithm and the Covid-19 app and all are foregrounded and excluded as appropriate. However, a similar focus on OpenAI in the ChatGPT discourse is not found, perhaps due to the system performing as expected and users expressing a greater level of trust.

Despite these shared patterns, the extent of responsibility varies among the systems, influenced by factors such as the decision-making process and potential biases. This relates the exploration by Archer (2000) of being and becoming, which adds a philosophical dimension to the understanding of responsibility. This could show the dynamic nature of how entities, including algorithms, navigate their perceived roles in society. Additionally, this also links to the work of Miller (2001), which contributes to the nuanced discussion by foregrounding the need for accountability frameworks to address potential biases inherent in decision-making algorithms (Diakopoulos, 2016). These variations in responsibility attribution underscore the complexity of public perception and ethical considerations surrounding decision-making algorithms, prompting a call for context-specific approaches to accountability in the ever-evolving landscape of human-AI interactions.

### 7.3.2 *Passivisation*

Examining passive presentations across the A Level algorithm, Covid app and ChatGPT case studies reveals a nuanced interplay between decision-making algorithms and their portrayal in Twitter discourses. Passive constructions, occasionally relegating the technology to the background, surfaced across all systems, although remained a relatively small proportion of the overall tweets. This strategic backgrounding aligns with linguistic principles discussed by Comrie (1977) and, more specifically related to the theme of this thesis, the exploration of design-focused communication by Holford (2022). The deliberate emphasis on other responsible parties or the impact on users and stakeholders serves as a linguistic strategy to shape the narrative surrounding these algorithms. This approach highlights a broader trend in discourse analysis, where passive constructions play a role in framing the narrative by foregrounding certain elements while downplaying others.

Passive constructions within these discussions also attempted to obscure or shift blame, employing techniques like backgrounding and exclusion, as outlined by Clark (1998). By doing so, the direct accountability of the technology becomes diluted, leading to a more complex landscape of trust and blame attribution (Garfinkel et al., 2017). This linguistic manoeuvring reflects the intricate dance between responsibility and obfuscation within the public discourse surrounding algorithmic decision-making.

This is not to say that it was only the decision-making algorithms that were backgrounded, excluded or presented passively: in fact, all three discourses showcased that those behind the development of these systems were also found to be included in similar structures. The intentional use of passive constructions complicates straightforward attribution of responsibility, shaping public perception in a way that may obscure the true agency behind these technologies: in this case Ofqual, the UK government, Serco or OpenAI.

Moreover, passive presentations facilitated a shift in the focus of blame. Rather than directly attributing blame to the technology itself, passive constructions directed attention toward external entities or impacted individuals. This shift in focus downplayed the direct agency of the system and redirects blame to those in proximity or affected by its decisions (Feier, Gogoll, and Uhl, 2021). The observation of passive constructions may indicate a nuanced manipulation of blame attribution dynamics. By doing so, the technology becomes a passive player in the narrative, highlighting a rhetorical strategy employed to shape public discourse and perceptions of accountability.

In essence, the analysis of passive presentations within the discourse surrounding decision-making algorithms underscores the complex strategies at play in the investigation into social media discourses concerned with these systems. These linguistic choices, as observed in passive constructions, go beyond mere grammatical preferences. Instead, they might be seen as powerful tools in framing the discourse around agency, trust and blame (which will be discussed next), perhaps revealing unconsciously how Twitter users perceive these systems. Alternatively, the deliberate use of passive constructions may also reflect a more conscious effort to influence the online narrative surrounding these technologies, contributing to discourses that decision-making algorithms do not hold the capability of a human social actor.

### 7.3.3 *Trust*

The exploration of trust within the discourse surrounding decision-making algorithms, as observed in the A Level algorithm, Covid-19 app and ChatGPT case studies, unveils a multifaceted narrative deeply entwined with the dynamics of agency, responsibility and accountability. At its core, the trust users place in these systems is intricately connected to the interplay between the algorithms and their

interactors, a perspective aligned with works by Muir (1987) and Madhavan and Wiegmann (2007). This relational lens underscores the pivotal role of agency and responsibility in shaping the level of trust that users invest in decision-making algorithms.

A recurrent theme across all three systems is the discernible fluctuations in trust perceptions over time, a phenomenon that finds resonance in the extensive literature by Mayer, Davis, and Schoorman (1995). These variations in trust align with key events or system releases, reflecting the nuanced changes in sentiment and emotion detection (Bonnefon, Shariff, and Rahwan, 2016). The ebb and flow of trust underscores the dynamic nature of potential public feeling, showcasing its responsiveness to external factors and its impact on the evolving narrative surrounding decision-making algorithms. Importantly, users' willingness to trust automated systems appears to be influenced by these fluctuations, echoing broader discussions on the social acceptance of AI technologies.

The attribution of responsibility emerges as a critical factor influencing trust perceptions, relating to the emphasis Shahrdad and Amirani (2018) put on the ubiquity of systems. When decision-making algorithms are portrayed as proactive social actors, they tend to bear a higher burden of responsibility, particularly during significant events, highlighting their profound impact on user welfare (Bonnefon, Shariff, and Rahwan, 2016). This perceived responsibility aligns closely with the portrayal of active agency, emphasising the reciprocal relationship between how agency is represented and the trust users place in decision-making algorithms.

Further influencing trust dynamics is the variability in the presentation of decision-making algorithms, whether depicted as active social actors or passive tools (Lyons et al., 2017). Active portrayals, which accentuate the role and influence of these systems, contribute to a heightened perception of their capabilities, thereby impacting trust. On the flip side, passive depictions that downplay the agency of these systems have the potential to diminish their perceived impact, in-

fluencing trust dynamics (Alaieri and Vellino, 2016). These nuanced choices in presentation underscore the malleability of trust and its sensitivity to how decision-making algorithms are portrayed in public discourse.

Within the domain of AI-generated content, users increasingly rely on systems like ChatGPT for information and content generation (Bonnefon, Shariff, and Rahwan, 2016). However, the trust placed in these systems is not impervious to fluctuations. Events, particularly those triggering concerns about bias, misinformation and post-launch limitations, can significantly sway user trust. Notably, trust appears to be lower for the two older systems, potentially attributed to the lack of optional use compared to the more recent ChatGPT, exemplifying the nuanced nature of trust within the temporal evolution of AI systems.

In conclusion, the exploration of trust within the online discourse surrounding decision-making algorithms unravels a complex tapestry woven by agency, responsibility and presentation dynamics. The fluctuating nature of trust, its reliance on the attribution of responsibility and the impact of varying presentations collectively contribute to the intricate landscape of public sentiment towards these systems. Trust in decision-making algorithms is not static but a dynamic construct, influenced by multifaceted factors that shape and reshape the evolving narrative surrounding AI technologies.

### 7.3.4 *Blame*

Exploring the attribution of blame within the discourse surrounding decision-making algorithms unveils a complex interplay of agency, responsibility and accountability. Blame, often realised through the lenses of agency, responsibility and accountability, emerges as a central theme across all three systems under scrutiny. Active attributions of blame were evident across the board, with the intensity of blame

more pronounced for older systems like A Level algorithm and the Covid app, compared to the relatively lesser blame directed towards ChatGPT, linking to the ideas of Burrell (2016) and Taddeo and Floridi (2018). This highlighted that issues of opacity and the perceived unfulfilment of ethical obligations may lead to blurred agency and transparency, potentially perpetuating bias and evoking blame. This disparity underscores the evolving nature of public perceptions regarding the responsibility of decision-making algorithms, with newer systems potentially benefiting from less entrenched blame narratives.

The presentation of blame within the discourse exhibits dynamism, characterised by the occasional use of passive constructions to background the technology and highlight other responsible parties or the impact on users (Jobin, Ienca, and Vayena, 2019). This is not to say, however, that passivisation does not eradicate blame entirely from the system or other backgrounded entity. This dynamic presentation of blame, often linked with perceived ethics, underscores the nuanced nature of blame attribution within the public discourse surrounding decision-making algorithms. Moreover, the fulfilment of ethical responsibilities varied among the systems, with ChatGPT appearing to fulfil these responsibilities to a greater extent compared to the others (Whittlestone et al., 2019).

While all systems faced some blame attribution, the nature of blame differed significantly. Ofqual and the UK government were predominantly attributed blame for algorithmic decisions, while blame for the Covid app was directed towards its role in informing, instructing or disrupting (Mittelstadt et al., 2016). This nuanced differentiation in blame attribution reflects the contextual intricacies surrounding the perceived roles and responsibilities of decision-making algorithms within different domains.

Perceived responsibility also varied among the systems, with the Covid app bearing a higher burden of responsibility, especially during crucial events such as public health emergencies (Burrell, 2016). Conversely, for ChatGPT, varying presentations influenced responsi-

bility perceptions, with many presentations absolving the algorithm of blame, leading to what is termed as a 'responsibility gap' (Mittelstadt et al., 2016). This disparity in perceived responsibility highlights the nuanced interplay between public discourse, technological design and societal expectations regarding accountability.

The impact of blame attribution within public discourses was particularly pronounced for A Level algorithm and the Covid-19 app, given the real-world consequences of their decisions, impacting students, schools or public health outcomes (Jobin, Ienca, and Vayena, 2019). These entities became subject to blame within the discourse, underscoring the significant societal implications of decision-making algorithms in critical domains. In contrast, while ChatGPT faced concerns about bias and ethics, these issues were addressed on a smaller scale than anticipated, possibly due to the nascent stage of discourse (Whittlestone et al., 2019). This divergence in the scale of impact underscores the evolving nature of public discourse surrounding newer AI technologies and the complexities inherent in attributing blame within this rapidly evolving landscape.

### 7.3.5 *Section Summary*

Overall, this section has detailed the discourse surrounding decision-making algorithms, as exemplified by the A Level algorithm, Covid-19 app and ChatGPT. The analysis here has revealed a dynamic interplay of agency, trust and blame attribution. In examining the depictions of active agency, the portrayal of these algorithms as active social actors influences societal perspectives on accountability, with blame evolving and influenced by socio-political dimensions. In subsection 7.3.1, all three of these case studies showcased consistent patterns of the systems being presented actively, contributing to their status as social actors. From the initial NLP-based analysis, it was possible to uncover surface-level insights into all three systems, espe-

cially when it came to the topics involved, the sentiment trajectories and the emotions detected. Twitter users discussed their roles and responsibilities and attributed active agency to the three algorithms, as exemplified in each of the CL and DA sections. However, there appeared to be variation in how each case study was presented. For example, the A Level algorithm was found to have blame attributed actively to the algorithm (plus various social actors), which intensified towards the discourse's end, suggesting an evolving stance on accountability. Similarly, the predominant portrayal of active agency in the discourse surrounding the Covid-19 app highlighted societal expectations of the app's responsibilities during critical events, while ChatGPT's portrayal was more versatile.

Conversely, the analysis of passive presentations across these case studies, seen in subsection 7.3.2, revealed a strategic backgrounding of decision-making algorithms, perhaps aimed at shaping the narrative surrounding their responsibilities. Although small in comparison to the active presentations, passive constructions within discourse obscured and shifted blame but also complicated straightforward attribution of responsibility, highlighting the nuanced manipulation of blame attribution dynamics. In all three case studies, backgrounding and exclusion techniques saw responsibility either removed from the system, or implicating the system by excluding those responsible for the development or deployment of the public-facing algorithm.

Subsection 7.3.3 explored the overall research findings regarding trust through the analysis of agency, responsibility and accountability. These factors shaped trust perceptions of Twitter users towards these systems over time, with fluctuations influenced by key events or releases. The decision-making algorithms, all depicted as active social actors, tended to bear more responsibility, especially during significant events. However, there was variability in the baseline trust in the discourses – with Twitter users exhibiting much more trust towards ChatGPT, through the grammatical portrayal of it as a trustworthy social actor, in comparison to the A Level algorithm and the NHS

Covid-19 app. Twitter users relied on AI systems like ChatGPT for information, but trust fluctuated based on events, particularly with concerns related to bias and misinformation, yet the two other case studies saw lower trust, potentially due to mandatory usage.

In examining blame attribution, in section 7.3.4, all three of the decision-making algorithms were presented as blameworthy due to grammatical structures seen in tweets. The two older systems received more intense blame compared to ChatGPT, which appeared to display better fulfilment of ethical responsibilities than other systems. Blame varied among systems, directed at governmental bodies for algorithmic decisions and at the Covid-19 app for its role in informing or disrupting. The Covid-19 app bore significant responsibility during critical events, while varying presentations affected responsibility and blame perceptions for ChatGPT, often absolving the algorithm. Additionally, it was important to note that the A Level algorithm and Covid-19 app faced extensive public scrutiny due to real-world consequences, impacting students, schools or public health, while ChatGPT faced less scrutiny.

This has culminated in the answering of the overarching research question of this thesis, which was that the Twitter discourses around decision-making algorithms tend to ascribe social agency, along with various degrees of trust and blame, to said systems via active grammatical structures. All three decision-making algorithms were portrayed as distinct social actors. While some patterns – such as the ascription of agency to the systems – were universal, the impact on trust and blame in the three discourses was varied, with ChatGPT being presented as more trustworthy and certainly less blameworthy. However, the other two case studies showcased how the agency that was ascribed to the systems from Twitter users unveiled a more overt degree of accountability and responsibility, resulting in decreased trust and clearer blame. Using the three computational and discursive linguistic approaches enabled a more holistic analysis and exploration than using one approach in isolation.

## 7.4 IMPLICATIONS FOR RESEARCH

This section will detail the most important implications for research that arise from this thesis. The implications drawn from the analysis of the roles and contexts of decision-making algorithms – embodied by the A Level algorithm, the NHS Covid-19 app and ChatGPT – underscore the diverse landscape in which these systems operate. The A Level algorithm, for instance, was intricately associated with flaws, suitability and its impact on students, schools and teachers. In contrast, the NHS Covid-19 app's discourse revolved around contact tracing, government involvement and app availability, reflecting its role in public health management. The Twitter discourse on ChatGPT, on the other hand, was linked to text generation, content creation, information dissemination and influence, highlighting its versatility and broad scope of application. The portrayal of these systems as social actors significantly influences user trust in AI-generated content (Engeström et al., 1999). Varying presentations of these systems impact perceptions of responsibility, thereby influencing trust and blame attributions in social media discourse. This reflection of human design choices emphasises the importance of understanding the nuanced interplay between technological design and user perceptions. As found in this thesis, significantly fewer tweets attribute agency to the developers and promoters of these systems, which does not reflect the recommendations by Floridi et al. (2018), who believe that human designers ultimately have agency and control over algorithmic decisions and outputs.

The implications extend beyond the social media discourses examined here; these findings have the ability to shape real-world impacts concerning responsibility and trust in decision-making algorithms (Burrell, 2016; Pasquale, 2015). The impact of trust, blame and responsibility attributions transcends mere discourse, influencing public perception, government decisions and potentially affecting future technological developments (Gillespie, 2014; Zarsky, 2016). Under-

standing these perceptions becomes paramount for developers, policymakers and promoters to enhance user experiences and maintain trust in decision-making algorithms and the content or decisions they produce (Mittelstadt et al., 2016; Selbst et al., 2019). Incorporating broader perspectives into system design, as advocated by Diakopoulos (2016) and Selbst et al. (2019), becomes essential in addressing the multifaceted implications arising from the public discourse surrounding decision-making algorithms (Burrell, 2016).

The utilisation of Twitter as a platform for commenting on these systems offers both opportunities and challenges. While commonplace and conducive to certain types of analysis (Bernard, 2018; Fadanelli, Dal Pozzo, and Fin, 2020; McGlashan, 2020), the veracity of comments made on Twitter can sometimes be questionable. However, despite potential inaccuracies, the platform serves as a springboard for discussion and debate, shedding light on public sentiments and concerns. Nevertheless, it remains uncertain to what extent Twitter users interacted with these systems directly, underscoring the need for further research to elucidate the nature and depth of such interactions and their implications on public discourse and technological development.

The complex interplay between social media discourse, public perception and real-world impact underscores the need for a comprehensive approach to research in this domain. By examining the roles and contexts of decision-making algorithms across different systems and platforms, this research has enabled those in the development and promotion of decision-making algorithms (and related systems) to gain deeper insights into the factors shaping user trust, responsibility attributions, blame and the broader societal implications of these technologies. This research has indicated that Twitter users have been shown to treat these systems as social actors in their own right, to various degrees of severity, implying that these algorithms are viewed as such beyond these discourses. These insights may be essential for informing policy decisions, guiding technological development and

fostering public trust in the increasingly pervasive role of decision-making algorithms in our lives.

Furthermore, the implications for research extend to the ethical considerations surrounding the design and deployment of decision-making algorithms. As emphasised by Mittelstadt et al. (2016) and Selbst et al. (2019), incorporating ethical principles and considerations into system design is essential for ensuring fairness, accountability and transparency. By addressing ethical concerns and considering broader societal implications, researchers can contribute to the development of decision-making algorithms that are not only effective but also socially responsible and ethically sound.

## 7.5 LIMITATIONS

The thesis, while providing valuable insights into the discourse surrounding decision-making algorithms, is not without its limitations, which warrant acknowledgement and discussion.

Methodological constraints posed challenges. Within NLP, grappling with the intricacies of language, such as sarcasm, negation and subtle nuances, presented a formidable obstacle, potentially skewing sentiment and emotion analysis. Moreover, the automated labelling of topics, sentiments and emotions lacked the requisite nuance and context, engendering discrepancies and inaccuracies in interpretation. Furthermore, contextual comprehension posed a perennial challenge for NLP tools, particularly when tweets lack explicit context, potentially impeding accurate classification and analysis.

Similarly, the approaches of CL and DA are susceptible to subjective interpretation, thereby risking potential biases or divergent conclusions among researchers (Baker, 2006). The multifaceted nature of the discourses analysed further complicated matters, as encapsulating the myriad linguistic and social elements comprehensively became an arduous task (Fairclough, 1993). It is important to note that

the analysis was conducted by a single researcher, in collaboration with a supervision team, which will likely have limited the range of perspectives considered. Despite efforts to mitigate subjectivity through combining DA with other approaches, the inherent interpretative nature of DA remains a factor to consider when evaluating the findings. Although some mitigation may have occurred, there will be many biases that remained. Acknowledging this potential for subjectivity is crucial for transparency. This even extends to the inclusion and exclusion of certain data. For example, another researcher may have taken the decision to include retweets as well as quote tweets in the analysis, as this tends to indicate support, even if the words are not directly from an individual. Additionally, the specificity of the corpora employed in CL and DA studies restricts their generalisability as they may not faithfully represent broader language usage patterns.

Integrating NLP, CL and DA approaches introduced another layer of complexity, while investigating some of shortcomings of individual approaches, which potentially compromised the coherence and consistency of analysis. Despite concerted efforts to amalgamate these approaches, the reliance on automated tools still runs the risk of overlooking nuanced language intricacies, thereby curtailing the ability of the thesis to fully capture the richness of social discourses.

Beyond methodological confines, broader limitations impinge upon the scope and implications of the thesis. The quality, quantity and representativeness of the datasets utilised may constrain the robustness of the conclusions drawn. Furthermore, examining discourse over specific periods might inadvertently curtail the understanding of evolving opinions or shifts in social narratives over time. Moreover, the broader socio-political context, cultural nuances and external events exert a significant influence on discourse, complicating efforts to isolate effects solely attributable to the analysed systems (Meyerson, Weick, Kramer, et al., 1996). Despite attempts to mitigate these factors, concerns remain regarding how well the thesis

encapsulated the full spectrum of social media discourses surrounding decision-making algorithms.

It is very important to note that this thesis does not claim that all people view public-facing decision-making algorithms in these ways. Rather, it is right to be explicit that the findings here contain a relatively small insight into specific discourses. While this does have the potential to reflect broader views, these results should realistically be discussed in their own specific contexts and used by those developing and promoting said systems as a reference point.

Another limitation relates to the notion of masculine agencies, which has formed the default agency in this thesis. In particular, the agency that is given to ChatGPT, for example, is associated with masculine qualities – such as being assertive, powerful and dominant. There has been little exploration – from a theoretical or practical angle – of feminine agencies in this thesis. Overlooking something such as this may have allowed for biases in individual interpretations to occur.

## 7.6 CHAPTER SUMMARY

Overall, this discussion chapter aimed to analyse and interpret the findings from Chapters 4, 5 and 6 to answer the sub-research questions and the overarching research question of this PhD thesis. Alongside this, it aimed to provide insights into the significance of the findings and address some of the limitations of the work. In section 7.2, the sub-research questions of the thesis were addressed individually. Firstly, the analysis of Twitter discourse surrounding the 2020 A Level algorithm revealed users predominantly holding the algorithm accountable as a social actor, aided by nuanced blame attribution techniques. Secondly, the examination of the NHS Covid-19 app discourse uncovered diverse topics and predominantly neutral sentiment, with trust and blame fluctuating based on app performance,

which showcased societal expectations of technology in public health. Finally, ChatGPT's portrayal in Twitter discourses was explored, revealing diverse topics and predominantly positive sentiment. ChatGPT was primarily presented as a social actor that created content, provided information and influenced users, which had a positive impact on trust based on its perceived agency and performance. This exploration answered each of the sub-research questions and addressed the individual study objectives of the thesis.

Next, section 7.3 provided a comprehensive overview of the discourse surrounding decision-making algorithms as a collective. All three systems were depicted as social actors with a strong sense of social agency (realised through grammatical agency); however, the implications of trust and blame were varied. Trust perceptions varied across the case studies, with ChatGPT generally receiving more trust. This was also reflected in the examination of blame as the A Level algorithm and Covid-19 app faced more intense scrutiny due to the levels of responsibility and accountability demonstrated. While active agency was predominant, passive presentations also shaped perceptions, backgrounding and excluding other actors within the discourses and, mainly, foregrounding the algorithms. Overall, this section concluded that, by using NLP, CL and DA as approaches to investigate, Twitter users tended to attribute social agency to decision-making algorithms, although the extent of trusting and blaming differed depending on the system. This underscored the importance of employing multiple analytical approaches for a more holistic understanding.

The implications of these findings on research were discussed in section 7.4. The fact that the system depicted distinct roles and contexts was important: the A Level algorithm was linked to flaws, suitability and its impact on students, schools and teachers; the NHS Covid-19 app focused on contact tracing, government involvement and app availability; while ChatGPT was associated with text generation, content creation, information dissemination and influence. With

these systems being depicted as social actors in Twitter discourses, this may reflect or impact trust – or blame – in AI-generated content and influence perceptions of responsibility and accountability. Such perceptions shaped through social media discourse can have significant real-world implications, influencing public perception, government decisions and future technological developments. Understanding these perceptions is crucial for developers, policymakers and promoters to enhance user experiences, maintain trust in decision-making algorithms and break down barriers, perhaps – most notably – via addressing algorithmic agency overtly. It was also important to note that, while using Twitter to comment on systems is commonplace and facilitates analysis, the level of interaction between Twitter users and these systems remained unknown.

Finally, section 7.5 acknowledged the limitations encountered during the work of this thesis. Methodological constraints within NLP, CL and DA, including language complexities and subjective interpretation, posed challenges, potentially affecting the coherence and consistency of the analysis. Moreover, the specificity of corpora and reliance on automated tools might have restricted the comprehensive capture of Twitter discourses. Broader limitations encompassed how representative the dataset was and the influence of socio-political context on discourse dynamics.

Overall, this chapter concludes by stating that the research question of this thesis, concerning the discourse surrounding decision-making algorithms on Twitter, has been addressed. Through using NLP, CL and DA, it was revealed that Twitter users tended to attribute social agency to these algorithms, portraying them as active social actors. While this portrayal was consistent across the studied algorithms, the impact on trust and blame varied. ChatGPT generally received more trust and less blame compared to the other two systems, suggesting a nuanced relationship between agency attribution and public perception that may be affected by context. Put simply, not all decision-making algorithms are perceived the same way. This

research adds valuable insights to the existing body of knowledge in the field of decision-making algorithms and social discourse by shedding light on the dynamics of agency, trust and blame attribution. Furthermore, it highlights the diverse roles and contexts attributed to different types of algorithms, providing a deeper understanding of their societal implications and the factors shaping user perceptions, which could inform more responsible promotion of these systems. Finally, although not the main aim of this thesis, it has also identified how limitations in NLP, CL and DA can be overcome by using them in a combined manner to investigate social media discourses.

# 8

# CONCLUSION

This chapter serves as the summary of this thesis, representing this research's culmination. Firstly, section 8.1 will provide a comprehensive overview of the PhD investigation as a whole, before detailing its key findings and main contributions in section 8.2. Following this, considering the limitations of the thesis seen in section 7.5, ideas for future work will be proposed in section 8.3.

## 8.1 THESIS SUMMARY

This thesis began in Chapter 1 by introducing the background and motivation of exploring the dynamics of agency, trust and blame in public-facing decision-making algorithms using an interdisciplinary lens. This chapter outlined the issue of recent concerns about decision-making algorithms, particularly when negative consequences arise, highlighting challenges in determining responsibility due to complexity and opacity. Moreover, this chapter introduced the focus on the three case studies and their relevance to the concern of the perceived social and algorithmic agency and the implications for trust and blame. This presented the research gap of little exploration currently existing on how trust and blame of these systems are impacted by perceived social agency, responsibility and accountability. To examine this, the relationship between grammatical and social agency was investigated using NLP, CL and DA as the combined approach to analyse Twitter discourses relating to the three systems. By integrating these approaches, insights into trust, blame and agency attribution in decision-making algorithm discourse on Twitter could be achieved.

The literature review in Chapter 2 explored how grammatical agency mirrors social agency within decision-making algorithms, touching on control, responsibility and social interactions. It introduced SAR for analysing grammatical structures, noting passive constructions and lexical choices as influential factors. Furthermore, this chapter discussed how agency, responsibility and accountability relate to debates on algorithmic autonomy and transparency challenges. In particular, trust's pivotal role in algorithmic acceptance, shaped by experience and context, was emphasised. Indeed, challenges in transparency and ethical principles persist, demanding ongoing research to bolster trust. Conversely, blame assignment involves a nuanced interplay of agency, responsibility and accountability, stressing fairness and transparency. Therefore, examining the case studies – the 2020 A Level algorithm, the NHS Covid-19 app and ChatGPT – in greater detail underscored their societal impact, ethical intricacies and user acceptance challenges. In this sense, these cases highlighted the need for trust, transparency and ethical considerations in algorithmic systems, calling for further exploration of societal perceptions online. Overall, this chapter revealed the complexities of agency, responsibility, trust and blame within decision-making algorithms and foregrounded that the portrayal of these systems' social agency on platforms like Twitter remained largely unexplored.

The following chapter, Chapter 3, introduced the analytical framework adopted in this thesis. This began with data collection using the Twitter for Academic Purposes API, which emphasised ethical considerations and data anonymisation practices. It then detailed the integration of various analytical approaches into the overarching methodology. NLP-based tools captured initial data trajectories, followed by CL analysis and DA, offering insights into grammatical structures, social agency and power dynamics within the three chosen discourses. Subsequent sections outlined the approaches' strengths and limitations but, more importantly, highlighted that they collectively enabled the analysis of large datasets, the uncovering of language patterns

and the exploration of social implications within texts. The chapter proposed integrating these approaches for robust analysis, combining computational tools' strengths with the interpretative depth of DA. Details on the overarching implementation of each approach were also documented, while specifics for each case study could be found in their individual chapters.

Chapter 4 explored the first of the three case studies: the 2020 A Level grade calculation algorithm. In summary, the NLP analysis identified key topics in Twitter discussions, focusing on government involvement and algorithmic flaws. Sentiment analysis showed a mixed trend, with challenges in categorising tweets. Even though emotion analysis detected predominant feelings of 'trust' and 'fear', interpreting 'trust' tweets, in particular, remained uncertain. CL and DA findings showed Twitter users blaming the algorithm for A Level results, with blame extending to government figures and regulatory bodies like Ofqual. Passive constructions obscured direct accountability, with blame increasingly directed at specific individuals over time. Overall, this chapter contributes a detailed analysis of Twitter discourse on the A Level results controversy, highlighting blame attribution dynamics and demonstrating the complementary role of CL and DA alongside NLP-based tools.

Chapter 5 focused on the case study of the NHS Covid-19 contact-tracing app. Firstly, the NLP analysis identified three main topics in Twitter discussions: contact tracing and isolation, government involvement and data management, and app download and availability. Tweet sentiment ranged from slightly positive to slightly negative, fluctuating over time, with dips during key events. Moreover, emotion detection showed fluctuating levels of 'trust', 'fear' and 'anticipation', with trust direction again unclear. Further analysis showed that Twitter users presented the app as a social actor with clear agency. The app was predominantly portrayed actively, implying agency through techniques like personalisation and determination. Twitter users perceived the app as responsible for their welfare, espe-

cially during significant events like its launch, the second lockdown and the 'pingdemic'. The study contributes to understanding the social impact of the NHS Covid-19 App, highlighting its perceived responsibility for user welfare and, therefore, blame. This study offered insights into public responses to decision-making algorithm interventions, particularly in healthcare, whilst again demonstrating the strengths of combining NLP, CL and DA approaches.

Chapter 6 analysed the ChatGPT discourse collected. Topics included consistent trends such as text generation and chatbot development, alongside fluctuating topics such as cryptocurrency discussions. Sentiment was mostly positive, though it dipped around the launch of ChatGPT Plus, potentially due to user frustration. Emotion detection highlighted dominant 'trust' and 'fear' patterns, perhaps reflecting concerns about biases and misinformation. Furthermore, Twitter users predominantly depicted ChatGPT as a social actor, mainly active in approximately 86% of cases, emphasising its roles in content creation and influence. Again, by combining NLP, CL and DA, this study offered valuable insights into the social impact of ChatGPT. In fact, these findings have implications for AI developers, like enhancing transparency, educating users about its limitations, as well as policymakers, who could ensure accountability and promote informed use. Overall, this provided insight into public perceptions and expectations of ChatGPT regarding trust and responsibility.

The discussion in Chapter 7 consolidated the key findings from the thesis chapters on the 2020 A Level algorithm, the NHS Covid-19 app and ChatGPT. The overarching analysis revealed patterns across the case studies: Twitter users predominantly attributed social agency to the decision-making algorithms in question, although trust and blame varied among the systems. ChatGPT generally garnered more trust and less blame compared to the A Level algorithm and the Covid-19 app, and vice-versa. The implications of these findings were discussed, highlighting the importance of understanding user perceptions of algorithmic agency, trust and blame. Such perceptions

can influence public opinion, government policies and technological developments. Recognising the diverse roles and contexts attributed to different algorithms is crucial for promoting responsible use and fostering trust. Also, the chapter acknowledges the methodological constraints in using NLP, CL and DA and the potential biases inherent in Twitter data. Moreover, this section explicitly states that the findings from this thesis are not broadly representative of the views of the wider public, but an insight into specific discourses that could be used to support those who develop and advertise these systems to consider alternate strategies for promotion and adoption.

## 8.2 MAIN CONTRIBUTIONS

As outlined at the start of this thesis, this PhD project intended to accomplish the following objectives:

a Demonstrate how sentiment analysis, topic modelling and emotion detection provide insight into public discourses surrounding decision-making algorithms.

b Demonstrate how corpus linguistics, particularly collocation, provides insight into public discourses surrounding the agency of decision-making algorithms.

c Demonstrate how Discourse Analysis provides insight into public discourses surrounding the agency, trust and blame of decision-making algorithms.

d Identify the strengths and limitations of using the three approaches to investigate public discourses surrounding decision-making algorithms.

In turn, the accomplishment of these objectives has culminated in the achievement of the primary contribution of the thesis, which was the insights into views expressed about decision-making algorithms, with particular regard to their agency and implications for

trust and blame, on social media. Through the examination of the three case studies, this research analysed the views expressed about public-facing decision-making algorithms on Twitter. By scrutinising the discourse surrounding these algorithms, the thesis has shed light on the attribution of agency, responsibility and accountability within these systems, thereby elucidating implications for trust and blame. This contribution extended the broader discourse on the societal implications of decision-making technologies by examining how views are expressed about these systems on social media and the influence that this may have on shaping broader attitudes and perceptions. This contribution was achieved through the analysis of Twitter discourses using NLP-based tools, CL and DA. Therefore, this research has offered a comprehensive understanding of linguistic patterns and communicative dynamics, as well as facilitating the identification of nuanced perspectives. Ultimately, through the incorporation of real-time, user-generated content, this thesis makes a considerable contribution to research by reporting on the portrayal of these systems and illuminating misconceptions and barriers to use or adoption.

Additionally, a subsidiary contribution of this thesis lies in its approach to language analysis, which integrated and complemented existing techniques. By combining NLP, CL and DA, this interdisciplinary lens not only enhanced the findings presented in this PhD thesis but also held the potential for broader application by other researchers seeking to study complex phenomena on social media platforms. This contribution was realised through the synthesis of the diverse approaches, which enabled a more comprehensive exploration of language use and interaction patterns in online discourse. By bridging the gap between these existing analyses, the research has contributed to methodological innovation within the field of social media studies and HCI. Moreover, it is also possible to say that this approach has facilitated a better understanding of the complexities inherent in online communication and enabled researchers to uncover

subtle nuances in language use and discourse dynamics, despite subjectivities being present in the analysis process.

Whilst not a core contribution of the thesis, the findings detailed in this research may have proven beneficial for policymakers, developers and promoters of decision-making algorithms, too. In the future, insights from this analysis may inform strategies to overcome barriers to adoption, enhance user experiences and foster responsible development and deployment of these systems. By highlighting the societal implications of decision-making algorithms, this research has aimed to contribute to the ongoing dialogue surrounding the ethical, social and technical dimensions of algorithmic decision-making by raising awareness of the societal implications of such systems and fostering dialogue among stakeholders.

## 8.3 FUTURE WORK

Following on from the limitations of the thesis that were outlined in section 7.5, several avenues of future work can be pursued from the foundations set in this PhD thesis.

Firstly, exploring other social media platforms beyond Twitter could provide a more comprehensive understanding of how the discourse surrounding decision-making algorithms varies across different online spaces. Platforms such as Facebook or Reddit may present unique nuances in user interactions and perceptions. Additionally, more longitudinal studies tracking changes in public discourse over time could offer more comprehensive insights into the evolution of perceptions towards public-facing decision-making algorithms. By analysing discourse at regular intervals, researchers could capture shifting trends and sentiments.

Future work stemming from the methodological constraints identified in this thesis could focus on refining and enhancing existing approaches in NLP, CL and DA. Specifically, research efforts could be

directed towards improving the contextual comprehension of NLP tools by combining these with CL and DA, particularly in situations where tweets lack explicit context, to enhance the accuracy of classification and analysis. To add to this, efforts to address the limitations of corpus specificity in CL and DA studies could include expanding the range and diversity of datasets used, thereby enhancing the generalisability of findings to broader language usage patterns.

Furthermore, this thesis sets the groundwork for integrating NLP, CL and DA approaches in future research. The development of this would involve exploring innovative methodologies that leverage the strengths of each approach, while mitigating their respective shortcomings. This could include refining automated tools to better capture nuanced language intricacies and developing frameworks for comprehensive and coherent analysis across multiple analytical dimensions.

Additionally, applying cross-cultural analysis presents an opportunity to examine how the discourse surrounding decision-making algorithms differs across various cultural contexts. For example, comparing discourse in different countries or regions could uncover cultural differences in trust, blame attribution and agency perception, foregrounding the influence of societal factors (Phillips and Cassidy, 2024). Further to this, comparative studies could analyse discourse surrounding different types of decision-making algorithms, such as those used in finance or criminal justice. This approach could provide insights into domain-specific challenges and perceptions, informing tailored approaches to addressing societal concerns. This may account for the influence of socio-political context, cultural nuances and external events on discourse dynamics. Additionally, efforts to explore alternative theoretical frameworks, such as feminine agencies, could help mitigate biases and enrich interpretations of social media discourses surrounding decision-making algorithms.

Moreover, user studies, involving in-depth interviews or surveys, could complement this analysis by providing data on individu-

als' perceptions and attitudes towards future public-facing decision-making algorithms. This approach could offer further insights into the factors driving trust, blame attribution and agency perception.

Finally and, perhaps, most importantly, as previously stated, these findings have the potential to be used by the developers and promoters of public-facing decision-making algorithms to overcome barriers to adoption. While only capturing views from three different Twitter discourses, these findings could springboard the development and evaluation of interventions aimed at improving public understanding of decision-making algorithms. This valuable area for future research could manifest in many different ways. For example, educational programs or tools designed to enhance algorithmic literacy and critical thinking skills could empower individuals to engage more meaningfully with AI technologies and avoid misconceptions.

In summary, these potential avenues for future work could further our understanding of how public-facing decision-making algorithms are perceived, discussed and, ultimately, impact society. By building on the foundations of this PhD thesis, research in this area can contribute to more informed policy-making and technology development in the field of AI.

## 8.4 CONCLUDING REMARKS

Overall, this PhD thesis has aimed to investigate views expressed about decision-making algorithms on Twitter, using NLP-based tools, CL and DA, to understand agency, trust and blame attribution. Of the case studies examined, all three systems were portrayed as social actors in their respective captured discourses due to the active grammatical agency and presentation, which showcased various degrees of responsibility and accountability. All three were implicated as blameworthy, although the extent for ChatGPT was considerably less than the other two systems, perhaps due to the context in which

these appeared to be public-facing in the first instance. The inverse was also found for trust. The main contribution of this thesis includes insights into the online discourse surrounding decision-making algorithms, with a secondary contribution of an approach to social media exploration that combines existing analyses. It now hoped that these findings have practical implications for policymakers, developers and promoters of decision-making algorithms, potentially assisting in overcoming adoption barriers.

Part IV

# APPENDIX

# A

# APPENDIX A: PRIVACY NOTICE

# PRIVACY NOTICE


The University of Nottingham is committed to protecting your personal data and informing you of your rights in relation to that data. The University will process your personal data in accordance with the General Data Protection Regulation (GDPR) and the Data Protection Act 2018 and this privacy notice is issued in accordance with GDPR Articles 13 and 14.

The University of Nottingham, University Park, Nottingham, NG7 2RD is registered as a Data Controller under the Data Protection Act 1998 (registration No. Z5654762, https://ico.org.uk/ESDWebPages/Entry/Z5654762**).**

The University has appointed a Data Protection Officer (DPO). The DPO's postal address is:

Data Protection Officer,
Legal Services
A5, Trent Building,
University of Nottingham,
University Park,
Nottingham
NG7 2RD

The DPO can be emailed at dpo@nottingham.ac.uk

**Why we collect your personal data**. We collect personal data under the terms of the University's Royal Charter in our capacity as a teaching and research body to advance education and learning. Specific purposes for data collection on this occasion are to develop insight into the public perception of Autonomous Systems.

**The legal basis for processing your personal data under GDPR.** Under the General Data Protection Regulation, the University must establish a legal basis for processing your personal data and communicate this to you. The legal basis for processing your personal data on this occasion is Article 6(1e) processing is necessary for the performance of a task carried out in the public interest.

*Note: Article 6(1e) public interest should be used by default whenever possible, as this fits the University's role as a teaching and research body to advance education and learning. This does not mean that you do not need to obtain consent from research participants, only that consent does not provide the legal basis for processing participant's data. In exceptional cases, where the public interest clause does not apply, e.g., if you are doing research on behalf of an external organisation (such as a commercial company), then* Article 6(1a) consent of the data subject *should be used instead.*

**Where the University receives your personal data from.** On this occasion, personal data is obtained from publicly available sources on Twitter.

**How long we keep your data**. The University may store your data for up to 25 years and for a period of no less than 7 years after the research project finishes. The researchers who gathered or processed the data may also store the data indefinitely and reuse it in future research.

**Who we share your data with.** Your data may be shared with researchers from other collaborating institutions and organisations who are involved in the research. Extracts of your data may be disclosed in published works that are posted online for use by the scientific community. Your data may also be stored indefinitely by members of the researcher team and/or be stored on external data repositories (e.g., the UK Data Archive) and be further processed for archiving purposes in the public interest, or for historical, scientific or statistical purposes.

**How we keep your data safe**. We keep your data securely and put measures in place to safeguard it. These safeguards include an encryption protocol, pseudonymisation procedure and anonymisation of data. The data is stored securely on University of Nottingham servers and Microsoft Teams.

Your data may be transferred for processing within the University of Nottingham.

**Your rights as a data subject**. GDPR provides you, as a data subject, with a number of rights in relation to your personal data. Subject to some exemptions, you have the right to:

- withdraw your consent at any time where that is the legal basis of our processing, and in such circumstances you are not obliged to provide personal data for our research.
- object to automated decision-making, to contest the decision, and to obtain human intervention from the controller.
- access (i.e., receive a copy of) your personal data that we are processing together with information about the purposes of processing, the categories of personal data concerned, recipients/categories of recipient, retention periods, safeguards for any overseas transfers, and information about your rights.
- have inaccuracies in the personal data that we hold about you rectified and, depending on the purposes for which your data is processed, to have personal incomplete data completed
- be forgotten, i.e., to have your personal data erased where it is no longer needed, you withdraw consent and there is no other legal basis for processing your personal data, or you object to the processing and there is no overriding legitimate ground for that processing.
- in certain circumstances, request that the processing of your personal data be restricted, e.g., pending verification where you are contesting its accuracy or you have objected to the processing.
- obtain a copy of your personal data which you have provided to the University in a structured, commonly used electronic form (portability), and to object to certain processing activities such as processing based on the University's or someone else's legitimate interests, processing in

the public interest or for direct marketing purposes. In the case of objections based on the latter, the University is obliged to cease processing.

- complain to the Information Commissioner's Office about the way we process your personal data.

If you require advice on exercising any of the above rights, please contact the University's data protection team: data-protection@nottingham.ac.uk

---

*Notes*

*1. The provision of the statutory information contained in this privacy notice to the data subject when processing is legally based on public interest is not necessary* ***if*** *the provision of such information proves impossible or involves a disproportionate effort.*

*2. If it is not impossible or disproportionate to provide data subjects with the statutory information contained in this privacy notice, then it* ***must*** *be provided to data subjects within* ***1 month of obtaining their personal data****.*

# B

# APPENDIX B: PROJECT INFORMATION

# PROJECT INFORMATION

University of Nottingham
UK | CHINA | MALAYSIA

**Date:** 25/01/2021

**Project:** Developing Natural Language analysis tools to capture public discourse of decision-making algorithms

**School of Computer Science Ethics Reference:** CS-2020-R33

**Funded by:** Horizon CDT

**Purpose of the research.** This project aims to investigate the perception of decision-making algorithms, including the NHS Covid-19 contact tracing mobile application and Ofqual A Level algorithm, on Twitter by using a new Natural Language analysis tool to extract sentiment over time. This is important research as it will provide insight as to the public perception of decision-making algorithms as an Autonomous System must be trustworthy by design and perception in order to be accepted into everyday life.

**Nature of participation.** Participation in this research is voluntary. Data has already been provided by a third party, for which consent is being sought.

**Participant engagement.** Social media data has already been provided and no further participant engagement is required.

**Benefits and risks of the research.** Your data may help us understand the perception of decision-making algorithms. There is no risk that your data would identify you in research reports or publications.

**Use of your data.** Data gathered in this research process will be used in project meetings and project reports. The results of this research will be shared in a presentation to the Horizon CDT 2020 cohort, as well as in research reports, doctoral thesis and publications.

**Future use of your data.** Your data may be archived and reused in future for purposes that are in the public interest, or for historical, scientific or statistical purposes. Data will be stored securely on University of Nottingham servers and their accompanying Microsoft Teams interface.

**Procedure for withdrawal from the research** You may withdraw from the study at any time and do not have to give reasons for why you no longer want to take part. If you wish to withdraw please contact the researcher who gathered the data. If you receive no response from the researcher, please contact the School of Computer Science's Ethics Committee.

**Contact details of the ethics committee.** If you wish to file a complaint or exercise your rights, you can contact the Ethics Committee at the following address: cs-ethicsadmin@cs.nott.ac.uk

# BIBLIOGRAPHY


Abbas, Akhtar and Tehseen Zahra (2021). "Corpus Driven Critical Discourse Analysis of 2020 Presidential Election Campaign Tweets of Donald Trump and Joe Biden." In: *Hayatian Journal of Linguistics and Literature* 5.1, pp. 26–47. URL: http://111.68.104.137/index.php/HJLL/article/view/12.

Abbasi, Kamran (2021). *Covid-19: The UK's political gamble that bodes ill for health and the health service*. DOI: 10.1136/bmj.n1848. eprint: https://www.bmj.com/content/374/bmj.n1848.full.pdf. URL: https://www.bmj.com/content/374/bmj.n1848.

Abdullah, Malak, Alia Madain, and Yaser Jararweh (2022). "ChatGPT: Fundamentals, applications and social impacts." In: *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, pp. 1–8. DOI: 10.1109/SNAMS58071.2022.10062688.

Adams, Barbara D., Lora E. Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann (2003). "Trust in automated systems." In: *Toronto: Ministry of National Defence*. URL: https://cradpdf.drdc-rddc.gc.ca/PDFS/unc13/p520342.pdf.

Adolphs, Svenja and Phoebe MS Lin (2011). "Corpus linguistics." In: *The Routledge handbook of applied linguistics*. Routledge, pp. 597–610. DOI: 10.4324/9780203835654.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau (2011). "Sentiment analysis of twitter data." In: *Proceedings of the workshop on language in social media (LSM 2011)*, pp. 30–38. URL: https://aclanthology.org/W11-0705.pdf.

Agarwal, Basant, Namita Mittal, Pooja Bansal, and Sonal Garg (2015). "Sentiment analysis using common-sense and context in-

formation." In: *Computational intelligence and neuroscience* 2015.1, p. 715730. DOI: 10.1155/2015/715730.

Ahearn, Laura M (1999). "Agency." In: *Journal of Linguistic Anthropology* 9.1/2, pp. 12–15.

Ahmed, Wasim, Peter A. Bath, and Gianluca Demartini (2017). "Using Twitter as a data source: An overview of ethical, legal, and methodological challenges." In: *The Ethics of Online Research*. ISSN: 1787144860. DOI: 10.1108/S2398-601820180000002004.

Aiyappa, Rachith, Jisun An, Haewoon Kwak, and Yong-yeol Ahn (2023). "Can we trust the evaluation on ChatGPT?" In: *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Al Lily, Abdulrahman Essa, Abdelrahim Fathy Ismail, Fathi M Abunaser, Firass Al-Lami, and Ali Khalifa Atwa Abdullatif (2023). "ChatGPT and the rise of semi-humans." In: *Humanities and Social Sciences Communications* 10.1, pp. 1–12. DOI: 10.1057/s41599-023-02154-3.

Alaieri, Fahad and André Vellino (2016). "Ethical Decision Making in Robots: Autonomy, Trust and Responsibility: Autonomy Trust and Responsibility." In: *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, pp. 159–168. DOI: 10.1007/978-3-319-47437-3_16.

Alamoodi, AH, Bilal Bahaa Zaidan, Aws Alaa Zaidan, Osamah Shihab Albahri, KI Mohammed, Rami Qays Malik, EM Almahdi, MA Chyad, Zaidoon Tareq, Ahmed Shihab Albahri, et al. (2021). "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review." In: *Expert systems with applications* 167, p. 114155. DOI: 10.1016/j.eswa.2020.114155.

Ali, Rohaid, Oliver Y Tang, Ian D Connolly, Jared S Fridley, John H Shin, Patricia L Zadnik Sullivan, Deus Cielo, Adetokunbo A Oyelese, Curtis E Doberstein, Albert E Telfeian, et al. (2022). "Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank." In: *Neurosurgery*, pp. 10–1227. DOI: 10.1227/neu.0000000000002551.

Aljanabi, Mohammad (2023). "ChatGPT: Future directions and open possibilities." In: *Mesopotamian journal of Cybersecurity* 2023, pp. 16–17. URL: https://www.iasj.net/iasj/article/303952.

Aljarallah, Rayya Sulaiman (2017). *A critical discourse analysis of twitter posts on the perspectives of women driving in Saudi Arabia*. Tech. rep. Arizona State University. URL: https://keep.lib.asu.edu/items/155782.

Amoussou, Franck and Ayodele A Allagbe (2018). "Principles, theories and approaches to critical discourse analysis." In: *International Journal on Studies in English Language and Literature* 6.1, pp. 11–18. DOI: 10.20431/2347-3134.0601002.

Antaki, Fares, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval (2023). "Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings." In: *Ophthalmology Science*, p. 100324. DOI: 10.1016/j.xops.2023.100324.

Ante, Lennart and Ender Demir (2024). "The ChatGPT effect on AI-themed cryptocurrencies." In: *Economics and Business Letters* 13.1, pp. 29–38. DOI: 10.2139/ssrn.4350557.

Anthony, Laurence (2013). "A critical look at software tools in corpus linguistics." In: *Linguistic Research* 30.2, pp. 141–161.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese (2020). "In AI we trust? Perceptions about automated decision-making by artificial intelligence." In: *AI & Society* 35.3, pp. 611–623. DOI: 10.1007/s00146-019-00931-w.

Archer, Margaret Scotford (2000). *Being human: The problem of agency*. Cambridge University Press.

Arianto, Bagus Wicaksono and Gangga Anuraga (2020). "Topic Modeling for Twitter Users Regarding the "Ruangguru" Application." In: *Jurnal ILMU DASAR*. DOI: 10.19184/jid.v21i2.17112.

Aribowo, Agus Sasmito and Siti Khomsah (2021). "Implementation Of Text Mining For Emotion Detection Using The Lexicon

Method (Case Study: Tweets About Covid-19)." In: *Telematika*. DOI: 10.31315/telematika.v18i1.4341.

Arnett, Jeffrey J (2015). "The neglected 95%: Why American psychology needs to become less American." In: *Methodological issues and strategies in clinical research (4th ed.)* Washington: American Psychological Association, pp. 115–132. DOI: 10.1037/14805-008.

Arnold, Austin, Anne Cafer, John Green, Seena Haines, Georgianna Mann, and Meagen Rosenthal (2021). "Perspective: Promoting and fostering multidisciplinary research in universities." In: *Research Policy* 50.9, p. 104334. DOI: 10.1016/j.respol.2021.104334.

Atteveldt, Wouter van, Mariken ACG van der Velden, and Mark Boukes (2021). "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." In: *Communication Methods and Measures* 15.2, pp. 121–140. DOI: 10.1080/19312458.2020.1869198.

BBC (2020). *A-levels and GCSEs: U-turn as teacher estimates to be used for exam results*. URL: https://www.bbc.co.uk/news/uk-53810655.

Baker, Paul (2006). *Using corpora in discourse analysis*. A&C Black.

— (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.

— (2012). "Acceptable bias? Using corpus linguistics methods with critical discourse analysis." In: *Critical Discourse Studies* 9.3, pp. 247–256. DOI: 10.1080/17405904.2012.688297.

Baker, Paul and Erez Levon (2015). "Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity." In: *Discourse & Communication* 9.2, pp. 221–236. DOI: 10.1177/1750481314568542.

Balakrishnan, Vimala and Wandeep Kaur (2019). "String-based Multinomial Naïve Bayes for Emotion Detection among Facebook Diabetes Community." In: *Procedia Computer Science* 159. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019, pp. 30–

37. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.09.157. URL: https://www.sciencedirect.com/science/article/pii/S1877050919313353.

Balakrishnan, Vimala, Marian Cynthia Martin, Wandeep Kaur, and Amir Javed (2019). "A comparative analysis of detection mechanisms for emotion detection." In: *Journal of Physics: Conference Series*. Vol. 1339. 1. IOP Publishing, p. 012016. DOI: 10.1088/1742-6596/1339/1/012016.

Bandura, Albert (2001). "Social cognitive theory: An agentic perspective." In: *Annual Review of Psychology* 52.1, pp. 1–26. DOI: 10.1146/annurev.psych.52.1.1.

Bardosh, Kevin, Alex De Figueiredo, Rachel Gur-Arie, Euzebiusz Jamrozik, James Doidge, Trudo Lemmens, Salmaan Keshavjee, Janice E Graham, and Stefan Baral (2022). "The unintended consequences of COVID-19 vaccine policy: why mandates, passports and restrictions may cause more harm than good." In: *BMJ Global Health* 7.5, e008684. DOI: 10.1136/bmjgh-2022-008684.

Barocas, Solon and Andrew D Selbst (2016). "Big data's disparate impact." In: *Calif. L. Rev.* 104, pp. 671–732. DOI: 10.15779/Z38BG31.

Baumeister, Roy F (1996). *Evil: Inside human cruelty and violence.* WH Freeman/Times Books/Henry Holt & Co.

Baumeister, Roy F, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs (2001). "Bad is stronger than good." In: *Review of general psychology* 5.4, pp. 323–370. DOI: 10.1037/1089-2680.5.4.323.

Baumeister, Roy F and Mark R Leary (2017). "The need to belong: Desire for interpersonal attachments as a fundamental human motivation." In: *Interpersonal development*. Routledge, pp. 57–89.

Beer, David Gareth (2017). "The Social Power of Algorithms." In: *Information, Communication and Society* 20.1, pp. 1–13. DOI: 10.1080/1369118X.2016.1216147.

Benjamin, Garfield (2022). "# FuckTheAlgorithm: algorithmic imaginaries and political resistance." In: *ACM Conference on Fairness,*

*Accountability, and Transparency 2022*. DOI: 10.1145/3531146.3533072.

Bernard, Taryn (2018). "The Discursive Representation of Social Actors in the Corporate Social Responsibility (CSR) and Integrated Annual (IA) Reports of Two South African Mining Companies." In: *Critical Approaches to Discourse Analysis across Disciplines* 10.1, pp. 81–97.

Bhopal, Kalwant and Martin Myers (2020). *The impact of COVID-19 on A level students in England*. DOI: 10.31235/osf.io/j2nqb. URL: osf.io/preprints/socarxiv/j2nqb.

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

Bishop, Libby and Daniel Gray (2017). "Ethical challenges of publishing and sharing social media research data." In: *The ethics of online research*. Vol. 2. Emerald Publishing Limited, pp. 159–187. DOI: 10.1108/S2398-601820180000002007.

Biswas, Som (2023). "Will ChatGPT take my Job? Replies and Advice by ChatGPT." In: *Replies and Advice by ChatGPT.(May 3, 2023)*. DOI: 10.2139/ssrn.4437405.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation." In: *Journal of machine Learning research* 3.Jan, pp. 993–1022. DOI: 10.5555/944919.944937.

Bloor, Meriel and Thomas Bloor (2013). *The practice of critical discourse analysis: An introduction*. Routledge. DOI: 10.4324/9780203775660.

Bollen, Johan, Huina Mao, and Alberto Pepe (2011). "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 5. 1, pp. 450–453. DOI: 10.1609/icwsm.v5i1.14171.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan (2016). "The social dilemma of autonomous vehicles." In: *Science* 352.6293, pp. 1573–1576. DOI: 10.1126/science.aaf2654.

Boucher, Abigail, Marcello Giovanelli, Chloe Harrison, Robbie Love, and Caroline Godfrey (2024). *Reading Habits in the COVID-19 Pandemic: An Applied Linguistic Perspective*. Springer Nature.

Bovens, Marcus Alphons Petrus, Mark Bovens, et al. (1998). *The quest for responsibility: Accountability and citizenship in complex organisations*. Cambridge University Press.

Bovens, Mark (2007). "Analysing and assessing accountability: A conceptual framework 1." In: *European law journal* 13.4, pp. 447–468.

Bran, Emanuela, Cosima Rughiniş, Gheorghe Nadoleanu, and Michael G Flaherty (2023). "The Emerging Social Status of Generative AI: Vocabularies of AI Competence in Public Discourse." In: *2023 24th International Conference on Control Systems and Computer Science (CSCS)*. IEEE, pp. 391–398. DOI: 10.1109/CSCS59211.2023.00068.

Brookes, Gavin (2023). "Killer, thief or companion? A corpus-based study of dementia metaphors in UK tabloids." In: *Metaphor and Symbol* 38.3, pp. 213–230.

Brookes, Gavin and Tony McEnery (2020). "Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse." In: *Discourse & Society* 31.4, pp. 351–373.

Brown, Gillian (1983). "Discourse analysis." In: *Cambridge: Cambridge University Press*.

Bryson, Joanna J (2020). "The artificial intelligence of the ethics of artificial intelligence." In: *The Oxford handbook of ethics of AI*, p. 1.

Bucher, Taina (2017). "The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms." In: *Information, communication & society* 20.1, pp. 30–44. DOI: 10.1080/1369118X.2016.1154086.

Bucholtz, Mary (June 2001). "Reflexivity and Critique in Discourse Analysis." In: *Critique of Anthropology* 21.2, pp. 165–183. ISSN: 0308-275X. DOI: 10.1177/0308275x0102100203. URL: https://doi.org/10.1177/0308275X0102100203.

Bullock, Justin B (2019). "Artificial intelligence, discretion, and bureaucracy." In: *The American Review of Public Administration* 49.7, pp. 751–761. DOI: 10.1177/0275074019856123.

Burrell, Jenna (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms." In: *Big data & society* 3.1, p. 2053951715622512. DOI: 10.1177/2053951715622512.

Busch, Peter André and Helle Zinner Henriksen (2018). "Digital discretion: A systematic literature review of ICT and street-level discretion." In: *Information Polity* 23.1, pp. 3–28. DOI: 10.3233/IP-170050.

Cao, Yi and Jia Zhai (2023). "Bridging the gap–the impact of ChatGPT on financial research." In: *Journal of Chinese Economic and Business Studies*, pp. 1–15. DOI: 10.1080/14765284.2023.2212434.

Carr, Gemma, Daniel P Loucks, and Günter Blöschl (2018). "Gaining insight into interdisciplinary research and education programmes: A framework for evaluation." In: *Research Policy* 47.1, pp. 35–48. DOI: 10.1016/j.respol.2017.09.010.

Carroll, Archie B (1979). "A three-dimensional conceptual model of corporate performance." In: *Academy of management review* 4.4, pp. 497–505. DOI: 10.5465/amr.1979.4498296.

Chaithra, V. D. (2019a). "Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments." In: *International Journal of Electrical and Computer Engineering*. DOI: 10.11591/ijece.v9i5.pp4452-4459.

Chaithra, VD (2019b). "Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments." In: *International Journal of Electrical and Computer Engineering (IJECE)* 9.5, pp. 4452–4459. DOI: 10.11591/ijece.v9i5.pp4452-4459.

Chang, Jenny Hsiu-Ying, Honggang Yang, Kuang-Hui Yeh, and Shih-Chi Hsu (2016). "Developing trust in close personal relationships: Ethnic Chinese's experiences." In: *Journal of Trust Research* 6.2, pp. 167–193. DOI: 10.1080/21515581.2016.1207543.

Chauhan, Vipul Kumar, Ashish Bansal, and Amita Goel (2018). "Twitter sentiment analysis using vader." In: *International Journal of Advance Research, Ideas and Innovations in Technology* 4.1, pp. 485–489.

Chen, Hong, Kang Yuan, Yanjun Huang, Lulu Guo, Yulei Wang, and Jie Chen (2023). "Feedback is all you need: from ChatGPT to autonomous driving." In: *Science China Information Sciences* 66.6, pp. 1–3. DOI: 10.1007/s11432-023-3740-x.

Cho, Yoon Jik and Hanjun Park (2011). "Exploring the relationships among trust, employee satisfaction, and organizational commitment." In: *Public Management Review* 13.4, pp. 551–573. DOI: 10.1080/14719037.2010.525033.

Chong, W. Y., B. Selvaretnam, and L. Soon (2014). "Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets." In: *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, pp. 212–217. DOI: 10.1109/icaiet.2014.43.

Choudhury, Avishek and Hamid Shamszare (2023). "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis." In: *Journal of Medical Internet Research* 25, e47184. DOI: 10.2196/47184.

Clark, William Roberts (1998). "Agents and structures: two views of preferences, two views of institutions." In: *International Studies Quarterly* 42.2, pp. 245–270. DOI: 10.1111/1468-2478.00081.

Clerwall, Christer (2017). "Enter the robot journalist: Users' perceptions of automated content." In: *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty*. Routledge, pp. 165–177.

Coates, D Justin and Neal A Tognazzini (2013). *Blame: Its nature and norms*. Oxford University Press on Demand.

Coeckelbergh, Mark (2020a). *AI Ethics*. MIT Press. DOI: 10.7551/mitpress/12549.001.0001.

Coeckelbergh, Mark (2020b). "Artificial intelligence, responsibility attribution, and a relational justification of explainability." In: *Sci-*

*ence and engineering ethics* 26.4, pp. 2051–2068. DOI: 10.1007/s11948-019-00146-8.

Comrie, Bernard (1977). "In defense of spontaneous demotion: The impersonal passive." In: *Grammatical relations*. Brill, pp. 47–58. DOI: 10.1163/9789004368866_004.

Cook, G. (1989). *Discourse*. Oxford University Press.

Cook, Karen S and Karen A Hegtvedt (1983). "Distributive justice, equity, and equality." In: *Annual Review of Sociology* 9.1, pp. 217–241. DOI: 10.1146/annurev.so.09.080183.001245.

Coughlan, Sean (2020). *A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco*. URL: https://www.bbc.co.uk/news/education-53923279.

Crang, Mike and Stephen Graham (2007). "Sentient cities ambient intelligence and the politics of urban space." In: *Information, Communication & Society* 10.6, pp. 789–817. DOI: 10.1080/13691180701750991.

Crisp, Victoria, Gill Elliott, Emma Walland, and Lucy Chambers (2024). "A structured discussion of the fairness of GCSE and A level grades in England in summer 2020 and 2021." In: *Research Papers in Education*, pp. 1–28. DOI: 10.1080/02671522.2024.2318046.

Cruse, David Alan (1986). *Lexical semantics*. Cambridge university press.

Cushing, Judy and Rachel Hastings (2009). "Introducing computational linguistics with NLTK (Natural Language Toolkit)." In: *Journal of Computing Sciences in Colleges*. DOI: 10.5555/1619221.1619254.

Dai, Yun, Ang Liu, and Cher Ping Lim (2023). "Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education." In: *Procedia CIRP* 119, pp. 84–90. DOI: 10.1016/j.procir.2023.05.002.

De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Cate-

rina Rizzo (2023). "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health." In: *Frontiers in Public Health* 11, p. 1166120. DOI: 10.3389/fpubh.2023.1166120.

De Silva, Daswin et al. (2018). "Machine learning to support social media empowered patients in cancer care and cancer treatment decisions." In: *PloS one* 13.10, e0205855. DOI: 10.1371/journal.pone.0205855.

Deignan, Alice and Robbie Love (2021). "Using corpus methods to identify subject specific uses of polysemous words in English secondary school science materials." In: *Corpora* 16.2, pp. 165–189.

Devitt, S (2018). "Trustworthiness of autonomous systems." In: *Foundations of trusted autonomy (Studies in Systems, Decision and Control, Volume 117)*, pp. 161–184. DOI: 10.1007/978-3-319-64816-3_9.

Devlin, Kate (2023). "Power in AI: Inequality Within and Without the Algorithm." In: *The Handbook of Gender, Communication, and Women's Human Rights*, pp. 123–139. DOI: 10.1002/9781119800729.ch8.

Diakopoulos, Nicholas (2016). "Accountability in algorithmic decision making." In: *Communications of the ACM* 59.2, pp. 56–62. DOI: 10.1145/2844110.

Dietz, Graham, Nicole Gillespie, and Georgia T Chao (2010). "Unravelling the complexities of trust and culture." In: *Organizational trust: A cultural perspective*, pp. 3–41. DOI: 10.1017/CBO9780511763106.002.

Dignum, Virginia (2020). "AI is multidisciplinary." In: *AI Matters* 5.4, pp. 18–21.

Dirks, Kurt T and Donald L Ferrin (2002). "Trust in leadership: meta-analytic findings and implications for research and practice." In: *Journal of applied psychology* 87.4, p. 611. DOI: 10.1037/0021-9010.87.4.611.

Dönmez, İsmail, IDİN Sahin, and Salih Gülen (2023). "Conducting academic research with the ai interface ChatGPT: Challenges and

opportunities." In: *Journal of STEAM Education* 6.2, pp. 101–118. DOI: 10.55290/steam.1263404.

Donovan, Joan, Robyn Caplan, Jeanna Matthews, and Lauren Hanson (2018). *Algorithmic accountability: A primer*. Tech. rep. URL: http://portaldaprivacidade.com.br/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf.

Dörr, Konstantin Nicholas (2016). "Mapping the field of algorithmic journalism." In: *Digital journalism* 4.6, pp. 700–722. DOI: 10.1080/21670811.2015.1096748.

Doshi, Rushabh H, Simar S Bajaj, and Harlan M Krumholz (2023). "ChatGPT: temptations of progress." In: *The American Journal of Bioethics* 23.4, pp. 6–8. DOI: 10.1080/15265161.2023.2180110.

Dowthwaite, Liz, Joel Fischer, Elvira Perez Vallejos, Virginia Portillo, Elena Nichele, Murray Goulden, and Derek McAuley (2021). "Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study." In: *Journal of Medical Internet Research* 23.9, e29085. DOI: 10.2196/29085.

Dowthwaite, Liz, Hanne Gesine Wagner, Camilla May Babbage, Joel E. Fischer, Pepita Barnard, Elena Nichele, Elvira Perez Vallejos, Jeremie Clos, Virginia Portillo, and Derek McAuley (Oct. 2022). "The relationship between trust and attitudes towards the COVID-19 digital contact-tracing app in the UK." In: *PLOS ONE* 17.10, pp. 1–24. DOI: 10.1371/journal.pone.0276661. URL: https://doi.org/10.1371/journal.pone.0276661.

Duranti, Alessandro (2008). *A Companion to Linguistic Anthropology*. John Wiley & Sons.

Dwivedi, Yogesh K, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. (2021). "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy." In: *In-*

*ternational Journal of Information Management* 57, p. 101994. DOI: 10.1016/j.ijinfomgt.2019.08.002.

Edwards, Chris (2021). "Let the algorithm decide?" In: *Communications of the ACM* 64.6, pp. 21–22. DOI: 10.1145/3460216.

Emirbayer, Mustafa and Ann Mische (1998). "What is agency?" In: *American journal of sociology* 103.4, pp. 962–1023.

Engeström, Yrjö et al. (1999). "Activity theory and individual and social transformation." In: *Perspectives on activity theory* 19.38, pp. 19–30. DOI: 10.1017/CBO9780511812774.003.

Fadanelli, Sabrina Bonqueves, Daniela Fátima Dal Pozzo, and Claudia Cristina Fin (2020). "The representation of social actors in the tweets of Jair Messias Bolsonaro." In: *Antares* 12, pp. 74–99. DOI: 10.18226/19844921.v12.n25.04.

Fairclough, Norman (1993). "Critical discourse analysis and the marketization of public discourse: The universities." In: *Discourse & society* 4.2, pp. 133–168. DOI: 10.1177/0957926593004002002.

Fallman, Daniel (2008). "The interaction design research triangle of design practice, design studies, and design exploration." In: *Design issues* 24.3, pp. 4–18. DOI: 10.1162/desi.2008.24.3.4.

Fang, Eric, Robert W Palmatier, Lisa K Scheer, and Ning Li (2008). "Trust at different organizational levels." In: *Journal of marketing* 72.2, pp. 80–98. DOI: https://doi.org/10.1509/jmkg.72.2.80.

Fast, Ethan, Binbin Chen, and Michael S. Bernstein (2016). "Empath: Understanding Topic Signals in Large-Scale Text." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 4647–4657. ISBN: 9781450333627. URL: https://doi.org/10.1145/2858036.2858535.

Feier, Till, Jan Gogoll, and Matthias Uhl (2021). "Hiding Behind Machines: When Blame Is Shifted to Artificial Agents." In: arXiv: 2101.11465 [cs.CY]. URL: https://arxiv.org/abs/2101.11465.

Feng, Yunhe, Sreecharan Vanam, Manasa Cherukupally, Weijian Zheng, Meikang Qiu, and Haihua Chen (2023). "Investigating

Code Generation Performance of Chat-GPT with Crowdsourcing Social Data." In: *Proceedings of the 47th IEEE Computer Software and Applications Conference*, pp. 1–10. DOI: 10.1109/COMPSAC57700.2023.00117.

Fernández-Cruz, Javier and Antonio Moreno-Ortiz (2023). "Tracking diachronic sentiment change of economic terms in times of crisis: Connotative fluctuations of 'inflation'in the news discourse." In: *Plos one* 18.11, e0287688. DOI: 10.1371/journal.pone.0287688.

Ferrara, Emilio (Nov. 2023). "Should ChatGPT be biased? Challenges and risks of bias in large language models." In: *First Monday*. ISSN: 1396-0466. DOI: 10.5210/fm.v28i11.13346. URL: http://dx.doi.org/10.5210/fm.v28i11.13346.

Finlay, Linda (2008). "Reflecting on 'Reflective practice'." In: *Practice-based Professional Learning Paper 52, The Open University*. URL: https://oro.open.ac.uk/68945/.

Firat, Mehmet (2023). *How ChatGPT Can Transform Autodidactic Experiences and Open Education?* DOI: 10.31219/osf.io/9ge8m. URL: osf.io/9ge8m.

Fischer, Joel E (2023). "Generative AI Considered Harmful." In: *Proceedings of the 5th International Conference on Conversational User Interfaces*. CUI '23. Eindhoven, Netherlands: Association for Computing Machinery. ISBN: 9798400700149. DOI: 10.1145/3571884.3603756. URL: https://doi.org/10.1145/3571884.3603756.

Floridi, Luciano and Josh Cowls (2022). "A unified framework of five principles for AI in society." In: *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545. DOI: 10.1002/9781119815075.ch45.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. (2018). "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations." In: *Minds and Machines* 28, pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

Fujioka, Takuya, Dario Bertero, Takeshi Homma, and Kenji Nagamatsu (2019). *Addressing Ambiguity of Emotion Labels Through Meta-Learning*. arXiv: 1911.02216 [eess.AS]. URL: https://arxiv.org/abs/1911.02216.

Gallagher, Shaun (2000). "Philosophical conceptions of the self: implications for cognitive science." In: *Trends in Cognitive Sciences* 4.1, pp. 14–21. DOI: 10.1016/S1364-6613(99)01417-5.

Gallifant, Jack, Amelia Fiske, Yulia A Levites Strekalova, Juan S Osorio-Valencia, Rachael Parke, Rogers Mwavu, Nicole Martinez, Judy Wawira Gichoya, Marzyeh Ghassemi, Dina Demner-Fushman, et al. (2024). "Peer review of GPT-4 technical report and systems card." In: *PLOS Digital Health* 3.1, e0000417. DOI: 10.1371/journal.pdig.0000417.

Garfinkel, Simson, Jeanna Matthews, Stuart S Shapiro, and Jonathan M Smith (2017). "Toward algorithmic transparency and accountability." In: *Communications of the ACM* 60.9, pp. 5–5. DOI: 10.1145/3125780.

Gee, James Paul (2014). *An introduction to discourse analysis: Theory and method*. Routledge.

George, A Shaji and AS Hovan George (2023). "A review of ChatGPT AI's impact on several business sectors." In: *Partners Universal International Innovation Journal* 1.1, pp. 9–23. DOI: 10.5281/zenodo.7644359.

Gibbs, G. and Further Education Unit (1988). *Learning by doing: A guide to teaching and learning methods*. English. Further Education Unit. URL: shop.brookes.ac.uk/browse/extra_info.asp?prodid=935.

Giddens, Anthony (1984). *The constitution of society: Outline of the theory of structuration*. Polity.

— (1986). *The constitution of society: Outline of the theory of structuration*. Vol. 349. Univ of California Press.

Giddens, Anthony (2007). "The Consequences of Modernity. 1990." In: *Contemporary sociological theory*. Ed. by Craig J. Calhoun. Blackwell, pp. 2–243.

Gilbert, Daniel T and Patrick S Malone (1995). "The correspondence bias." In: *Psychological bulletin* 117.1, p. 21. DOI: 10.1037/0033-2909.117.1.21.

Gill, Rosalind (2000). "Discourse analysis." In: *Qualitative researching with text, image and sound* 1, pp. 172–190. DOI: 10.4135/9781849209731.

Gillespie, Tarleton (2014). "The relevance of algorithms." In: *Media technologies: Essays on communication, materiality, and society* 167.2014, p. 167. DOI: 10.7551/mitpress/9042.003.0013.

Goatly, Andrew (2007). *Washing the brain: Metaphor and hidden ideology*. Vol. 23. John Benjamins Publishing. URL: http://digital.casalini.it/9789027292933.

González-Ibánez, Roberto, Smaranda Muresan, and Nina Wacholder (2011). "Identifying sarcasm in twitter: a closer look." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 581–586. DOI: 10.5555/2002736.2002850.

Graefe, Andreas and Nina Bohlken (2020). "Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news." In: *Media and Communication* 8.3, pp. 50–59. DOI: 10.17645/mac.v8i3.3019.

Grange, Camille (2022). "Algorithmically Controlled Automated Decision-Making and Societal Acceptability: Does Algorithm Type Matter?" In: *HICSS*, pp. 1–10. URL: http://hdl.handle.net/10125/79967.

Green, Melanie C. (2007). "Trust and social interaction on the Internet." In: *The Oxford handbook of Internet psychology* 56, pp. 43–51. DOI: 10.1093/oxfordhb/9780199561803.013.0004.

Grillitsch, Markus, Josephine V Rekers, and Markku Sotarauta (2021). "Investigating agency: Methodological and empirical challenges."

In: *Handbook on city and regional leadership*. Edward Elgar Publishing. DOI: 10.4337/9781788979689.00028.

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang (2019). "XAI—Explainable artificial intelligence." In: *Science Robotics* 4.37, p. 7120. DOI: 10.1126/scirobotics.aay7120.

Gupta, Itisha and Nisheeth Joshi (2021). "A Review on Negation Role in Twitter Sentiment Analysis." In: *International Journal of Healthcare Information Systems and Informatics*. DOI: 10.4018/ijhisi.20211001oa06.

Gupta, Vibhuti and Rattikorn Hewett (2020). "Real-time tweet analytics using hybrid hashtags on twitter big data streams." In: *Information* 11.7, p. 341. DOI: 10.3390/info11070341.

Gutiérrez, Jorge Luis Morton (2023). "On actor-network theory and algorithms: ChatGPT and the new power relationships in the age of AI." In: *AI and Ethics*, pp. 1–14. DOI: 10.1007/s43681-023-00314-4.

Haleem, Abid, Mohd Javaid, and Ravi Pratap Singh (2022). "An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges." In: *BenchCouncil transactions on benchmarks, standards and evaluations* 2.4, p. 100089. DOI: 10.1016/j.tbench.2023.100089.

Halliday, Michael Alexander Kirkwood (1994). "Spoken and written modes of meaning." In: *Media texts: Authors and readers* 7, pp. 51–73. DOI: 10.1163/9789004653436_006.

Haque, Mubin Ul, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad (2022). *"I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*. arXiv: 2212.05856 [cs.CL]. URL: https://arxiv.org/abs/2212.05856.

Hariri, Walid (2024). *Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Fu-*

*ture Directions in Natural Language Processing*. arXiv: 2304.02017 [cs.CL]. URL: https://arxiv.org/abs/2304.02017.

Hart, Christopher (2008). "Critical discourse analysis and metaphor: Toward a theoretical framework." In: *Critical discourse studies* 5.2, pp. 91–106. DOI: 10.1080/17405900801990058.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp (2023). "More than a Feeling: Accuracy and Application of Sentiment Analysis." In: *International Journal of Research in Marketing* 40.1, pp. 75–87. ISSN: 0167-8116. DOI: https://doi.org/10.1016/j.ijresmar.2022.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0167811622000477.

Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation*. arXiv: 2301.01768 [cs.CL]. URL: https://arxiv.org/abs/2301.01768.

Hassani, Hossein and Emmanuel Sirmal Silva (2023). "The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field." In: *Big data and cognitive computing* 7.2, p. 62. DOI: 10.3390/bdcc7020062.

Heaton, Dan, Jérémie Clos, Elena Nichele, and Joel E Fischer (2023a). "The Social Impact of Decision-Making Algorithms: Reviewing the Influence of Agency, Responsibility and Accountability on Trust and Blame." In: *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, pp. 1–11. DOI: 10.1145/3597512.3599706.

Heaton, Dan, Jeremie Clos, Elena Nichele, and Joel E Fischer (2024a). ""The ChatGPT bot is causing panic now–but it'll soon be as mundane a tool as Excel": analysing topics, sentiment and emotions relating to ChatGPT on Twitter." In: *Personal and Ubiquitous Computing*, pp. 1–20. DOI: 10.1007/s00779-024-01811-x.

Heaton, Dan, Jeremie Clos, Elena Nichele, and Joel Fischer (2023b). "Critical reflections on three popular computational linguistic ap-

proaches to examine Twitter discourses." In: *PeerJ Computer Science* 9, e1211. DOI: 10.7717/peerj-cs.1211.

Heaton, Dan, Elena Nichele, Jeremie Clos, and Joel E. Fischer (July 2023c). ""The algorithm will screw you": Blame, social actors and the 2020 A Level results algorithm on Twitter." In: *PLOS ONE* 18.7, pp. 1–29. DOI: 10.1371/journal.pone.0288662. URL: https://doi.org/10.1371/journal.pone.0288662.

Heaton, Dan, Elena Nichele, Jeremie Clos, and Joel E Fischer (2023d). ""The pingdemic has been a greater challenge than Covid itself": semantic prosodies in UK newspaper articles during the pandemic." In: *SN Social Sciences* 3.9, p. 146. DOI: 10.1007/s43545-023-00740-5.

Heaton, Dan, Elena Nichele, Jérémie Clos, and Joel E Fischer (2024b). "Perceptions of the Agency and Responsibility of the NHS COVID-19 App on Twitter: Critical Discourse Analysis." In: *J Med Internet Res* 26, e50388. ISSN: 1438-8871. DOI: 10.2196/50388. URL: https://www.jmir.org/2024/1/e50388.

Heaton, Dan, Elena Nichele, Jeremie Clos, and Joel E Fischer (2024c). ""ChatGPT says no": agency, trust, and blame in Twitter discourses after the launch of ChatGPT." In: *AI and Ethics*, pp. 1–23. DOI: 10.1007/s43681-023-00414-1.

Hecht, Yannique (2020). *UK's Failed Attempt to Grade Students by an Algorithm — pub.towardsai.net*. https://pub.towardsai.net/ofqual-algorithm-5ecbe950c264. [Accessed 18-02-2021].

Hidayatullah, Ahmad Fathan, Silfa Kurnia Aditya, Syifa Tri Gardini, et al. (2019). "Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA)." In: 482.1, p. 012033. DOI: 10.1088/1757-899x/482/1/012033.

Hoey, Michael (2007). "Grammatical creativity: A corpus perspective." In: Bloomsbury, pp. 31–56. URL: http://digital.casalini.it/9781441194855.

Holford, W David (2022). "'Design-for-responsible'algorithmic decision-making systems: a question of ethical judgement and

human meaningful control." In: *AI and Ethics* 2.4, pp. 827–836. DOI: 10.1007/s43681-022-00144-w.

Holmes, Janet and Nick Wilson (2022). *An introduction to Sociolinguistics*. Routledge.

Howard, Andrea L (2021). "A guide to visualizing trajectories of change with confidence bands and raw data." In: *Advances in Methods and Practices in Psychological Science* 4.4, p. 25152459211047228. DOI: 10.1177/25152459211047228.

Hu, Andrea, Stevie Chancellor, and Munmun De Choudhury (2019). "Characterizing Homelessness Discourse on Social Media." In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Chi Ea '19. Glasgow, Scotland Uk: Association for Computing Machinery, 1–6. ISBN: 9781450359719. DOI: 10.1145/3290607.3313057. URL: https://doi.org/10.1145/3290607.3313057.

Hu, Xiaomeng, Pin-Yu Chen, and Tsung-Yi Ho (2023). "Radar: Robust ai-text detection via adversarial learning." In: *Advances in Neural Information Processing Systems* 36, pp. 15077–15095. DOI: 10.5555/3666122.3666784.

Hudson, Richard A (1996). *Sociolinguistics*. Cambridge University Press.

Hunston, Susan (2010). "How can a corpus be used to explore patterns?" In: *The Routledge handbook of corpus linguistics*. Routledge, pp. 140–154.

Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550.

Jacob, Steve and Justin Lawarée (2021). "The adoption of contact tracing applications of COVID-19 by European governments." In: *Policy Design and Practice* 4.1, pp. 44–58. DOI: 10.1080/25741292.2020.1850404.

Jasanoff, S (2020). "Temptations of technocracy in the century of engineering." In: *The Bridge* 50.supplement, pp. 8–10.

Jaworska, Sylvia (2017). "Corpus approaches: Investigating linguistic patterns and meanings." In: *The Routledge handbook of language and media*. Routledge, pp. 93–108.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao (2019). "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." In: *Multimedia tools and applications* 78, pp. 15169–15211. DOI: 10.1007/s11042-018-6894-4.

Jiang, Jialun "Aaron", Jed R. Brubaker, and Casey Fiesler (2017). "Understanding Diverse Interpretations of Animated GIFs." In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Chi Ea '17. Denver, Colorado, USA: Association for Computing Machinery, 1726–1732. ISBN: 9781450346566. DOI: 10.1145/3027063.3053139. URL: https://doi.org/10.1145/3027063.3053139.

Jiang, Weijie and Zachary A Pardos (2021). "Towards equity and algorithmic fairness in student grade prediction." In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 608–617. DOI: 10.1145/3461702.3462623.

Jianqiang, Z. (2015). "Pre-processing Boosting Twitter Sentiment Analysis?" In: *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 748–753. DOI: 10.1109/SmartCity.2015.158.

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "The global landscape of AI ethics guidelines." In: *Nature Machine Intelligence* 1.9, pp. 389–399. DOI: 10.1038/s42256-019-0088-2.

Johnson, Melissa N. P., Ethan McLean, and Audrey Kobayashi (2020). "Discourse Analysis." In: *International Encyclopedia of Human Geography (Second Edition)*. Oxford: Elsevier, pp. 377–383. DOI: 10.1016/b978-0-08-102295-5.10814-5. URL:

https://www.sciencedirect.com/science/article/pii/B9780081022955108145.

Kalla, Dinesh and Nathan Smith (2023). "Study and Analysis of Chat GPT and its Impact on Different Fields of Study." In: *International Journal of Innovative Science and Research Technology* 8.3. URL: https://ssrn.com/abstract=4402499.

Kapidzic, Sanja, Christoph Neuberger, Stefan Stieglitz, and Milad Mirbabaie (2019). "Interaction and Influence on Twitter: Comparing the discourse relationships between user types on five topics." In: *Digital Journalism* 7.2, pp. 251–272. ISSN: 2167-0811. DOI: 10.1080/21670811.2018.1522962.

Kellerman, Aharon (2023). "Chatbots and information mobility: An agenda for thought and study." In: *Environment and Planning B: Urban Analytics and City Science*, p. 23998083231181595. DOI: 10.1177/23998083231181595.

Kelly, Anthony (2021). "A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020." In: *British Educational Research Journal* 47.3, pp. 725–741. DOI: 10.1002/berj.3705.

Kelly, Samantha Murphy (2023). *This AI chatbot is dominating social media with its frighteningly good essays | CNN business*. [Accessed 18-04-2023]. URL: https://edition.cnn.com/2022/12/05/tech/chatgpt-trnd/index.html.

Kendall, Gavin et al. (2007). "What is critical discourse analysis?" In: *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*. Vol. 8. 2. DOI: 10.17169/fqs-8.2.255.

Kennedy, Graeme (2014). *An introduction to corpus linguistics*. Routledge. DOI: 10.4324/9781315843674.

Kent, C (2020). *A comedy of errors: the UK's contact-tracing apps*. URL: https://www.medicaldevice-network.com/features/uk-contact-tracing-app-problems/.

Khalil, Mohammad and Erkan Er (2023). "Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection." In: *International*

*Conference on Human-Computer Interaction*. Springer, pp. 475–487. DOI: 10.1007/978-3-031-34411-4_32.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell (2008). "The Sketch Engine." In: *Practical Lexicography: A Reader*. Citeseer, pp. 297–306.

Kirmani, Ahmad R (2022). "Artificial Intelligence-Enabled Science Poetry." In: *ACS Energy Letters* 8, pp. 574–576. DOI: 10.1021/acsenergylett.2c02758.

Kitishat, Amal Riyadh, Murad Al Kayed, and Mohammad Al-Ajalein (2020). "A Corpus-Assisted Critical Discourse Analysis of the Syrian Refugee Crisis in Jordanian Newspapers." In: *International Journal of English Linguistics* 10.6. DOI: 10.5539/ijel.v10n6p195.

Klein, Helen Altman, Mei-Hua Lin, Norma L Miller, Laura G Militello, Joseph B Lyons, and Jessica Grace Finkeldey (2019). "Trust across culture and context." In: *Journal of Cognitive Engineering and Decision Making* 13.1, pp. 10–29. DOI: 10.1177/1555343418810936.

Kocaballi, A. Baki (2023). *Conversational AI-Powered Design: ChatGPT as Designer, User, and Product*. arXiv: 2302.07406 [cs.HC]. URL: https://arxiv.org/abs/2302.07406.

Kochenderfer, Mykel J, Tim A Wheeler, and Kyle H Wray (2022). *Algorithms for decision making*. MIT press.

Kolkman, Daan (2020). "F** k the algorithm?: what the world can learn from the UK's A-level grading fiasco." In: *Impact of Social Sciences Blog*. URL: https://eprints.lse.ac.uk/106366/.

Koohang, Alex, Jeretta Horn Nord, Keng-Boon Ooi, Garry Wei-Han Tan, Mostafa Al-Emran, Eugene Cheng-Xi Aw, Abdullah Mohammed Baabdullah, Dimitrios Buhalis, Tat-Huei Cham, Charles Dennis, et al. (2023). "Shaping the metaverse into reality: a holistic multidisciplinary understanding of opportunities, challenges, and avenues for future investigation." In: *Journal of Computer Information Systems* 63.3, pp. 735–765. DOI: 10.1080/08874417.2023.2165197.

Kopaczyk, Joanna and Jukka Tyrkkö (2018). *Applications of Pattern-driven Methods in Corpus Linguistics*. Vol. 82. John Benjamins Publishing Company. DOI: 10.1075/scl.82.

Kopf, Susanne and Elena Nichele (2018). "Es-tu Charlie?" In: *Doing Politics: Discursivity, performativity and mediation in political discourse* 80, p. 211. URL: http://digital.casalini.it/9789027263148.

Koppell, Jonathan GS (2005). "Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder"." In: *Public administration review* 65.1, pp. 94–108. DOI: 10.1111/j.1540-6210.2005.00434.x.

Korkmaz, Adem, Cemal Aktürk, and Tarık Talan (2023). "Analyzing the User's Sentiments of ChatGPT Using Twitter Data." In: *Iraqi Journal For Computer Science and Mathematics* 4.2, pp. 202–214. DOI: 10.52866/ijcsm.2023.02.02.018.

Kreis, Ramona (2017). "# refugeesnotwelcome: Anti-refugee discourse on Twitter." In: *Discourse & Communication* 11.5, pp. 498–514. DOI: 10.1177/1750481317714121.

Kretzschmar, Mirjam E., Ganna Rozhnova, Martin C. J. Bootsma, Michiel van Boven, Janneke H. H. M. van de Wijgert, and Marc J. M. Bonten (2020). "Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study." In: *The Lancet Public Health* 5.8, e452–e459. ISSN: 2468-2667. DOI: 10.1016/S2468-2667(20)30157-2.

Kumar, Arun HS (2023). "Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain." In: *Biology, Engineering, Medicine and Science Reports* 9.1, pp. 24–30. DOI: 10.5530/bems.9.1.5.

Kumar, M Naveen and R Suresh (2012). "Emotion detection using lexical chains." In: *International Journal of Computer Applications* 57.4, pp. 1–4.

Kumar, Shamanth, Fred Morstatter, and Huan Liu (2014). *Twitter data analytics*. Springer. DOI: 10.1007/978-1-4614-9372-3.

Kwiatkowska, MMarta and Morteza Lahijanian (2016). "Social trust: a major challenge for the future of autonomous systems." In: *AAAI Fall Symposium on Cross-Disciplinary Challenges for Autonomous Systems*. Association for the Advancement of Artificial Intelligence.

Laitinen, Arto and Otto Sahlgren (2021). "AI systems and respect for human autonomy." In: *Frontiers in artificial intelligence* 4, p. 151. DOI: 10.3389/frai.2021.705164.

Lamanna, Camillo and Lauren Byrne (2018). "Should artificial intelligence augment medical decision making? The case for an autonomy algorithm." In: *AMA journal of ethics* 20.9, pp. 902–910. DOI: 10.1001/amajethics.2018.902..

Lamsal, Rabindra (2021). "Design and analysis of a large-scale COVID-19 tweets dataset." In: *Applied Intelligence* 51.5, pp. 2790–2804. DOI: 10.1007/s10489-020-02029-z.

Lang, Daniel J, Arnim Wiek, Matthias Bergmann, Michael Stauffacher, Pim Martens, Peter Moll, Mark Swilling, and Christopher J Thomas (2012). "Transdisciplinary research in sustainability science: practice, principles, and challenges." In: *Sustainability science* 7, pp. 25–43. DOI: 10.1007/s11625-011-0149-x.

Lee, Min Kyung (2018). "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management." In: *Big Data & Society* 5.1, p. 2053951718756684. DOI: 10.1177/2053951718756684.

Lee, Nicol Turner, Paul Resnick, and Genie Barton (2019). "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms." In: *Brookings Institute: Washington, DC, USA* 2. URL: https://policycommons.net/artifacts/4141276/algorithmic-bias-detection-and-mitigation/4949849/.

Lee, Sang-Woong, Sadaf Hussain, Ghassan F Issa, Sagheer Abbas, Taher M Ghazal, Tanweer Sohail, Munir Ahmad, and Muhammad Adnan Khan (2021). "Multi-dimensional trust quantification by artificial agents through evidential fuzzy multi-criteria deci-

sion making." In: *IEEE Access* 9, pp. 159399–159412. DOI: 10.1109/ACCESS.2021.3131521.

Leiter, Christoph, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger (2024). "Chatgpt: A meta-analysis after 2.5 months." In: *Machine Learning with Applications* 16, p. 100541. DOI: 10.1016/j.mlwa.2024.100541.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck (2018). "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges." In: *Philosophy & Technology* 31, pp. 611–627. DOI: 10.1007/s13347-017-0279-x.

Leslie, Alan M (1993). *A theory of agency*. Citeseer. URL: https://psycnet.apa.org/record/1995-98256-005.

Leung, Yvonne W et al. (2021). "Natural language processing–based virtual cofacilitator for online cancer support groups: Protocol for an algorithm development and validation study." In: *JMIR research protocols* 10.1, e21453. DOI: 10.2196/21453.

Li, Bing, Shiji Chen, and Vincent Larivière (2023). "Interdisciplinarity affects the technological impact of scientific research." In: *Scientometrics*, pp. 1–33. DOI: 10.1007/s11192-023-04846-8.

Liao, Shiyang, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng (2017). "CNN for situations understanding based on sentiment analysis of twitter data." In: *Procedia computer science* 111, pp. 376–381. DOI: 10.1016/j.procs.2017.06.037.

Liimatta, Aatu (2020). "Using lengthwise scaling to compare feature frequencies across text lengths on Reddit." In: *Corpus approaches to social media*, pp. 111–130. URL: http://digital.casalini.it/9789027260499.

Liu, Bing (2010). "Sentiment analysis and subjectivity." In: *Handbook of natural language processing* 2.2010, pp. 627–666.

Liu, Bing and Lei Zhang (2012). "A survey of opinion mining and sentiment analysis." In: *Mining text data*. Springer, pp. 415–463. DOI: 10.1007/978-1-4614-3223-4_13.

Loria, Steven (2018). *textblob Documentation*. URL: https://readthedocs.org/projects/textblob/downloads/pdf/dev/.

Love, Robbie (2021). "Swearing in informal spoken English: 1990s–2010s." In: *Text & Talk* 41.5-6, pp. 739–762.

Lu, Yujie, Kotaro Sakamoto, Hideyuki Shibuki, and Tatsunori Mori (2017). "Are deep learning methods better for twitter sentiment analysis." In: *Proceedings of the 23rd annual meeting of natural language processing (Japan)*, pp. 787–790.

Luhmann, Niklas (1979). *Trust and Power*. John A. Wiley and Sons.

Lyall, Catherine (2019). *Being an interdisciplinary academic: How institutions shape university careers*. Springer. DOI: 10.1007/978-3-030-18659-3.

Lyons, Joseph B., Matthew A. Clark, Alan R. Wagner, and Matthew J. Schuelke (2017). "Certifiable trust in autonomous systems: Making the intractable tangible." In: *AI Magazine* 38.3, pp. 37–49. ISSN: 2371-9621. DOI: 10.1609/aimag.v38i3.2717.

Maclean, S (2016). "A new model for social work reflection: whatever the weather." In: *Professional Social Work* 1, pp. 28–29.

Madhavan, Poornima and Douglas A. Wiegmann (2007). "Similarities and differences between human-human and human-automation trust: an integrative review." In: *Theoretical Issues in Ergonomics Science* 8.4, pp. 277–301. ISSN: 1463-922X. DOI: 10.1080/14639220500337708.

Mahmud, Hasan, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander (2022). "What influences algorithmic decision-making? A systematic literature review on algorithm aversion." In: *Technological Forecasting and Social Change* 175, p. 121390. DOI: 10.1016/j.techfore.2021.121390.

Maier, Daniel, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. (2018). "Applying LDA topic modeling in communication research: Toward a valid and reli-

able methodology." In: *Communication Methods and Measures* 12.2-3, pp. 93–118.

Mallett, Bruno (2023). "Reviewing the impact of OFQUAL's assessment 'algorithm'on racial inequalities." In: *COVID-19 and Racism*. Policy Press, pp. 187–198. DOI: 10.51952/9781447366751.ch012.

Marková, Ivana (2003). *Dialogicality and social representations: The dynamics of mind*. Cambridge University Press. URL: https://psycnet.apa.org/record/2004-00023-000.

Marsh, Kimberly, Emily Griffiths, Johanna J Young, Carrie-Anne Gibb, and Jim McMenamin (2021). "Contributions of the EURO 2020 football championship events to a third wave of SARS-CoV-2 in Scotland, 11 June to 7 July 2021." In: *Eurosurveillance* 26.31, p. 2100707. DOI: 10.2807/1560-7917.ES.2021.26.31.2100707.

Mason, Shannon and Lenandlar Singh (2022). "Reporting and discoverability of "Tweets" quoted in published scholarship: current practice and ethical implications." In: *Research Ethics* 18.2, pp. 93–113. DOI: 10.1177/17470161221076948.

Mathur, Amrita, Purnima Kubde, and Sonali Vaidya (2020). "Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19." In: *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, pp. 845–848. DOI: 10.1109/ICCES48766.2020.9138079.

Mautner, Gerlinde (2007). "Mining large corpora for social information: The case of elderly." In: *Language in Society* 36.1, pp. 51–72. DOI: 10.1017/S0047404507070030.

Mayer, Roger C, James H Davis, and F David Schoorman (1995). "An integrative model of organizational trust." In: *Academy of management review* 20.3, pp. 709–734. DOI: 10.5465/amr.1995.9508080335.

Mbwogge, Mathew (2021). *Mass Testing With Contact Tracing Compared to Test and Trace for the Effective Suppression of COVID-19 in the United Kingdom: Systematic Review*. DOI: 10.2196/27254. URL: http://www.ncbi.nlm.nih.gov/pubmed/33857269.

McCormick, Tyler H, Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma S Spiro (2017). "Using Twitter for demographic and social science research: Tools for data collection and processing." In: *Sociological methods & research* 46.3, pp. 390–421. DOI: 10.1177/0049124115605339.

McEnery, Tony and Andrew Hardie (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. DOI: 10.1093/oxfordhb/9780199276349.013.0024.

McGlashan, Mark (2020). "Collective identity and discourse practice in the followership of the Football Lads Alliance on Twitter." In: *Discourse & Society* 31.3, pp. 307–328. DOI: 10.1177/0957926519889128.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy (2014). "Sentiment analysis algorithms and applications: A survey." In: *Ain Shams Engineering Journal* 5.4, pp. 1093–1113. ISSN: 2090-4479. DOI: 10.1016/j.asej.2014.04.011. URL: https://www.sciencedirect.com/science/article/pii/S2090447914000550.

Meisner, Colten, Brooke Erin Duffy, and Malte Ziewitz (2024). "The labor of search engine evaluation: Making algorithms more human or humans more algorithmic?" In: *New Media & Society* 26.2, pp. 1018–1033. DOI: 10.1177/14614448211063860.

Meyerson, Debra, Karl E Weick, Roderick M Kramer, et al. (1996). "Swift trust and temporary groups." In: *Trust in organizations: Frontiers of theory and research* 166, p. 195. DOI: doi.org/10.4135/9781452243610.

Michaux, Clarisse (2023). *Can Chat GPT Be Considered an Author? I Met with Chat GPT and Asked Some Questions About Philosophy of Art and Philosophy of Mind*. DOI: http://dx.doi.org/10.2139/ssrn.4439607.

Miller, Dale T (2001). "Disrespect and the experience of injustice." In: *Annual Review of Psychology* 52.1, pp. 527–553. DOI: 10.1146/annurev.psych.52.1.527.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi (2016). "The ethics of algorithms: Mapping the debate." In: *Big Data & Society* 3.2, p. 2053951716679679. DOI: 10.1177/2053951716679679.

Mogashoa, Tebogo (2014). "Understanding critical discourse analysis in qualitative research." In: *International Journal of Humanities Social Sciences and Education* 1.7, pp. 104–113.

Mohammad, Saif M and Peter D Turney (2013). "Crowdsourcing a word–emotion association lexicon." In: *Computational intelligence* 29.3, pp. 436–465. DOI: 10.1111/j.1467-8640.2012.00460.x.

Mohseni, Sina, Niloofar Zarei, and Eric D Ragan (2021). "A multidisciplinary survey and framework for design and evaluation of explainable AI systems." In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4, pp. 1–45. DOI: 10.1145/3387166.

Morales, Manuel et al. (2021). *COVID-19 Tests Gone Rogue: Privacy, Efficacy, Mismanagement and Misunderstandings*. arXiv: 2101.01693 [cs.CY]. URL: https://arxiv.org/abs/2101.01693.

Morgan, Angela (2010). "Discourse analysis: An overview for the neophyte researcher." In: *Journal of Health and Social Care Improvement* 1.1, pp. 1–7.

Morris, Michael W, Oliver J Sheldon, Daniel R Ames, and Maia J Young (2007). "Metaphors and the market: Consequences and preconditions of agent and object metaphors in stock market commentary." In: *Organizational behavior and human decision processes* 102.2, pp. 174–192. DOI: 10.1016/j.obhdp.2006.03.001.

Mpofu, Phillip (2022). "Indigenous media and social media convergence: Adaptation of storytelling on Twitter, SoundCloud and YouTube in Zimbabwe." In: *Journal of Asian and African Studies* 57.6, pp. 1199–1213. DOI: 10.1177/00219096211049176.

Muir, Bonnie M. (1987). "Trust between humans and machines, and the design of decision aids." In: *International journal of man-machine studies* 27.5-6, pp. 527–539. ISSN: 0020-7373. DOI: 10.1016/S0020-7373(87)80013-5.

Mulgan, Richard (2000). "'Accountability': an ever-expanding concept?" In: *Public administration* 78.3, pp. 555–573. DOI: 10.1111/1467-9299.00218.

Munch, Lauritz, Jakob Mainz, and Jens Christian Bjerring (2023). "The value of responsibility gaps in algorithmic decision-making." In: *Ethics and Information Technology* 25.1, p. 21. DOI: 10.1007/s10676-023-09699-6.

Mustaqim, T, Tanzilal Mustaqim, K Umam, M A Muslim, and Much Aziz Muslim (2020). "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm." In: *Journal of Physics: Conference Series*. DOI: 10.1088/1742-6596/1567/3/032024.

NHS England (2021). *Data for Contact Tracing*. URL: https://www.covid19.nhs.uk/privacy-and-data/data-for-contact-tracing.html.

Najafali, Daniel, Justin M Camacho, Erik Reiche, Logan G Galbraith, Shane D Morrison, and Amir H Dorafshar (2023). "Truth or lies? The pitfalls and limitations of ChatGPT in systematic review creation." In: *Aesthetic Surgery Journal* 43.8, pp. 654–655. DOI: 10.1093/asj/sjad093.

Nartey, Mark and Isaac N Mwinlaaru (2019). "Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis." In: *Corpora* 14.2, pp. 203–235. DOI: 10.3366/cor.2019.0169.

Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters (2020). "How we do things with words: Analyzing text as social and cultural data." In: *Frontiers in Artificial Intelligence* 3, p. 62. DOI: 10.3389/frai.2020.00062.

Nikolenko, Sergey I., Sergei Koltcov, and Olessia Koltsova (Feb. 2017a). "Topic Modelling for Qualitative Studies." In: *J. Inf. Sci.*

43.1, 88–102. ISSN: 0165-5515. DOI: 10.1177/0165551515617393. URL: https://doi.org/10.1177/0165551515617393.

Nikolenko, Sergey I, Sergei Koltcov, and Olessia Koltsova (2017b). "Topic modelling for qualitative studies." In: *Journal of Information Science* 43.1, pp. 88–102. DOI: 10.1177/0165551515617393.

Norton, Bonny and Kelleen Toohey (2011). "Identity, language learning, and social change." In: *Language teaching* 44.4, pp. 412–446.

Nowotny, Helga (2021). *In AI we trust: Power, illusion and control of predictive algorithms*. John Wiley & Sons.

Nugraha, Intan Siti, Eva Tuckyta S Sujatna, and Sutiono Mahdi (2021). "Corpus Linguistic Study of Tweets Using #CHARLIEHEBDO Hashtag." In: *JALL (Journal of Applied Linguistics and Literacy)* 5.1, pp. 54–70. DOI: 10.25157/jall.v5i1.4965.

Ofqual (2020). *Awarding GCSE, AS & A levels in summer 2020: interim report*. URL: https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

Oktar, Lütfiye (2001). "The ideological organization of representational processes in the presentation of us and them." In: *Discourse & Society* 12.3, pp. 313–346. DOI: 10.1177/0957926501012003003.

Olhede, Sofia and Patrick J Wolfe (2020). "Blame the algorithm?" In: *Significance* 17.5, pp. 12–12. DOI: 10.1111/1740-9713.01441.

Olson, Hope (1995). "Quantitative" versus" qualitative research: The wrong question." In: *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*. DOI: 10.29173/cais414.

Osei Fordjour, Nana Kwame (2021). "A multimodal social semiotic analysis of an African Vice President on Twitter." In: *Visual Communication Quarterly* 28.4, pp. 227–239. DOI: 10.1080/15551393.2021.1986829.

Ottovordemgentschenfelde, Svenja (2017). "'Organizational, professional, personal': An exploratory study of political journalists and their hybrid brand on Twitter." In: *Journalism* 18.1, pp. 64–80. DOI: 10.1177/1464884916657524.

Pak, Alexander, Patrick Paroubek, et al. (2010). "Twitter as a corpus for sentiment analysis and opinion mining." In: *LREc*. Vol. 10. 2010, pp. 1320–1326.

Palmer, Carole L (1999). "Structures and strategies of interdisciplinary science." In: *Journal of the American society for information science* 50.3, pp. 242–253. DOI: 10.1002/(SICI)1097-4571(1999)50:3<242::AID-ASI7>3.0.CO;2-7.

— (2013). *Work at the boundaries of science: Information and the interdisciplinary research process*. Springer Science & Business Media. DOI: 10.1007/978-94-015-9843-9.

Panagiotopoulos, Panos, Bram Klievink, and Antonio Cordella (2019). "Public value creation in digital government." In: *Government Information Quarterly* 36.4, p. 101421. DOI: 10.1016/j.giq.2019.101421.

Pandit, Jay A, Jennifer M Radin, Giorgio Quer, and Eric J Topol (2022). "Smartphone apps in the COVID-19 pandemic." In: *Nature Biotechnology* 40.7, pp. 1013–1022. DOI: 10.1038/s41587-022-01350-x.

Park, Jaehyuk, Giovanni Luca Ciampaglia, and Emilio Ferrara (2016). "Style in the Age of Instagram: Predicting Success within the Fashion Industry Using Social Media." In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. CSCW '16. San Francisco, California, USA: Association for Computing Machinery, 64–73. ISBN: 9781450335928. DOI: 10.1145/2818048.2820065. URL: https://doi.org/10.1145/2818048.2820065.

Pasquale, Frank (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. DOI: 10.4159/harvard.9780674736061.c8.

Paucar, Luis H Garcia, Nelly Bencomo, Alistair Sutcliffe, and Pete Sawyer (2022). "A Bayesian network-based model to understand the role of soft requirements in technology acceptance: the case of the NHS COVID-19 test and trace app in England and Wales." In:

*Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pp. 1327–1336. DOI: 10.1145/3477314.3507147.

Peeters, Rik (2020). "The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making." In: *Information Polity* 25.4, pp. 507–522. DOI: 10.3233/IP-200253.

Peñalvo, Francisco José García (2023). "The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic?" In: *Education in the knowledge society (EKS)* 24, p. 1. DOI: 10.14201/eks.31279.

Peng, Roger D and Elizabeth Matsui (2016). *The Art of Data Science: A guide for anyone who works with Data*. Skybrude consulting LLC.

Pepper, Cecily, Gisela Reyes-Cruz, Ana Rita Pena, Liz Dowthwaite, Camilla M Babbage, Hanne Wagner, Elena Nichele, Joel E Fischer, et al. (2022). "Understanding Trust and Changes in Use After a Year With the NHS COVID-19 Contact Tracing App in the United Kingdom: Longitudinal Mixed Methods Study." In: *Journal of Medical Internet Research* 24.10, e40558. DOI: 10.2196/40558.

Petrović, Vladimir M (2018). "Artificial intelligence and virtual worlds–toward human-level ai agents." In: *IEEE Access* 6, pp. 39976–39988. DOI: 10.1109/ACCESS.2018.2855970.

Pettit, Philip (2001). *A theory of freedom: from the psychology to the politics of agency*. Oxford University Press on Demand.

Phillips, Peter and Tony Cassidy (2024). "Social representations and symbolic coping: a cross-cultural discourse analysis of the covid-19 pandemic in newspapers." In: *Health Communication* 39.3, pp. 451–459. DOI: 10.1080/10410236.2023.2169300.

Picazo-Vela, Sergio, Isis Gutiérrez-Martínez, and Luis Felipe Luna-Reyes (2012). "Understanding risks, benefits, and strategic alternatives of social media applications in the public sector." In: *Government information quarterly* 29.4, pp. 504–511. DOI: 10.1016/j.giq.2012.07.002.

Piorkowski, David, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy (2021). "How ai developers overcome communication challenges in a multidisciplinary team: A case study." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, pp. 1–25. DOI: 10.1145/3449205.

Pokharel, Bishwo Prakash (2020). "Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal." In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3624719.

Post, Ger, Vincent Visser, and Joris Buis (2017). "13. Reflection." In: *Academic Skills for Interdisciplinary Studies*. Amsterdam University Press, pp. 116–123.

Prihatini, P M, Pantjawarni Prihatini, I K Suryawan, and IN Mandia (2018). "Feature extraction for document text using Latent Dirichlet Allocation." In: *Journal of Physics: Conference Series*. DOI: 10.1088/1742-6596/953/1/012047.

Radomska, Joanna, Przemysław Wołczek, Letycja Sołoducho-Pelc, and Susana Silva (2019). "The impact of trust on the approach to management—A case study of creative industries." In: *Sustainability* 11.3, p. 816. DOI: 10.3390/su11030816.

Rathore, Bharati (2023). "Future of AI & Generation Alpha: ChatGPT beyond Boundaries." In: *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal* 12.1, pp. 63–68.

Ray, Abhishek, Hossein Ghasemkhani, and Cesar Martinelli (2024). "Competition and Cognition in the Market for Online News." In: *Journal of Management Information Systems* 41.2, pp. 367–393. DOI: 10.1080/07421222.2024.2340824.

Ray, Partha Pratim (2023). "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope." In: *Internet of Things and Cyber-Physical Systems* 3, pp. 121–154. DOI: 10.1016/j.iotcps.2023.04.003.

Razis, Gerasimos, Ioannis Anagnostopoulos, and Petr Saloun (2016). "Thematic labeling of Twitter accounts using DBpedia properties." In: *2016 11th International Workshop on Semantic and Social*

*Media Adaptation and Personalization (SMAP)*. IEEE, pp. 106–111. DOI: 10.1109/SMAP.2016.7753393.

Reeves, Byron and Clifford Nass (1997). "The Media Equation: How People Treat Computers, Television,? New Media Like Real People? Places." In: *Computers and Mathematics with Applications* 5.33, p. 128.

Reeves, Scott, Ayelet Kuper, and Brian David Hodges (2008). "Qualitative research methodologies: ethnography." In: *BMJ* 337. DOI: 10.1136/bmj.a1020.

Rehůřek, Radim and Petr Sojka (2011). *Gensim–statistical semantics in python*. URL: https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh (2020). *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*. DOI: 10.18653/v1/2020.acl-main.442.

Richardson, Peter, Charles M Mueller, and Stephen Pihlaja (2021). *Cognitive linguistics and religious language: an introduction*. Routledge. DOI: 10.4324/9781003041139.

Riegler, Carolyn (2019). "The Moral Decision-Making Capacity of Self-Driving Cars: Socially Responsible Technological Development, Algorithm-driven Sensing Devices, and Autonomous Vehicle Ethics." In: *Contemp. Readings L. & Soc. Just.* 11, p. 15. DOI: 10.22381/CRLSJ11120192.

Rimmer, Abi (2021). "Sixty seconds on... the pingdemic." In: *BMJ (Clinical Research ed.)* 374, n1822–n1822. DOI: 10.1136/bmj.n1822.

Rish, Irina et al. (2001). "An empirical study of the naive Bayes classifier." In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22, pp. 41–46.

Roesslein, Joshua (2009). *tweepy Documentation*. URL: http://tweepy.readthedocs.io/en/v3.

Rogers, Everett M (1995). "Diffusion of Innovations: modifications of a model for telecommunications." In: *Die diffusion von innovatio-*

*nen in der telekommunikation*, pp. 25–38. DOI: 10.1007/978-3-642-79868-9_2.

Roose, Kevin (2022). "The brilliance and weirdness of ChatGPT." In: *The New York Times* 5.

— (2023). "GPT-4 is exciting and scary." In: *The New York Times* 15.

Rosamond, Emily (2020). "What Was to Have Happened? Tenses for a Cancelled Future." In: *Metropolis M* 2020.6. URL: https://research.gold.ac.uk/id/eprint/29358.

Rose, Elizabeth Robertson (2017). "A Month of Climate Change in Australia: A Corpus-Driven Analysis of Media Discourse." In: *Text-Based Research and Teaching*. Springer, pp. 37–53. DOI: 10.1057/978-1-137-59849-3_3.

Ross, Lee (1977). "The intuitive psychologist and his shortcomings: Distortions in the attribution process." In: *Advances in experimental social psychology*. Vol. 10. Elsevier, pp. 173–220. DOI: 10.1016/S0065-2601(08)60357-3.

Rousseau, Denise M, Sim B Sitkin, Ronald S Burt, and Colin Camerer (1998). "Not so different after all: A cross-discipline view of trust." In: *Academy of management review* 23.3, pp. 393–404. DOI: 10.5465/amr.1998.926617.

Rout, Jitendra Kumar, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena, and Karen L. Williams (2018). "A model for sentiment and emotion analysis of unstructured social media text." In: *Electronic Commerce Research* 18.1, pp. 181–199. ISSN: 1572-9362. DOI: 10.1007/s10660-017-9257-8.

Royal Academy of Engineering (2017). *Algorithms in decision-making*. URL: https://www.raeng.org.uk/publications/responses/algorithms-in-decision-making.

Rozin, Paul, Maureen Markwith, and Caryn Stoess (1997). "Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust." In: *Psychological science* 8.2, pp. 67–73. DOI: 10.1111/j.1467-9280.1997.tb00685.x.

Rubel, Alan, Clinton Castro, and Adam Pham (2020). "Algorithms, agency, and respect for persons." In: *Social theory and practice*, pp. 547–572. DOI: 10.5840/soctheorpract202062497.

Rudolph, Jürgen, Shannon Tan, and Samson Tan (2023). "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education." In: *Journal of Applied Learning and Teaching* 6.1, pp. 364–389. DOI: 10.37074/jalt.2023.6.1.23.

Russo, Katherine E and Arianna Grasso (2022). "Coping with dis-/ableism in Twitter discourse: A corpus-based critical appraisal analysis of the Hidden Disabilities Sunflower Lanyard case." In: *International Journal of Language Studies* 16.4.

Saggu, Aman and Lennart Ante (2023). "The influence of ChatGPT on artificial intelligence related crypto assets: Evidence from a synthetic control analysis." In: *Finance Research Letters*, p. 103993. DOI: 10.1016/j.frl.2023.103993.

Sailunaz, Kashfia, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj (2018). "Emotion detection from text and speech: a survey." In: *Social Network Analysis and Mining* 8.1, pp. 1–26. DOI: https://doi.org/10.1007/s13278-018-0505-2.

Samuel, Gabrielle, SL Roberts, A Fiske, F Lucivero, S McLennan, A Phillips, S Hayes, and SB Johnson (2021). "COVID-19 contact tracing apps: UK public perceptions." In: *Critical Public Health*, pp. 1–13. DOI: 10.1080/09581596.2021.1909707.

Sanderson, Katharine (2023). "GPT-4 is here: what scientists think." In: *Nature* 615.7954, p. 773. DOI: 10.1038/d41586-023-00816-5.

Saura, Jose Ramon, Domingo Ribeiro-Soriano, and Pablo Zegarra Saldaña (2022). "Exploring the challenges of remote work on Twitter users' sentiments: From digital technology development to a post-pandemic era." In: *Journal of Business Research* 142, pp. 242–254. DOI: 10.1016/j.jbusres.2021.12.052.

Schedler, Andreas, Larry Jay Diamond, and Marc F Plattner (1999). *The self-restraining state: power and accountability in new democracies*. Lynne Rienner Publishers.

Schiffrin, Deborah (2001). "Discourse markers: Language, meaning, and context." In: *The handbook of discourse analysis* 1, pp. 54–75. DOI: 10.1002/9780470753460.

Schoenherr, Jordan Richard and Robert Thomson (2024). "When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions." In: *IEEE Transactions on Technology and Society*. DOI: 10.1109/TTS.2024.3370095.

Schofield, Alexandra and David Mimno (2016). "Comparing Apples to Apple: The Effects of Stemmers on Topic Models." In: *Transactions of the Association for Computational Linguistics* 4, pp. 287–300. DOI: 10.1162/tacl_a_00099. URL: https://aclanthology.org/Q16-1021.

Schuetz, Peter NK (2021). "Fly in the Face of Bias: Algorithmic Bias in Law Enforcement's Facial Recognition Technology and the Need for an Adaptive Legal Framework." In: *Law & Ineq.* 39, p. 221. URL: https://scholarship.law.umn.edu/lawineq/vol39/iss1/8.

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi (2019). "Fairness and abstraction in sociotechnical systems." In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68. DOI: 10.1145/3287560.3287598.

Semino, Elena, Tara Coltman-Patel, William Dance, Alice Deignan, Zsófia Demjén, Claire Hardaker, and Alison Mackey (2024). "Narratives, information and manifestations of resistance to persuasion in online discussions of HPV vaccination." In: *Health Communication* 39.10, pp. 2123–2134.

Sengers, Phoebe, John McCarthy, and Paul Dourish (2006). "Reflective HCI: articulating an agenda for critical practice." In: *CHI'06 extended abstracts on Human factors in computing systems*, pp. 1683–1686. DOI: 10.1145/1125451.1125762.

Sengupta, Subhasree (2019). "What Are Academic Subreddits Talking About? A Comparative Analysis of r/Academia and r/Gradschool." In: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. CSCW '19. Austin, TX, USA: Association for Computing Machinery, 357–361. ISBN: 9781450366922. DOI: 10.1145/3311957.3359491. URL: https://doi.org/10.1145/3311957.3359491.

Shafeeg, Abdulla, Ilman Shazhaev, Dimitry Mihaylov, Arbi Tularov, and Islam Shazhaev (2023). "Voice Assistant Integrated with Chat GPT." In: *Indonesian Journal of Computer Science* 12.1, pp. 22–31. DOI: 10.33022/ijcs.v12i1.3146.

Shahrdad, Mehdi and Mehdi Chehel Amirani (2018). "Detection of preterm labor by partitioning and clustering the EHG signal." In: *Biomedical Signal Processing and Control* 45, pp. 109–116. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2018.05.044.

Shahrdar, Shervin, Luiza Menezes, and Mehrdad Nojoumian (2019). "A Survey on Trust in Autonomous Systems." In: *Intelligent Computing*. Ed. by Kohei Arai, Supriya Kapoor, and Rahul Bhatia. Cham: Springer International Publishing, pp. 368–386. DOI: 10.1007/978-3-030-01177-2_27.

Shijie, Song, Zhao Yuxiang, and Zhu Qinghua (2023). "From ELIZA to ChatGPT: AI-Generated Content (AIGC) Credibility Evaluation in Human-Intelligent Interactive Experience." In: *Information and Documentation Services* 44.4, pp. 35–42.

Siedlok, Frank and Paul Hibbert (2014). "The organization of interdisciplinary research: modes, drivers and barriers." In: *International Journal of Management Reviews* 16.2, pp. 194–210. DOI: 10.1111/ijmr.12016.

Silver, Crystal A, Benjamin W Tatler, Ramakrishna Chakravarthi, and Bert Timmermans (2021). "Social Agency as a continuum." In: *Psychonomic Bulletin & Review* 28.2, pp. 434–453. DOI: 10.3758/s13423-020-01845-1.

Sivalakshmi, P, P Udhaya Kumar, M Vasanth, R Srinath, and M Yokesh (2021). "COVID-19 vaccine public sentiment analysis using Python's Textblob approach." In: *International journal of current research and review* 13.11, pp. 166–172. DOI: 10.31782/ijcrr.2021.sp218.

Smith, Helen (2020). "Algorithmic bias: should students pay the price?" In: *AI & society* 35.4, pp. 1077–1078. ISSN: 1435-5655. DOI: 10.1007/s00146-020-01054-3.

Smith, Jenifer AE, Susan Hopkins, Charlie Turner, Kyle Dack, Anna Trelfa, Jerlyn Peh, and Paul S Monks (2022). "Public health impact of mass sporting and cultural events in a rising COVID-19 prevalence in England." In: *Epidemiology & Infection* 150, e42. DOI: 10.1017/S0950268822000188.

Song, SH, JY Min, HJ Kim, and KB Min (2019). "Topic modeling to mind illegal compensation for occupational injuries." In: *European Journal of Public Health* 29.Supplement 4, pp. 186–317. DOI: 10.1093/eurpub/ckz186.317.

Srinivasan, Balasubramanian and Kishan Mohan Kumar (2019). "Flock the similar users of twitter by using latent Dirichlet allocation." In: *Int. J. Sci. Technol. Res* 8, pp. 1421–1425.

Stine, Robert A (2019). "Sentiment analysis." In: *Annual Review of Statistics and Its Application* 6, pp. 287–308. DOI: 10.1146/annurev-statistics-030718-105242.

Suchomel, Vít (2020). "Better Web Corpora For Corpus Linguistics And NLP [online]." Dissertation. Masaryk University, Faculty of Informatics, Brno. URL: https://is.muni.cz/th/u4rmz/.

Sulalah, Anis (2020). "The semantic prosody analysis of 'increase' in COVID-19: A corpus-based study." In: *Lire Journal (Journal of Linguistics and Literature)* 4.2, pp. 237–246. DOI: 10.33019/lire.v4i2.92.

Sundar, S Shyam (2020). "Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII)." In:

*Journal of Computer-Mediated Communication* 25.1, pp. 74–88. DOI: 10.1093/jcmc/zmz026.

Sundar, S Shyam and Mengqi Liao (2023). "Calling BS on ChatGPT: Reflections on AI as a Communication Source." In: *Journalism & Communication Monographs* 25.2, pp. 165–180. DOI: 10.1177/15226379231167135.

Sveinson, Katherine and Rachel Allison (2021). ""Something Seriously Wrong With US Soccer": A Critical Discourse Analysis of Consumers' Twitter Responses to US Soccer's Girls' Apparel Promotion." In: *Journal of Sport Management* 1, pp. 1–13. DOI: 10.1123/jsm.2021-0127.

Symeonidis, Symeon, Dimitrios Effrosynidis, and Avi Arampatzis (2018). "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis." In: *Expert Systems with Applications* 110, pp. 298–310. DOI: 10.1016/j.eswa.2018.06.022.

Taddeo, Mariarosaria and Luciano Floridi (2018). "How AI can be a force for good." In: *Science* 361.6404, pp. 751–752. DOI: 10.1126/science.aat5991.

Taecharungroj, Viriya (2023). ""What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter." In: *Big Data and Cognitive Computing* 7.1, p. 35. DOI: 10.3390/bdcc7010035.

Tan, Felix B and Paul Sutherland (2004). "Online consumer trust: a multi-dimensional model." In: *Journal of Electronic Commerce in Organizations (JECO)* 2.3, pp. 40–58. DOI: 10.4018/jeco.2004070103.

Tang, Chien-Lin, Jingxian Liao, Hao-Chuan Wang, Ching-Ying Sung, Yu-Rong Cao, and Wen-Chieh Lin (2020). "Supporting Online Video Learning with Concept Map-Based Recommendation of Learning Path." In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Chi Ea '20. Honolulu, HI, USA: Association for Computing Machinery, 1–8. ISBN:

9781450368193. DOI: 10.1145/3334480.3382943. URL: https://doi.org/10.1145/3334480.3382943.

Teh, Yee, Michael Jordan, Matthew Beal, and David Blei (2004). "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes." In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press, pp. 1–8. URL: https://proceedings.neurips.cc/paper_files/paper/2004/file/fb4ab556bc42d6f0ee0f9e24ec4d1af0-Paper.pdf.

Tenorio, Encarnacion Hidalgo (2011). "Critical discourse analysis, an overview." In: *Nordic journal of English studies* 10.1, pp. 183–210. DOI: 10.35360/njes.247.

Tetlock, Philip E (1992). "The impact of accountability on judgment and choice: Toward a social contingency model." In: *Advances in experimental social psychology*. Vol. 25. Elsevier, pp. 331–376. DOI: 10.1016/S0065-2601(08)60287-7.

Timmins, Nicholas (2021). *Schools and coronavirus: The government's handling of education during the pandemic*. Tech. rep. Institute for Government.

Tiwary, Neelam, A Subaveerapandiyan, and A Vinoth (2023). *Netizens, Academicians, and Information Professionals' Opinions About AI With Special Reference To ChatGPT*. arXiv: 2302.07136 [cs.CY]. URL: https://arxiv.org/abs/2302.07136.

Tobi, Hilde and Jarl K Kampen (2018). "Research design: the methodology for interdisciplinary research framework." In: *Quality & quantity* 52, pp. 1209–1225. DOI: 10.1007/s11135-017-0513-8.

Tognini-Bonelli, Elena (2001). *Corpus linguistics at work*. Vol. 6. John Benjamins Publishing. URL: http://digital.casalini.it/9789027285447.

Tollon, Fabio (2023). "Responsibility gaps and the reactive attitudes." In: *AI and Ethics* 3.1, pp. 295–302. DOI: 10.1007/s43681-022-00172-6.

Tourish, Dennis and Owen Hargie (2012). "Metaphors of failure and the failures of metaphor: A critical study of root metaphors used

by bankers in explaining the banking crisis." In: *Organization Studies* 33.8, pp. 1045–1069. DOI: 10.1177/0170840612453528.

Tsoukias, Alexis (2021). "Social responsibility of algorithms: an overview." In: *EURO Working Group on DSS: A Tour of the DSS Developments Over the Last 30 Years*, pp. 153–166. DOI: 10.1007/978-3-030-70377-6_9.

Tucker, Gordon, Guowen Huang, Lise Fontaine, and Edward McDonald (2020). *Approaches to systemic functional grammar: convergence and divergence*. DOI: https://orca.cardiff.ac.uk/id/eprint/139878.

Turton, William (2017). *The algorithm is innocent*. https://theoutline.com/post/2362/the-algorithm-is-innocent. [Accessed 22-01-2023].

Ulfa, Maria Arista, Budi Irmawati, and Ario Yudo Husodo (2018). "Twitter Sentiment Analysis using Na¨ ive Bayes Classifier with Mutual Information Feature Selection." In: *Journal of Computer Science and Informatics Engineering (J-Cosine)* 2.2, pp. 106–111. DOI: 10.29303/jcosine.v2i2.120.

Van Dijk, Teun (1997). "What is political discourse analysis?" In: *Belgian journal of linguistics* 11.1, pp. 11–52. DOI: 10.1075/bjl.11.03dij.

— (2001). *Discourse, ideology and context*. Walter de Gruyter, Berlin/New York Berlin, New York. DOI: 10.1515/flin.2001.35.1-2.11.

Van Leeuwen, Theo (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford university press. DOI: 10.1002/9781118584194.ch22.

Velkova, Julia and Anne Kaun (2021). "Algorithmic resistance: Media practices and the politics of repair." In: *Information, Communication & Society* 24.4, pp. 523–540. DOI: 10.1080/1369118X.2019.1657162.

Ventola, C Lee (2014). "Social media and health care professionals: benefits, risks, and best practices." In: *Pharmacy and therapeutics* 39.7, p. 491.

Verma, Pranshu and Rachel Lerman (2023). *What is ChatGPT? Everything you need to know about chatbot from OpenAI*. https://www.washingtonpost.com/technology/2022/12/06/what-is-chatgpt-ai/. [Accessed 19-07-2023].

Viera, Anthony J and Joanne M Garrett (2005). "Understanding interobserver agreement: the kappa statistic." In: *Fam med* 37.5, pp. 360–363.

Villena-Román, Julio and Janine García-Morera (2013). "TASS 2013—workshop on sentiment analysis at SEPLN 2013: An overview." In: *Proceedings of the TASS workshop at SEPLN*, pp. 112–125.

Vyas, Vishal and Baskar V. Uma (2018). "An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner." In: *Procedia Computer Science* 125, pp. 329–335. ISSN: 1877-0509. DOI: 10.1016/j.procs.2017.12.044.

Waddell, T Franklin (2019). "Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility." In: *Journalism & mass communication quarterly* 96.1, pp. 82–100. DOI: 10.1177/1077699018815891.

Wagner, Ben (2019). "Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems." In: *Policy & Internet* 11.1, pp. 104–122. DOI: 10.1002/poi3.198.

Wallace, R Jay (1998). *Responsibility and the Moral Sentiments: reprint edition*. Cambridge, MA: Harvard University Press.

Wang, Lucy Lu et al. (2020). *CORD-19: The COVID-19 Open Research Dataset*. arXiv: 2004.10706 [cs.DL]. URL: https://arxiv.org/abs/2004.10706.

Wang, Qiaosi, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel (2021). "Towards Mutual Theory of Mind in Human-AI

Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. ISBN: 9781450380966. URL: https://doi.org/10.1145/3411764.3445645.

Wang, Shuai, Harrisen Scells, Bevan Koopman, and Guido Zuccon (2023). "Can ChatGPT write a good boolean query for systematic review literature search?" In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1426–1436. DOI: 10.1145/3539618.3591703.

Watanabe, Kohei (2021). "Latent semantic scaling: A semisupervised text analysis technique for new domains and languages." In: *Communication Methods and Measures* 15.2, pp. 81–102. DOI: 10.1080/19312458.2020.1832976.

Weber-Lewerenz, Bianca (2021). "Corporate digital responsibility (CDR) in construction engineering—ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice." In: *SN Applied Sciences* 3, pp. 1–25. DOI: 10.1007/s42452-021-04776-1.

Weber, Linda, Linda R. Weber, and Allison I. Carter (2003). *The social construction of trust*. Springer Science & Business Media. DOI: 10.1007/978-1-4615-0779-6.

Weber, Max (1978). *Max Weber: selections in translation*. Cambridge University Press. DOI: 10.1017/CBO9780511810831.031.

Wee, Alicia and Mark Findlay (2021). "Digital Contact Tracing–An Examination of Uptake in UK and Germany." In: *SMU Centre for AI & Data Governance Research Paper* 10. DOI: 10.2139/ssrn.3915303.

Weller, Katrin, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (2013). *Twitter and society*. Peter Lang New York. URL: https://eprints.qut.edu.au/66322/.

Wetherell, Margaret and Jonathan Potter (1988). "Analysing everyday explanation: A casebook of methods." In: SAGE. Chap. Discourse analysis and the identification of interpretative repertoires.

Whalen, Jeromie, Chrystalla Mouza, et al. (2023). "ChatGPT: Challenges, Opportunities, and Implications for Teacher Education." In: *Contemporary Issues in Technology and Teacher Education* 23.1, pp. 1–23. URL: https://www.learntechlib.org/primary/p/222408/..

Whannel, Kate (2022). *Could a chatbot answer prime minister's questions?* URL: https://www.bbc.co.uk/news/uk-politics-64053550.

Whittaker, Adam (2021). "Teacher perceptions of A-level music: Tension, dilemmas and decline." In: *British Journal of Music Education* 38.2, pp. 145–159. DOI: 10.1017/S0265051720000352.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave (2019). "The role and limits of principles in AI ethics: towards a focus on tensions." In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200. DOI: 10.1145/3306618.3314289.

Wicke, Philipp and Marianna M Bolognesi (2021). "Covid-19 discourse on twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time." In: *Frontiers in Communication* 6, p. 651997. DOI: 10.3389/fcomm.2021.651997.

Williams, Matthew L, Pete Burnap, and Luke Sloan (2017). "Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation." In: *Sociology* 51.6, pp. 1149–1168. DOI: 10.1177/0038038517708140.

Williams, Simon N, Christopher J Armitage, Tova Tampe, and Kimberly Dienes (2021). "Public attitudes towards COVID-19 contact tracing apps: A UK-based focus group study." In: *Health Expectations* 24.2, pp. 377–385. DOI: 10.1111/hex.13179.

Wodak, Ruth (2007). "Pragmatics and critical discourse analysis: A cross-disciplinary inquiry." In: *Pragmatics & cognition* 15.1, pp. 203–225. ISSN: 0929-0907. DOI: 10.1075/pc.15.1.13wod.

Woodfield, Kandy, Gareth Morrell, Katie Metzler, Grant Blank, Janet Salmons, Jerome Finnegan, and Mithu Lucraft (2013). *Blurring the Boundaries? New social media, new social research: Developing a network to explore the issues faced by researchers negotiating the new research landscape of online social media platforms - NCRM EPrints Repository — eprints.ncrm.ac.uk*. https://eprints.ncrm.ac.uk/id/eprint/3168/. [Accessed 22-05-2022].

Wymant, Chris, Luca Ferretti, Daphne Tsallis, Marcos Charalambides, Lucie Abeler-Dörner, David Bonsall, Robert Hinch, Michelle Kendall, Luke Milsom, Matthew Ayres, et al. (2021). "The epidemiological impact of the NHS COVID-19 app." In: *Nature* 594.7863, pp. 408–412. DOI: 10.1038/s41586-021-03606-z.

Xie, Yi, Ishith Seth, David J Hunter-Smith, Warren M Rozen, Richard Ross, and Mathew Lee (2023). "Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT." In: *Aesthetic Plastic Surgery*, pp. 1–9. DOI: 10.1007/s00266-023-03338-7.

Yang, Sidi and Haiyi Zhang (2018). "Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis." In: *International Journal of Computer and Information Engineering* 12.7, pp. 525–529. DOI: 10.5281/zenodo.1317350.

Yatoo, Mudasir A and Faiza Habib (2023). "ChatGPT, a friend or a foe?" In: *MRS Bulletin* 48.4, pp. 310–313. DOI: 10.1557/s43577-023-00520-9.

Ye, Rizwan (2023). *The Power of Prompting: Navigating The Future Of AI And Machine Learning*. Rizwan Ye.

Young, Matthew M, Justin B Bullock, and Jesse D Lecy (2019). "Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public adminis-

tration." In: *Perspectives on Public Management and Governance* 2.4, pp. 301–313. DOI: 10.1093/ppmgov/gvz014.

Yousafzai, Shumaila, John Pallister, and Gordon Foxall (2009). "Multi-dimensional role of trust in Internet banking adoption." In: *The Service Industries Journal* 29.5, pp. 591–605. DOI: 10.1080/02642060902719958.

Yu, Hao (2023). "A Cogitation on the ChatGPT Craze from the Perspective of Psychological Algorithm Aversion and Appreciation." In: *Psychology Research and Behavior Management*, pp. 3837–3844. DOI: 10.2147/PRBM.S430936.

Zarsky, Tal (2016). "The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making." In: *Science, Technology, & Human Values* 41.1, pp. 118–132. DOI: 10.1177/0162243915605575.

Zhang, Borui (2023). "ChatGPT, an Opportunity to Understand More About Language Models." In: *Medical Reference Services Quarterly* 42.2, pp. 194–201. DOI: 10.1080/02763869.2023.2194149.

Zhang, Xiuzhen, Lishan Cui, and Yan Wang (2013). "Commtrust: Computing multi-dimensional trust by mining e-commerce feedback comments." In: *IEEE transactions on knowledge and data engineering* 26.7, pp. 1631–1643. DOI: 10.1109/TKDE.2013.177.

Zhou, Jianlong, Heimo Müller, Andreas Holzinger, and Fang Chen (2023). *Ethical ChatGPT: Concerns, Challenges, and Commandments*. arXiv: 2305.10646 [cs.AI]. URL: https://arxiv.org/abs/2305.10646.

Zhuo, Terry Yue, Yujin Huang, Chunyang Chen, and Zhenchang Xing (2023). *Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity*. arXiv: 2301.12867 [cs.CL]. URL: https://arxiv.org/abs/2301.12867.

Ziewitz, Malte (2016). "Governing algorithms: Myth, mess, and methods." In: *Science, Technology, & Human Values* 41.1, pp. 3–16. DOI: 10.1177/0162243915608948.

Zimmerman, Barry J (2000). "Self-efficacy: An essential motive to learn." In: *Contemporary educational psychology* 25.1, pp. 82–91. DOI: 10.1006/ceps.1999.1016.