Don't be Fooled by Fake News? Mapping the Social, Cognitive, and Political Mechanisms of Misinformation Beliefs

Fabio Torreggiani

Supervised by Prof. Gabriele Ballarino
Co-supervised by Dr. Áron Székely

Department of Social and Political Sciences
University of Milan
Academic Year: 2022-2023 - XXXVI Cycle
PhD Program Director: Prof. Marco Guerci

A thesis presented in fulfilment of the requirements for the degree of
Doctor of Philosophy in Economic Sociology and Labour Studies
SPS/09

Acknowledgments

Firstly, I would like to express my gratitude to Dr. Áron Székely. He always believed in me, sometimes more than I did. Working with him was always a great pleasure.

I would like to thank Prof. Moreno Mancosu and Prof. Giuliano Bobba for giving me the chance to implement two empirical studies without asking anything in return.

I am grateful to Francesco Renzini, Carlo Debernardi, Chiara Perin and Margherita Criveller, the wonderful colleagues who helped me translate and back-translate my questionnaires.

Other wonderful people showed me the way by listening to my ideas. In particular, I would like to thank Dr. Amalia Alvarez Benjumea and the other researchers who welcomed me to Madrid, Prof. Diego Gambetta, Dr. Juan Morales, Dr. Fabian Neuner, Dr. Vicente Valentim, and everyone who listened to my presentations. Their comments and questions helped improve this project.

A special thank you is due to the people with whom I shared my working hours and who made them bearable and lighter: Lollo, Chichi, the friends who welcomed me at the Bunker, and those who did the same in the Collegio PhD Room.

I would also like to thank my family for always being there, ready to listen and help. A special thanks also to Fra and Dani, without whom Turin would feel less like home.

Finally, I would like to thank all the musicians who, without even knowing me, immensely helped me with their music over the years.

To my family.

To my friends.

To the time we spend together.

Alla mia famiglia.

Alle mie amiche ed ai miei amici.

Al nostro tempo insieme.

Abstract

Why do people believe in fake news? I explore this question using five studies on more than 4200 participants. While there is extensive research on this question, most studies have primarily tried to identify specific individual mechanisms, e.g. cognitive styles, without considering the possible interplay of various mechanisms. Instead, this dissertation has jointly considered five drivers of misinformation beliefs: cognitive styles, motivated reasoning, anti-elitism, institutional trust, and social norms. The second contribution of this doctoral thesis is the inclusion of these latter three explanations (one of which - social norms - is also experimentally manipulated), as the literature often focused on cognitive drivers while only rarely exploring social ones. This dissertation also aimed to make relevant methodological contributions. First, instead of using artificial or semi-artificial stimuli to test participants' truth discernment, it collected and validated a new dataset of 80 existing social media posts. Second, it also tested its hypotheses in Italy, outside the widely explored context of English-speaking countries. The results of these studies indicate that cognitive styles, such as the tendency to rely on intuition or analytical thinking, significantly influence participants' ability to discern truth, even when using real social media posts as stimuli. However, the effects of other drivers were less pronounced. Motivated reasoning, for instance, had only marginal effects, possibly due to the low perceived partisanship of the collected posts. Similarly, I found no consistently significant effects of anti-elitism and institutional trust, except for trust in universities. Regarding social norms, I observed that individual motivations for information accuracy were associated with higher truth discernment, while social expectations had no discernible impact. The

social norm treatment message yielded mixed results but was largely inconclusive. My findings have two main insights: for methodological issues, my thesis testifies to the importance of using more realistic stimuli selection in experimental research on misinformation; for substantive issues, my study shows that cognitive drivers, although crucial, cannot be considered the sole determinants for misinformation beliefs.

Table of Contents

1 Introduction **10**

2 Literature Review **18**

2.1 Concepts and Definitions: What is a Fake news? 18

2.1.1 Misinformation, Disinformation and Fake news 19

2.1.2 What is the Focus of this Thesis and Why 20

2.2 Factors and Mechanisms Explaining Misinformation Perception 22

2.2.1 Factors Linked with Misinformation Believing 23

2.2.2 Mechanisms Explaining Misinformation Beliefs 24

2.3 Methodological Approaches 26

2.3.1 The Central Role of Experiments 27

2.3.2 Existing Works Using a Survey Approach 29

2.3.3 Other Methodological Approaches 30

2.4 Gaps in the Literature 31

2.4.1 Manipulation, Not Mapping 31

2.4.2 Lack of Social Mechanisms 35

2.4.3 Stimuli Selection: Artificial, Semi-artificial and "Natural" Stimuli 36

2.4.4 Scarce Evidence Outside Anglo-Saxon Countries 39

2.4.5 Theoretical Contributions 41

3 Theoretical Framework **45**

3.1 The Role of Analytical Thinking 47

3.2 The Role of Motivated Reasoning 48

3.3 The Role of Anti-Elitism 50

3.3.1 Defining Anti-elitism 51

3.3.2 Anti-Elitism and Misinformation Believing: One Link, Many Paths 52

3.4 The Role of Institutional (Dis)Trust 57

3.4.1 Defining Institutional (Dis)Trust 57

3.4.2 Institutional (Dis)Trust and Misinformation Believing: a Parallel with Anti-Elitism? 59

3.5 The Role of Social Norms 61

3.5.1 Concepts and Definitions: What Do We Mean with Social Norms? 63

3.5.2 Norms Governing "Bad" Beliefs and Their Expression 64

3.5.3 Social Norms About What? Believing, Reading and Sharing 65

3.5.4 What is Punished and What Are the Sanctions? 71

3.5.5 How the Social Norm Mechanism Works 74

3.6 Research Questions 75

4 Research Design **77**

4.1 Structure of the Survey Experiments 79

4.1.1 Measuring the Explanatory Variables 79

4.1.2 Measuring the Dependent Variable 83

4.1.3 The Treatment and the Treatment Groups 83

4.1.4 Other details of the survey 86

4.2 Structure of the Validation Studies 88

4.3 Selection of the Social Media News Posts 91

4.3.1 Choosing the Fake News Posts . . . 91
4.3.2 Choosing the Reliable News Posts . . . 93
4.3.3 Processing and Final Datasets . . . 94
4.4 Analytical Methods . . . 95
4.4.1 Analytical Methods of Validation Studies . . . 95
4.4.2 Analytical Methods of Main Studies . . . 96
4.5 Hypotheses . . . 101
5 Results of Validation Studies 106
5.1 Perceived Quality of the Social Media Posts . . . 108
5.1.1 Presence of Outliers . . . 108
5.1.2 Qualitative Differences Between Fake News and Reliable Posts . . . 111
5.1.3 Cross-Country Comparability . . . 113
5.2 Perceived Political Leaning of the Social Media Posts . . . 115
6 Results of the Convenience Sample 123
6.1 Cognitive Styles . . . 126
6.2 Motivated Reasoning . . . 127
6.3 Anti-elitism . . . 129
6.4 Institutional Trust . . . 130
6.5 Social Norm on the Importance of Accuracy . . . 132
6.5.1 Does an "Accurate Sharing" Social Norm Exist? . . . 132
6.5.2 Correlational Evidence on the Relation with Truth Discernment . . . 135
6.6 Treatment Effect . . . 136
7 Results of the Italian Quota Sample 140
7.1 Cognitive Styles . . . 141
7.2 Motivated Reasoning . . . 141

7.3 Anti-elitism . . . 144
7.4 Institutional Trust . . . 145
7.5 Social Norm on the Importance of Accuracy . . . 147
7.5.1 Does an "Accurate Sharing" Social Norm Exist? . . . 147
7.5.2 Correlational Evidence on the Relation with Truth Discernment . . . 150
7.6 Treatment Effect . . . 151
8 Results of the US Quota Sample 154
8.1 Cognitive Styles . . . 157
8.2 Motivated Reasoning . . . 157
8.3 Anti-elitism . . . 158
8.4 Institutional Trust . . . 159
8.5 Social Norm on the Importance of Accuracy . . . 160
8.5.1 Does an "Accurate Sharing" Social Norm Exist? . . . 160
8.5.2 Correlational Evidence on the Relation with Truth Discernment . . . 163
8.6 Treatment Effect . . . 164
9 Discussion 167
9.1 Retesting Cognitive Mechanisms . . . 167
9.1.1 Confirmatory Evidences on the Role of Cognitive Styles . . . 167
9.1.2 The Marginal Impact of Left-Right Motivated Reasoning in Everyday News . 168
9.2 Trust in "Trusted" Institutions is Positively Linked with Truth Discernment . . . 170
9.2.1 Considerations on Anti-elitism. . . . 173
9.3 Contextual Variability of the "Accurate Sharing" Social Norm and Its Disconnection from Truth Discernment . . . 174
9.3.1 Explaining the Ineffectiveness of Social Expectations and the Treatment . . . 176
9.4 General Considerations . . . 178

9.4.1 Difficulty of the Chosen Stimuli . . . 178
9.4.2 Inclusion of "I Don't Know" Option . . . 179
9.4.3 Erroneous Exclusion of Forza Italia Voters in *Main-ITA* . . . 180
10 Conclusion **181**
10.1 Contributions . . . 181
10.2 Limitations . . . 183
10.3 Future Research . . . 184
Bibliography **189**
A Regression Tables **198**
B Complete List of Utilized Social Media Posts **202**
C Questionnaire of Main Studies - English Version **217**
D Ethics Committee Certificate **252**
E Pre-registration **254**

Chapter 1

Introduction

The diffusion of false information, and even their deliberate creation, are not recent phenomena and are probably as old as society itself (Bloch, 1921). Historical examples can be dated back to the birth of the Roman Empire, although a proper inquiry would plausibly bring out even older episodes. During the Ides of March of 44 BC, a group of conspirators led by Brutus and Cassius assassinated Julius Caesar. The vacuum of power left by this sudden assassination paved the way for a harsh civil war. The members of the Second Triumvirate initially joined their forces together to defeat Caesar's assassins. However, Octavian, the adopted child of Caesar and his formal heir, and Mark Antony, Caesar's longtime and closest ally, both claimed the Republic's leadership. Octavian started a defamatory campaign to discredit his opponent to win this war. In the absence of mass media, he minted coins with slogans about Mark Antony, falsely depicting him as an immoral person, an alcoholic and a traitor of the Republic in favour of Cleopatra, the queen of Egypt[1].

A more recent example is the panic caused by the radio adaptation of "War of the Worlds", directed by Orson Welles. On the night of Halloween of 1938 a fictional bulletin aired on CBS Radio Network, narrating an ongoing alien invasion in the US. Despite being introduced by an announcement that clarified its fictional nature, the transmission caused widespread panic, according to

[1]The long history of disinformation during war. https://www.washingtonpost.com/outlook/2022/04/28/long-history-misinformation-during-war/ Visited on March 2, 2024.
A lesson in fake news from the info-wars of ancient Rome. https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6 Visited on March 2, 2024.

newspapers of the time. However, this latter information was probably exaggerated by newspapers greatly. The transmission had a relatively limited audience and there are no reports of episodes of public panic[2]. Interestingly, what seemed to be a textbook case of misinformation-induced public disorder turned out to be a case of poor journalism. Misinformation was not spread by the radio transmission, but by newspapers discussing it.

To summarise, false information is nothing new. Why, then, should we deal with it now? The issue of disinformation recently gained popularity due to the diffusion of online media and social media, combined with the occurrence of other political, social and philosophical changes (Lazer et al., 2018; McIntyre, 2018). These recent changes, combined, are argued to have amplified fake news' negative consequences (Kata, 2010). For instance, misinformation has been mentioned as one of the possible determining factors of the Capitol Hill riot following the 2020 US Presidential Election (Pennycook & Rand, 2021a) as well as the negative reactions to anti-COVID-19 measures (Pennycook, McPhetres, et al., 2020).

The concern linked with the spreading of online misinformation and its consequences on society caused a rapid and exponential increase in academic production. A new field of papers studying it originated by gathering scholars from various disciplines. For example, the number of documents registered on the citation database Scopus almost doubled between 2019 and 2020, after an accelerating trend in previous years (see Figure 1.1).

Despite being relatively recent, this burgeoning academic production on fake news already addresses different objectives. A common trait of these subfields is an attempt to stop misinformation, which is considered negative for contemporary societies.

The first strand of authors tried developing methods for the automatic detection of fake news through the use of algorithms (see, for example, Shu et al., 2017). The production and spreading rate of fake news is too high to allow for their manual checking. Automatic detection, instead could

[2]Orson Welles and the Birth of Fake News. https://www.nytimes.com/2018/10/30/opinion/orson-welles-war-of-the-worlds-fake-news.html Visited on March 2, 2024.

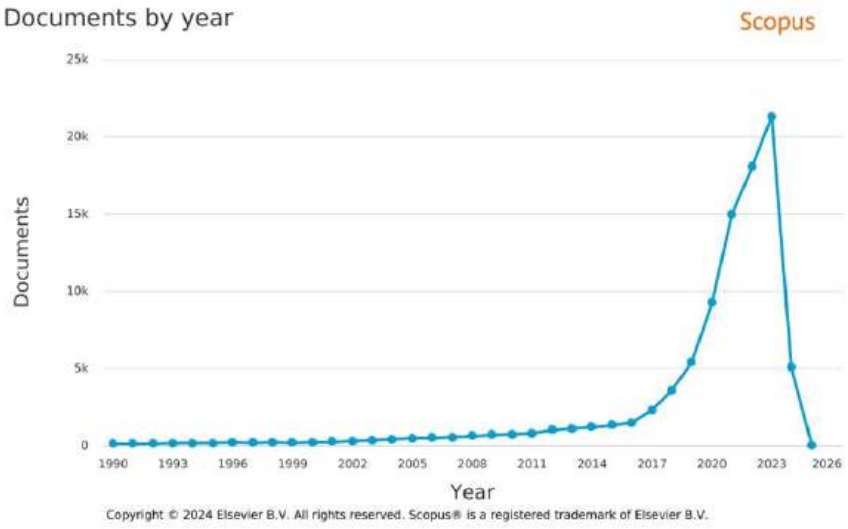


Figure 1.1: Documents collected on Scopus, by year. Analysis made with Scopus' advanced research tool on April 4, 2024. Query string: 'misinformation OR "fake news" OR disinformation'. The plot presents only documents published between 1990 and 2023. The dot representing 2024 only covers the first three months of the year. The dot representing 2025 could be over zero because of papers already registered for that year.

allow social media and other platforms to delete dangerous content at an effective rate.

Another important strand regards the definition of the concept of fake news and its related terminological nuances (see, for example, Tandoc et al., 2018). The definition of this concept, along with others which are semantically close to it, has still not reached a definitive consensus in the literature. Conceptual confusion can complicate communication, negatively impacting academic production and the public debate. This strand aims to solve this confusion.

This project, like many others, tries to contribute to exploring why internet users share and believe fake news (see, for example, Pennycook and Rand, 2021b). It aims to understand the factors that are linked to increased misinformation beliefs. Additionally, it tries to explain the interaction between individuals and information, trying to examine the mechanisms behind this interaction. The underlying perspective of this strand is that to stop misinformation, research needs to dissect how individuals interpret information and how false beliefs are formed and diffused.

In particular, I contribute to the literature by testing five different explanations of misinforma-

tion beliefs within the same research framework. I do this by implementing five different studies on a total of more than 4200 participants across Italy and the United States. Firstly, I re-evaluate the two main existing explanations of misinformation beliefs (cognitive styles and motivated reasoning), empirically testing them in two countries. Furthermore, I explore the role of three social drivers (anti-elitism, institutional trust and social norms), which received much less attention in the literature. The resulting evidence is not only correlational, but partially causal, as one of these drivers is also tested experimentally (social norms). Finally, on a more methodological side, I test all these drivers on a new and validated set of existing social media posts, instead of widely-used artificial or semi-artificial stimuli.

Past research already found relevant evidence on the role of various factors and mechanisms but rarely went outside the boundaries of cognitive explanations. According to one cognitive explanation, individuals believe in misinformation because they are not focused enough on the information they receive. As a consequence, they fail to recognise cues of its veracity. According to an alternative explanation, individuals are not interested in the accuracy of what they read or hear. Instead, they are primarily motivated to confirm their prior political and cultural ideas.

Cognitive drivers are clearly crucial in evaluating news. Information processing is an inherently individual process that is strictly linked to the way we think. However, it is also a largely social process as we rarely have the resources to analyse every needed information by ourselves. Instead, we rely on others in an infinite number of evaluations across our lives.

For example, think about a person evaluating whether to pick up and eat a potentially venomous mushroom. For sure, their decision will be influenced by how they evaluate information more broadly, e.g. by how much "analytical" or "intuitive" they are (see Section 2) or by their previous beliefs about that mushroom's potential harmfulness. However, if this person is part of a society and not living in a social vacuum, they could also ask to other humans.

Similarly, when evaluating news and information more broadly, our decisions will be influenced

by our cognitive tendencies and by our set of beliefs. However, we will also be influenced by others' opinions in the form of direct or indirect communications. For example, we could ask a peer that we trust if they know anything about it. Alternatively, we could also rely on indications spread by institutions and public figures, such as newspapers, politicians, ministries and so on.

The current literature on misinformation beliefs largely overlooked the potential role of peers and socially constructed institutions in news evaluation. Nevertheless, these social drivers could have distinct consequences on misinformation perceptions. Consequently, studying them could help us having a more nuanced understanding of this process and developing more effective countermeasures.

In particular, I propose the exploration of three social drivers: social norms, anti-elitism, and institutional trust. These relatively understudied perspectives all focuses on instances where individuals' assessments are influenced by information coming from *outside* themselves, instead of being determined by just prior beliefs and cognitive resources. In other words, in all these three drivers, I expect individuals to be influenced by cues about or coming from others, whether them being non-organized peers or more structured institutions.

Starting from the first driver, I argue that, independently of their cognitive characteristics, individuals are also influenced by their peers and the importance they give to the accuracy of circulating information. Other things being equal, individuals who believe that everybody else is prioritizing accuracy will put more effort in evaluating news[3]. Conversely, individuals who perceive their social context as not interested in accuracy will also feel less pushed to be committed in their evaluations.

On a more vertical perspective, I expect individuals to be also influenced by how much they trust institutions linked to the public discourse. Individuals can rarely access events mentioned in the news directly or contact people who did so. Instead, they often have to rely on institutions,

[3]I intentionally do not further define the term "everybody else", as understanding the relevant reference network of this potential social norm is an empirical question.

organization and public figures who indirectly mediate their access to this knowledge. Literature rarely studied the potential role of individuals' trust in these mediators. On the contrary, I propose to explore the role of anti-elitism and institutional trust. I expect individuals with high anti-elitist beliefs and/or low institutional trust to have lower truth discernment, proposing three possible theoretical justifications for this expectation.

Besides the chosen theoretical frameworks to explain misinformation beliefs, other limitations of the literature regard the methodological approaches used to do so. Many past studies addressed the above-mentioned research question by resorting to manipulation and experimental approaches. While useful, this approach left many questions unanswered. In particular, past papers rarely compared multiple explanations trying to test which is the most strongly tied to misinformation beliefs, or whether different individuals act according to different mechanisms.

Instead this project measures and maps cognitive, political and social drivers within the same framework, using the same samples and surveys. This allows for their simultaneous testing and the assessment of their relative strength. This focus on mapping, rather than manipulating variables, is mirrored by using quota samples. Quotas were selected to represent respective populations in various demographics, such as age, gender, and education.

The exploration of social drivers, along with the retesting of cognitive ones, was accomplished with the implementation of three survey experiments in Italy and the United States, involving over 3000 participants. Two of them used quota samples for a total of around 2400 participants. Instead, the third study involved an Italian convenience sample of around 600 participants. These three studies had the same structure and minimal contextual differences. Through standard survey items, they measured demographic, cognitive, political and social variables. Additionally, participants were asked to judge a series of news posts containing both false and reliable information, to assess their truth discernment.

However, in addition to mapping different drivers of misinformation beliefs, the project also

aims to manipulate perceptions of social norms. In particular, the three main survey experiments test the effects of an experimental intervention in the form of a social norm information message. The treatment informs participants about their peers' social expectations regarding the importance of accuracy in the news they share online. The aim is to test whether this can change individuals' social expectations and, ultimately, their ability to discern news.

Furthermore, existing research often utilized statements, modified newspaper screenshots, and other types of stimuli that only partially replicate what individuals encounter in their everyday interactions. Instead, more realistic stimuli are rarely used. When they are, the selection often follows unclear or unreported procedures and lacks proper validation.

To enhance the external validity of its conclusion, the present work utilizes stimuli collected through a transparent and replicable procedure. The resulting dataset is a pool of 80 existing Italian and US-based social media posts that either contain false information or do not. Posts were then minimally edited and rigorously validated. Two validation studies were run in the two countries with convenience samples of around 600 participants each. These studies measure participants' perceptions of the stimuli to avoid outliers and ensure comparability between the two contexts.

Finally, the existing literature is largely based on empirical evidence from the United States and other Anglo-Saxon countries. The baseline assumption that mechanisms at the individual levels can be generalised across cultures is rarely tested. In addition, academic production in other countries only occasionally builds on the wider English-based literature. Instead, this project closely replicates the same studies in the US and in Italy. The first country is chosen to allow strict comparability of my results with most existing papers, while the Italian replication allows for their testing outside the Anglo-Saxon context.

The present thesis will now outline the existing literature on fake news (Chapter 2), starting with the definition of relevant concepts and proceeding with its main findings and gaps. Chapter 3 is dedicated to the theoretical framework on which this project bases its potential theoretical

contribution, alongside the research questions that stem from it. Chapter 4 will outline the methodological details of the project, while Chapter 5 to 8 will show the main results of the five studies. These findings are then discussed more comprehensively in Chapter 9. The thesis concludes with some considerations of the effective contributions of this project, its limitations, and suggestions for future research.

Chapter 2

Literature Review

2.1 Concepts and Definitions: What is a Fake news?

Three key concepts in the study of false information are *misinformation*, *disinformation* and *fake news*. The literature and public discourse have used these concepts interchangeably and with some confusion.

Journalists use the term fake news to refer to many different concepts, sometimes only loosely related to its formal definition (Egelhofer et al., 2020). For example, they sometimes refer to attacks on legacy news media or use it as an empty buzzword to state that something is incorrect or debatable.

On their side, politicians did not help resolve this confusion by using the term "fake news" to discredit political opponents or mainstream media (Habgood-Coote, 2019). For example, ex-US President Donald Trump transformed this term into a slur that referred to people or organizations instead of pieces of information[1].

Academic research, too, has grappled with semantic ambiguity. For example, a review of 34 articles found that they used the term "fake news" with 6 different meanings (Tandoc, 2019). Those include news satire, news parody, fabrication, manipulation, advertising, and propaganda. This conceptual ambiguity also encompasses the term "misinformation". Some authors assert

[1] https://www.theguardian.com/us-news/2017/jan/11/trump-attacks-cnn-buzzfeed-at-press-conference Visited on April 6, 2024.

that the intention to deceive the reader is a crucial characteristic of misinformation (Wardle and Derakhshan, 2017). Others argue it refers to false information, regardless of the writer's intentions (Brown, 2018).

To address this confusion, I will define the concepts of misinformation, disinformation and fake news (Section 2.1.1). This will establish a stable definition of terminology. I will then clarify my work's positioning (Section 2.1.2). To anticipate, this work bases its empirical results on social media posts about news events, which I here consider a specific type of misinformation.

2.1.1 Misinformation, Disinformation and Fake news

Misinformation, the broadest category, encompasses all types of *"false or misleading information"* (Lazer et al., 2018). This concept's central characteristic is its low level of *facticity*. Tandoc (2019) uses this dimension to describe whether communication relies on facts.

The literature is discordant on whether deception is a crucial characteristic of misinformation. Tandoc et al. (2018) defines *deception* as the degree to which the communication creator intends to mislead[2]. Many authors limit misinformation to cases where falsehood is inadvertently created and diffused (Chadwick et al., 2018; Wardle & Derakhshan, 2017). Other authors do not make such a distinction, including content intentionally created to deceive (Brown, 2018; Lazer et al., 2018). Finally, some argue that misinformation is false or inaccurate information that is deliberately created and intentionally or unintentionally propagated (Wu et al., 2019). Deception is thus considered in the act of creation, but not diffusion.

Misinformation does not consider the format of information. Some papers use it to refer to false news that is shared online (Bakir & McStay, 2018). In this case, the focus is on specific *pieces of information*, which can have a format. In other cases, misinformation refers to false *beliefs*. This conceptualization focuses on beliefs held in people's minds. It is detached from any specific post, article or dialogue that caused them to arise (Arechar et al., 2023; Vegetti & Mancosu, 2020).

[2]In the original paper, these two dimensions are used to categorize fake news as defined in the 34 reviewed papers (Tandoc et al., 2018). In this Section, I use them to outline types of information more broadly.

Authors define *disinformation* as *"false information that is purposely spread to deceive people"* (Lazer et al., 2018). Unlike misinformation, it includes the element of intention to mislead the reader in addition to being false. The consensus on this concept is much broader, as definitions across various authors share its core characteristics of deception and low facticity (for example, see Bakir and McStay, 2018; Pantazi et al., 2021; Vegetti and Mancosu, 2020).

To conclude, authors define *fake news* as *"fabricated information that mimics news media content in form but not in organisational process or intent"* (Lazer et al., 2018). This subtype of disinformation introduces a third element: the format of the information. In essence, fake news is false information (low facticity), created to deceive (deception), and presented in a format that mimics traditional news media content.

2.1.2 What is the Focus of this Thesis and Why

This work gathered evidence on a specific type of information: social media posts discussing news events. I conceptualise this as a specific type of misinformation, given that its defining characteristic is the false information it contains (*facticity*).

Posts were selected solely based on their facticity. On the contrary, the selection was not based on deception and news resemblance. The reason for this choice is that facticity is the only readily observable characteristic of posts, thanks to the use of debunking sites. In particular, the primary emphasis is on social media news posts reporting events that did not occur or falsely representing actual events.

On the contrary, assessing the *deception* of social media posts is challenging. The intention to deceive seems plausible when the writer claims to be the primary source of non-existent events. Conversely, when writers of false social media posts report other sources, discerning whether they intended to deceive or were themselves deceived is more complex. A comprehensive decision on this matter would require studying each source individually. In contrast, this project uses posts evaluated by debunkers solely based on their facticity (see Section 4.3).

Similarly, it is challenging to establish whether each post's format mimics news media content enough to classify it as *fake news*. Some posts resemble those of mainstream news pages, while others exhibit less professional formatting. Notably, this distinction is not relevant to debunking sites, which, as said, categorise posts as fake news or not solely based on facticity.

Even if I consider social media news posts a specific type of misinformation, in some cases, this concept closely aligns with those of disinformation and fake news. Consequently, during the manuscript, I also rely on the literature on disinformation and fake news. Additionally, I use the term fake news to refer to the stimuli used in the survey experiments, when they contain false information. Nevertheless, the main focus remains on misinformation, as I do not draw conclusions about writers' willingness to deceive or on posts' specific formats. While the findings may have implications for other types of misinformation and disinformation, further research is needed to establish their generalizability.

Why Social Media News Posts

The choice of focusing on social media news posts contrasts with the literature, as Section 2.4.5 will explain. The primary rationale behind this choice is external validity. On social media, misinformation typically takes the form of posts. While it is possible to encounter more traditional news articles, it is arguably less common. Some papers explore misinformation beliefs by having participants judge de-contextualised statements about a specific fact, statistical figure or event (e.g., Arechar et al., 2023). Encountering this type of information on social media seems even less plausible.

A second, more practical reason for using social media posts is their prevalence as the most common type of fake news found in debunking sites. For example, news-like articles are considerably rarer in these archives. Consequently, constructing a satisfactory dataset of fake news articles without resorting to inventing items is challenging.

2.2 Factors and Mechanisms Explaining Misinformation Perception

In this section, I outline the academic literature on misinformation perception, exploring how people perceive news irrespective of its format and why they believe in fake news. In doing so, I will address academic production about fake news, misinformation and disinformation. These three sub-strands are closely connected. Authors often find evidence of similar explanations in more than one sub-field. I will thus illustrate these three sub-fields together to avoid continuous clarifications and distinctions.

Despite its recent emergence, the literature on fake news is rapidly expanding, capturing critical empirical findings. This swift growth underscores the urgency and importance of understanding misinformation perception, which draws researchers from various fields without a unifying approach, making systematisation a challenge.

A helpful way to systematise this literature is to divide it into two main strands: works focusing on factors and works about mechanisms. The first category encompasses papers examining characteristics linked with belief in misinformation without explaining how or why this occurs. These characteristics can refer either to the piece of information or its reader. The second category investigates the latent mechanisms underlying misinformation belief. This second strand is interested in explaining how information is processed and then eventually transformed into behaviours. These two strands are interconnected. On the one hand, evidence about mechanisms allows us to interpret evidence about factors. On the other hand, evidence about factors can shed light on the potential existence of unexplored mechanisms.

As I provide this overview, it is crucial to note that this division transcends the correlational-causal typology. In fact, both types of works can be found in both categories. Additionally, this division is introduced for better exposure clarity and should not be interpreted as a strict, formal, or institutionalised separation. This overview serves as an introduction to the academic context of

my work rather than an exhaustive account of the literature. See Ecker et al. (2022) and van der Linden (2022) for recent reviews on the topic.

2.2.1 Factors Linked with Misinformation Believing

Factors associated with misinformation beliefs can be divided into those referring to news characteristics or readers' characteristics. When it comes to news factors, characteristics of fake news are relevant either because they lead people to believe them or because they allow fake news to spread more rapidly than reliable news.

A content analysis of articles from fake news sources reveals that they use sensationalism, clickbait, misleading content and partisan bias (Mourão & Robertson, 2019). This latter characteristic, notably, is correlated with increased social media engagement. Articles with a stronger conservative or liberal leaning record a higher number of interactions, such as comments, likes and shares (Mourão & Robertson, 2019). Another characteristic that influences news credibility is its source. Knowing what organisation or person created a piece of information can dramatically change its believability. The literature studied source credibility long before the surge of fake news (for example, see Pornpitakpan, 2004). However, more recently, authors also addressed source credibility in fake news. Unsurprisingly, reliable sources are more credible than fake news sources (Bauer & Clemm von Hohenberg, 2021). Nevertheless, the latter can enhance their credibility by aligning the narratives of their articles with readers' political positions (Bauer & Clemm von Hohenberg, 2021). Subjects are more prone to believe a source that confirmed their attitudes in the past, even if these are fake news sources. Strictly linked to source credibility, other papers study the role of knowing who shared a piece of news. The believability of content can considerably change depending on which person, organisation, or social media page shares it. Sterrett et al. (2019) find that subjects believe more in the news shared by a more trusted public figure. Crucially, this is true whether the source which created the news is mainstream or fake.

Other authors explore factors linked to a faster spreading of news. This latter perspective is

slightly different, as social media users can share news without believing it. Vosoughi et al. (2018) find that, on average, fake news spreads faster than reliable news on Twitter. In addition, they evaluate factors that could explain this differential speed. The first is novelty. Fake news talks about events that have happened more recently. Thus, they appear both more surprising and more valuable. The second is the presence of negative emotions, such as fear and disgust, in fake news.

Apart from news characteristics, the literature also explored the role of various characteristics of news readers. Evidence shows that age and political affiliation consistently correlate with belief in misinformation. For example, subjects in older age groups interact more with fake news and rarely click on their link to read the whole article (Loos & Nijenhuis, 2020). Regarding political affiliation, a study finds that unsubstantiated rumours spread faster among conservatives than liberals (DeVerna et al., 2024). In addition, the former is more resistant to implemented corrections (DeVerna et al., 2024). For a list of papers finding evidence on age and political affiliation as relevant factors, see Pennycook and Rand (2021b). Another relevant characteristic of news readers is political sophistication. Participants with a better knowledge of the political system are better at distinguishing reliable from fake news (Vegetti & Mancosu, 2020). Similarly, media literacy is a crucial factor influencing people's ability to distinguish fake news from reliable sources. This concept is defined as the understanding of how the media system operates. Evidence suggests that higher media literacy is associated with lower misinformation believing (Hameleers, 2022). Another factor, information literacy, focuses more on the ability to navigate and locate information. Some papers find this dimension, not other literacies, to be the real driver of truth discernment (Jones-Jang et al., 2021).

2.2.2 Mechanisms Explaining Misinformation Beliefs

The literature has explored some relevant mechanisms that could be useful in explaining misinformation beliefs. Many of these mechanisms are cognitive, and they are based on how individuals process information. Two of these mechanisms regard the role of cognitive styles. The first, cham-

pioned in particular by works such as Pennycook and Rand (2019), posits that people fall for fake news due to a lack of analytical thinking. Individuals rely on their intuitive system (Kahneman, 2011) and often fail to think analytically when processing news for various reasons. In simpler terms, individuals may not be focused enough to recognise cues indicating that what they read is false.

The second strand, encapsulated in the concept of motivated reasoning, argues that analytical thinking can also lead to misinformation beliefs. Scholars like Kahan (2017) contend that analytical thinking is not inherently directed towards discovering objective truth. Instead, cognitive effort can be motivated by various factors, with accuracy being just one of them. Individuals may be motivated to confirm their political beliefs, leading them to fall for misinformation independently of their use of analytical thinking and, in some cases, precisely because of it.

Another explanation for misinformation beliefs is illusory truth. This mechanism focuses on individuals' tendency to mistake the ease of processing information for accuracy. Individuals are more prone to believe information that is easier to process. Many factors can determine this ease. For example, individuals are likelier to accept information they have encountered before, even if implausible (Pennycook et al., 2018). In support of this mechanism being a limitation of human cognition, evidence indicates that this effect is robust to incentives for accuracy (Brashier & Rand, 2021).

Again, on the role of cognitive drivers, papers explored cognitive failures. According to this mechanism, individuals believe fake news because of a series of mistakes in interpreting or remembering cues of their veracity. For example, evidence shows that individuals forget the source of information they receive (Rahhal et al., 2002). More recently, Greene et al. (2021) find the existence of "false memories" in participants exposed to fabricated news about Brexit. In other cases, individuals disregard information about the source altogether. Dias et al. (2020) find that interventions highlighting the source did not change participants' ability to recognise misinformation.

Emotions elicited by news can also affect individuals' ability to judge them. In particular, emotions distract the reader from other relevant cues of veracity. The emotional frame with which events are portrayed can change participants' judgment of the actors involved.

Regarding social drivers, very few papers already explored social norms in the field of fake news. In particular, two papers directly investigate this phenomenon. Andı and Akesson (2020) discovered that a priming message based on social norms significantly increases the number of people stating that they would refrain from sharing false articles. In a separate study, Gimpel et al. (2021) found that social norm-based treatments effectively encourage people to engage more in reporting behaviour. Other papers, while not directly addressing social norms, gather evidence that can be linked to this mechanism. Altay et al. (2022) observes that sharing fake news may lead to reputational consequences, indicating the potential presence of informal sanctions that individuals fear. Similarly, Talwar et al. (2019) found that social comparison motivates individuals to refrain from sharing fake news to safeguard their image. Qualitative exploration into interaction within WhatsApp groups sheds light on how individuals seek to limit the circulation of fake news by establishing informal rules and sanctions (Chadwick et al., 2023). Evidence regarding the role of peers in shaping beliefs also emerges from related academic fields. For instance, research indicates that individuals are more inclined to believe in conspiracy theories if they think others agree with them (Pennycook et al., 2022).

2.3 Methodological Approaches

In the previous paragraph, I summarised essential evidence in the literature on the question, "Why do people believe in fake news?". Now, let us examine how authors gathered these answers. Despite being relatively recent, the research on misinformation beliefs already accounts for various methodological approaches. The primary methodology is experiments (Section 2.3.1). Much of the earlier studies in this field use experiments to explore how people perceive news. Secondly, I will focus on

survey-based works (Section 2.3.2), as this is the other core methodological approach of this study. A third section will summarise works that have deployed other methodologies (Section 2.3.3).

2.3.1 The Central Role of Experiments

The exploration of *causal links and mechanisms* within misinformation beliefs predominantly relies on experimental studies. This approach's centrality arises from its inherent characteristics that align with the needs of academic inquiries in this field. I identify three main reasons for these methods' centrality: the suitability of experiments for causal inquiries, their reproducibility, and the contextual need to find fast solutions against fake news.

Firstly, experimental studies address the "fundamental problem of causal inferences" by leveraging random treatments, allowing the observation of how the dependent variable changes both with and without the influence of a manipulation (Holland, 1986; Imai et al., 2013). The randomisation of the treatment allows for the observation of how the dependent variable changes when only the manipulated variable mutates. Causation can thus be established by isolating the influence of one single explanatory variable at a time. For this reason, experiments are of primary importance in all inquiries that aim to explore causal mechanisms. Survey-based research, on the contrary, only allows for observational research. Associations resulting from regression analysis can illuminate possible links between measured variables. However, they cannot inform us of the causal mechanisms that create these links since they do not allow for the observation of counterfactual scenarios. Finally, qualitative research can give meaning to observed associations. However, it lacks the possibility of observing how variables mutate in isolation.

Secondly, the experimental strand of research, characterised by standardised practices, facilitates an incremental and replicable production of knowledge. In misinformation research, the accumulation of knowledge took two main paths. On the one hand, different studies address the same research question using multiple operationalisations and measurements of the dependent variable. This enhanced the results' reliability by mitigating the influence of technical differences specific to

each study. For example, Ross et al. (2021) and Pennycook and Rand (2019) both found evidence of the role of cognitive styles. However, they measure the dependent variable with slightly different formulations. The consistent use of Cognitive Reflection Tasks (CRTs) to measure the explanatory variable allows for the comparison of their results. The same reasoning can also be applied to studies outside the experimental realm of this field. Mosleh et al. (2021) similarly explore the role of cognitive styles, this time by connecting survey-measured CRTs with online behaviour.

On the other hand, consistent operationalisations and measurements of the dependent variables were used to explore the role of different explanatory variables. This consistency excludes the possibility that utilising always-changing items fictitiously creates new, eventually contradictory evidence. For instance, many studies utilise identical questions and answers to measure perceived accuracy across different experiments. In particular, they all employ the same core question: *"To the best of your knowledge, how accurate is the claim in the above headline?"*. Some examples of the explanatory variables studied with this item are the impact of warning tags (Clayton et al., 2020), the role of emotions (Martel et al., 2020), the "implied truth" effect (Pennycook, Bear, et al., 2020), and the effectiveness of attention-improving techniques (Pennycook et al., 2021).

A third reason for the prevalence of experiments is that they became a fitting approach to addressing potential solutions directly. This field of research abruptly emerged as a response to public discourse's concerns about the effects of misinformation. Consequently, the authors needed a methodology that allowed them to test the effectiveness of solutions rather than measuring existing variables. Driven by this urgency, authors resorted to experimental design to test potential interventions. Besides the already mentioned papers (Clayton et al., 2020), other papers similarly explore the role of accuracy prompts (for a review, see Pennycook and Rand, 2022), while others test literacy programs (Hameleers, 2022; Scheibenzuber et al., 2021) or economic incentives (Brashier et al., 2021) (for a review of papers focused on interventions, see Kozyreva et al., 2022).

2.3.2 Existing Works Using a Survey Approach

More recently, mapping emerged as an alternative approach to manipulation. An illustrative example of this switch is the pivotal study from Pennycook and Rand (2019), cited in the previous section. The authors employ an observational point of view instead of an experimental one. They collect data on the perceived accuracy of news and explanatory variables linked to two main explanations. They then map these two different mechanisms by observing how they are differently associated with the ability to recognise fake news. By doing so, they do not manipulate different variables, given that the study does not add any treatment. Instead, the authors compare conflicting explanations in the literature to establish their relative impact on the dependent variable. Interestingly, Pennycook and Rand (2019) do not gather observations by implementing a new survey. Instead, they collect observations from control groups of other experimental studies.

The switch from causal testing to description also determines the increase in publications using surveys. Instead of aggregating control groups from different studies (like in the Pennycook and Rand, 2019 paper), authors designed dedicated surveys to measure different explanatory variables and collect correlational evidence on their link with the perceived accuracy of news. For example, as a first step in this direction, Angelucci and Prat (2023) measure the *dimension* of the problem, i.e. how many individuals can distinguish real stories from fake news. The paper also tries to explain this ability in terms of socioeconomic variables.

Other authors address the same research question, exploring correlational links with various explanatory variables. To name a few examples, these include demographic variables and trust in institutions (Pickles et al., 2021), the use of social media and belief in conspiracy theories (Halpern et al., 2019), and partisanship (Faragó et al., 2020). A separated but closely linked sub-field of research is the one dedicated to misinformation about COVID-19 and the exploration of its determinants (see, for example, Roozenbeek et al., 2020 and Agley and Xiao, 2021).

Again on this line, an impressive academic effort pursued by many authors in various countries

resulted in the implementation of an extensive cross-country survey (Arechar et al., 2023). Implemented in 16 countries across six continents, this survey collects standardised information on many explanatory variables and how they relate to truth discernment. This project, more than just manipulating one mechanism, measured multiple explanations simultaneously. The aim was to establish the most influential mechanism when controlling others. Additionally, the project allows authors to test if the mechanisms act similarly across cultural contexts.

2.3.3 Other Methodological Approaches

Besides experiments and surveys, the literature on misinformation also employs other methodologies. One prevalent avenue of research is the analysis of social media traces. Social media sites like Twitter and Facebook inherently record vast amounts of information about the interactions they host. Researchers analyse those automatically-collected data to estimate the prevalence and potential audience of online misinformation. For example, Vosoughi et al. (2018) collected a set of rumours and then analysed how many times they were shared on Twitter. The authors then measured which characteristics were linked with longer "cascades" of sharing. Other studies combine social media data with survey data to test whether participants' stated and observed behaviours and beliefs coincided (Guess et al., 2019; Guess et al., 2020). Additionally, some designs combined digital traces with network analysis techniques to examine fake news' emergence and dissemination patterns online (Kopp et al., 2018).

In contrast, some studies analyse these textual data using qualitative content analysis. Instead of analysing large quantities of data through automated techniques, these papers focus on smaller samples of textual data. Authors then employed manual techniques to gain nuanced insights into the characteristics and themes of misinformation (Mourão & Robertson, 2019). In other cases, researchers utilise content analysis to explore how "fake news sites" present themselves (Robertson & Mourão, 2020) or how Facebook formulates its announcements to tackle disinformation (Iosifidis & Nicoli, 2020).

The availability of automatically recorded data does not preclude the adoption of qualitative methods. Especially more recently, a strand of studies embraced interviews and focus groups. For example, Chadwick et al. (2023) interviewed a sample of UK participants to understand how they manage fake news sharing in their WhatsApp group chats. In contrast, Steffens et al. (2019) used interviews to explore how Australian pro-vaccination organisations fight misinformation online. Finally, Papapicco et al. (2022) studied adolescents' perception of disinformation and racial hoaxes through focus groups. Additionally, some studies have combined different techniques to adopt mixed-method approaches. For instance, Tandoc et al. (2020) implemented a survey and a series of interviews to study how Singaporean social media users deal with misinformation.

2.4 Gaps in the Literature

While the literature addressing the question "Why do people believe in fake news?" has produced a varied and robust body of knowledge, it also exhibits specific gaps and limitations. Through my examination of existing papers, I identified four main gaps. Firstly, the majority of studies focus on manipulation and solution testing. On the contrary, measuring variables and mapping different mechanisms in the population are left mainly unexplored. Secondly, the exploration of potential explanations primarily focuses on cognitive ones, while social mechanisms receive much less attention. Thirdly, papers often test mechanisms with either unrealistic or arbitrarily chosen stimuli. Lastly, the literature is still at an early stage in studying misinformation across different cultures and languages.

2.4.1 Manipulation, Not Mapping

A large part of the literature tries explaining misinformation beliefs by resorting to experiments and, consequently, by *manipulating* certain variables to test their influences. This approach is valuable for establishing causation and replicable results. However, this type of academic production leaves some fundamental questions unanswered and, more concerning, not explored.

Firstly, The experimental approach alone is not particularly helpful in establishing the *average level* of misinformation beliefs. Before studying *why* people believe in fake news, it is crucial to grapple with preliminary inquiries such as: "Do people *believe* in fake news, and to what extent?". To answer the questions mentioned above the research needs to focus on the description of the population. On the contrary, experiments are not designed to describe the average level of participants' accuracy. Instead, they focus on measuring how that accuracy changes depending on whether participants are treated or not with an intervention. For example, they are not usually implemented on representative samples of the population. They thus do not allow conclusions to be drawn on populations' average levels of accuracy.

To measure the average level of accuracy, studies would also need to test participants' abilities on a "representative sample of items". In other words, a sample which appropriately mirrors information that is present online. This sample of items should include different types of information. In addition, it should reflect the frequency with which they are respectively encountered online. On the contrary, experiments usually measure their effect on a limited and rarely justified selection of items (see Section 2.4.3).

Secondly, the experimental paradigm is less adept at capturing the *distribution* of misinformation beliefs across the population. It struggles to answer questions like: "Do individuals hold similar beliefs? Are their reasons for believing (or not) the same?". Unlike experimental comparisons between treatment groups, these inquiries demand extensive measurements of multiple explanatory variables within a representative sample. Instead, experiments usually focus on the average effect of a single treatment. This setup does not allow for exploring whether different types of respondents behave according to different mechanisms.

Moreover, the urgency associated with the misinformation issue has led some experimental studies to attempt to "solve the problem" directly, potentially at the cost of a deeper understanding. In some cases, papers test potential solutions without measuring their distribution in the population.

A notable example is the limited subset of papers exploring the interplay between social norms and fake news. Instead of first examining the presence of these social norms, some papers in this subfield hasten to test the effectiveness of norm-based solutions without establishing their baseline existence (see, for example, Andı and Akesson, 2020).

Lastly, the experimental manipulation of single mechanisms prohibits their simultaneous comparison. Questions like "Which mechanism is the strongest?" or "Are some mechanisms dependent on others?" are better addressed through observational surveys rather than experiments. Only the measurements of multiple explanatory variables at once can provide insights into the relative strength and interdependence of different mechanisms.

To summarise, the predominant focus on manipulation left many research questions in the field of misinformation largely, but not completely, unexplored. In fact, some papers already approached misinformation beliefs using surveys, as we briefly mentioned (see Section 2.3.3). In some cases, they even tried to map different mechanisms at once. However, they share some limitations.

Firstly, a notable omission in existing "mapping" papers is the neglect of crucial mechanisms explored by the experimental literature. For example, surveys exploring explanations linked with political beliefs do not include measurements of cognitive effort and analytical thinking (see, for example, Faragó et al., 2020 or Goyanes and Lavin, 2018). This neglect does not allow us to link different explanations together. For instance, these papers cannot test whether partisanship's role is consistent when controlling for participants' cognitive styles. Additionally, they do not allow to test whether the effect of partisanship changes according to participants' cognitive styles.

In other instances, authors use different definitions of the dependent variables, hindering the integration of their results with those derived from experimental approaches. For instance, certain studies do not measure participants' ability to recognise fake news using an accuracy task. Instead, they gauge "recollection" of encountering and believing fake news (e.g. Chadwick et al., 2018). This approach assumes that participants, at some point, realized the information was fake. Nevertheless,

this is not always the case. In many instances, we believe in false information without realising it. Measuring only recollection precludes the exploration of these crucial cases.

Another prevalent issue in existing survey studies is their focus on specific topics. Consequently, they cannot generalise their conclusions to other topics or the issue of misinformation as a whole. Many studies focus on the COVID-19 pandemic, vaccinations, and related health issues. Take, for example, the above-mentioned work by Arechar et al. (2023). Despite its strengths, it draws conclusions about misinformation in general based solely on COVID-19-related information[3].

Lastly, many surveys measure perceived accuracy without allowing participants to express their hesitation (for example, see Arechar et al., 2023; or Pennycook and Rand, 2019). When measuring participants' ability to judge news through an accuracy task, these studies do not include any "I don't know" option (abbreviated to "DK option"). This choice can alter survey results by increasing the percentage of participants choosing wrong answers (Luskin et al., 2018). In fact, these surveys force respondents to choose an answer even when they have no clues about the veracity of what they are reading. Some answers will be correct by chance, but answers in all other levels will be wrong. Consequently, answers that would be DK if this option was available will be labelled more often as wrong than correct. In conclusion, this will result in an overestimation of misinformation beliefs.

Furthermore, this overestimation may interact with measured effects. For example, participants with different cognitive styles may have different propensities to choose random answers when DK options are absent. If true, this could have significant and unmeasured consequences on effects estimations.

[3]This is a conscious choice made by the authors. Their primary goal was to analyze misinformation comparatively across countries, and they argue that COVID-19 "provided a unique opportunity in this context as it allowed us to construct a set of true and false statements that were of global relevance" (Arechar et al., 2023).

2.4.2 Lack of Social Mechanisms

Another notable gap in the literature on fake news regards the explanations proposed. Much research delves into cognitive drivers, such as intuitive thinking, motivated reasoning, cognitive failures, and the illusory truth effect. Additionally, psychological drivers, including the role of emotions, and political explanations, such as personal views and partisanship, have been extensively explored (see Ecker et al., 2022 for a review).

In contrast, social explanations have received comparatively less attention. For example, a recent review mentions only one social driver when listing relevant areas of study (Ecker et al., 2022). This driver is source cues or the idea that individuals believe more in credible sources, elites, and in-group members. Other social drivers, like institutional trust and social norms, are only mentioned when proposing potential solutions. Questions like, "Is the interplay between people relevant?" or "Is the relationship between people and socially constructed institutions relevant?" have been less explored.

The imbalance in research development poses the risk of prematurely declaring the issue of fake news 'solved'. This risk involves slowing research before exploring all relevant avenues for explanation. Policymakers may then translate this incomplete understanding into policy.

This imbalance is not a problem per se. As long as we know at least one explanation of misinformation, we can start solving the problem. However, even stopping at the first solution could pose problems. First, relying solely on one solution, even if robustly proven effective, could lead to adverse effects if it disregards other pertinent but overlooked factors. For example, consider a hypothetical society where people believe in fake news because they lack trust in public institutions, beyond cognitive reasons. Proposing a state-based intervention to reinforce cognitive mechanisms, even if proven effective, may not work for those lacking trust in public institutions. It might even exacerbate the situation by further diminishing their trust. Second, science aims not only to solve problems but also to create knowledge. In this context, exploring new drivers should be intrinsically

valued for contributing to knowledge creation. Instead, the imbalanced focus on one explanation to rapidly transform it into an implementable solution may limit our comprehension of misinformation perception. While it is unrealistic to investigate every conceivable explanation comprehensively, I thus argue for exploring social drivers even in the presence of robust literature on cognitive drivers.

To be fair, some works already explored social drivers. In particular, Section 2.2.2 mentioned those that directly addressed social norms. Instead, the other two social drivers have not been addressed directly in past works. The literature on social norms in misinformation is emerging and in a phase of expansion. However, it is still scarce, recent, and limited.

Firstly, the two papers exploring social norms focus on analysing sharing (Andı & Akesson, 2020) and reporting behaviours (Gimpel et al., 2021). Instead, they neglect to examine the effects on the perceived accuracy of news. Thus, the impact of social norms on misinformation beliefs is largely unexplored. Secondly, they do not measure the presence and distribution of social norms, merely assessing the effectiveness of priming messages.

In addition, they have some methodological limitations. Andı and Akesson (2020) measure sharing behaviour for a single (false) news. This restriction prevents the determination of whether the effect is significant solely for fake news or also extends to reliable news. In other words, the absence of reliable news excludes the possibility of calculating participants' truth discernment.

Furthermore, the two papers do not use an existing reference network when nudging participants about a potential social norm. Instead, interventions refer to "most responsible people" (Andı & Akesson, 2020), or they simply state what would be appropriate to do (*"[...] It is therefore important that our users [...]"*, Gimpel et al., 2021).

2.4.3 Stimuli Selection: Artificial, Semi-artificial and "Natural" Stimuli

Moving to a more methodological perspective, another limitation in the present literature is the choice of stimuli. Many papers measured participants' accuracy using either artificial or semi-artificial stimuli. Using these types of stimuli creates problems with the external validity and

reliability of studies' conclusions.

To backtrack briefly, an accuracy task is a section of a survey or experiment in which participants must judge the accuracy of various stimuli. Many studies use this task to measure participants' truth discernment, i.e. their ability to discern false from accurate information. The stimuli in accuracy tasks can be news articles, social media posts, written statements, and others.

The selection of stimuli raises practical questions and trade-offs for authors. Realistic stimuli enhance external validity by measuring accuracy in a situation closely mimicking reality. They also do not allow for arbitrariness. Authors cannot modify natural stimuli to make them fit in a pre-established research framework. However, precisely for this reason, they may limit authors' control over the material presented to participants. For example, a study cannot use a stimulus if it does not present the feature which is the object of its inquiry. Furthermore, "natural" stimuli are more difficult to gather. They need to be manually collected, checked and validated.

In the early stages of this field, some notable papers used realistic stimuli, like articles' screenshots. A prominent example is Pennycook and Rand (2019), who collected a list of fake news articles from debunking sites and a relative list of articles from mainstream sources. The value of this dataset and the difficulty in its creation are confirmed by the fact that a series of related papers, like Bago et al. (2020) and Dias et al. (2020), used it in their measurements.

However, in many other papers, authors favoured control over external validity, creating statements from scratch (see, for example, Weeks, 2015). I refer to cases where authors create stimuli from scratch as *"artificial"* stimuli. In these cases, participants are not asked to judge information in a format that can be encountered in real life, such as articles, videos, posts, memes, and others. Instead, they are asked to judge "statements" or "claims". These are de-contextualised one-line sentences which contain a piece of factual information. Here is an example from Weeks (2015): *"Illegal immigrants benefit from U.S. public welfare programs"* (found in Supplementary Materials). Artificial stimuli give authors complete control over what they submit to participants. For

example, they can choose the message's characteristics, such as its topic, difficulty, and political leaning. Furthermore, the measurement is cleaned from all intervening factors. Participants can respond only based on what they know about this specific factual information, as they would on a school test. They can only respond based on the information or the misinformation they hold.

In other cases, authors used existing but modified pieces of information. I refer to those cases as *"semi-artificial"* stimuli. For example, Bauer and Clemm von Hohenberg (2021) use realistic Facebook posts. However, they create different versions of them. The authors present the same article with either an existing and known source or a fictional one. Similarly, they present it with either pro-immigration or anti-immigration content. These manipulations allow them to create different experimental treatments that vary on specific dimensions, all others being fixed.

To summarise, the use of artificial and semi-artificial stimuli allows researchers to have great control over what they submit to participants. In addition, it allows them to avoid costly procedures like stimuli selection, fact-checking and validation.

Nevertheless, these types of stimuli also pose a series of risks and limitations. The first risk lies in the possibility that a statement crafted by a researcher might be *structurally different* from a social media post encountered in everyday life. Even minor modifications aimed at creating new stimuli from existing ones could inadvertently result in unusual examples of information.

For example, Gimpel et al. (2021) submitted a news feed containing 15 posts to participants. Among these, five are fake news posts that circulated on Facebook. However, authors modified those posts to *"make it easier to identify them"*. This modification creates a problem of generalizability outside the scope of this specific study. In real life, that news would not have been so easily recognisable. In addition, it also questions the reliability of their conclusions, as it is unclear whether the effects they found would be replicated with "naturally difficult" fake news. Furthermore, "real news posts", or news that do not contain false information, were not modified to make them more reliable. This imbalance of intervention makes the study's conclusions even less reliable, especially

given that this choice is not justified or discussed.

Other than external validity, the problem with artificial and semi-artificial stimuli is that they are created by the same researchers using them to test their theories. This overlap introduces the risk of authors unintentionally or intentionally embedding the hypothesis to be tested into the material used for testing. When constructing these stimuli, authors interpret a (fake) news writer. In doing so, there is a risk that they interpret *their representation* of a (fake) news writer, potentially creating a piece of fake news as that fictional writer would. This choice is even more problematic if the authors do not report their stimuli creation procedure.

In conclusion, the use of artificial and semi-artificial stimuli in past works limits their external validity and, ultimately, the reliability of their results. This choice does not automatically invalidate their results. Those results would likely be replicated with more realistic stimuli. Furthermore, using controlled and easy-to-create stimuli is undoubtedly helpful for gathering initial evidence. However, the replicability of those results with more realistic and less arbitrarily chosen stimuli cannot be taken for granted. Instead, it would need to be tested empirically.

2.4.4 Scarce Evidence Outside Anglo-Saxon Countries

One notable methodological gap in the existing literature is the concentration of scientific production in English-speaking countries, particularly in experimental and survey-based studies (Pennycook & Rand, 2021b). For example, four out of five of the countries with the highest number of published documents on the citation database Scopus have English as their primary national language (see Figure 2.1).

This trend can be attributed to a combination of linguistic and socioeconomic factors that influence academic production. Firstly, institutions that lead research in other fields will plausibly be the first capable of financing research also in a globally emerging field. Many of these influential scientific institutions are based in English-speaking countries[4]. Furthermore, institutions based in

[4]It is difficult to objectively assess the quality of research in different contexts and academic fields. However, as an indication of the centrality of English-speaking countries in top positions of universities' ranking, see:

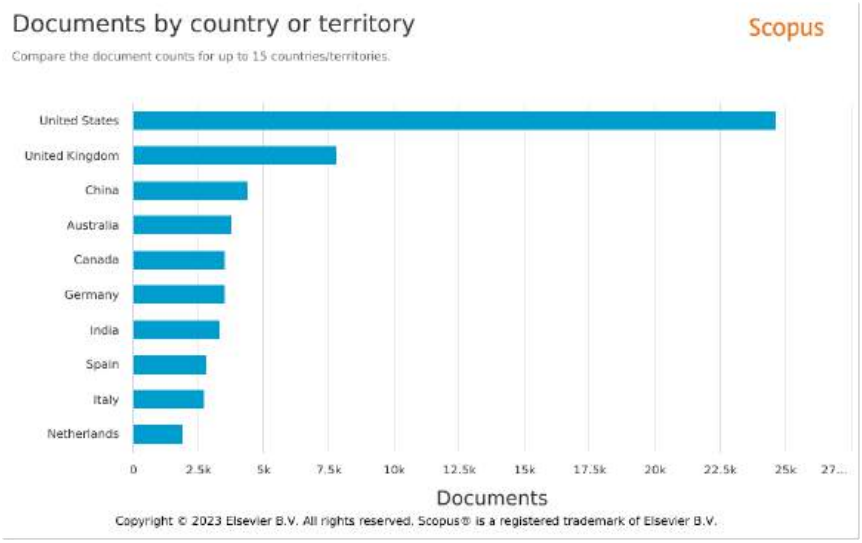


Figure 2.1: Documents collected on Scopus, by country or territory. Analysis made with Scopus' advanced research tool on April 14, 2023. Query string: *'misinformation OR "fake news" OR disinformation'*. The plot presents only documents published between 1990 and 2023.

English-speaking countries have a linguistic advantage. Despite the increased use of other languages, English is still the most used language online, representing 20% of online content alone (Pimienta et al., 2023).

It is reasonable to assume that the evidence from the literature can be applied globally, even if mainly based on English-speaking countries. Past studies often considered mechanisms at the individual level rather than the national or regional level. However, this assumption needs scrutiny. Unobserved national-level variables, such as language, political systems, and cultural dimensions, could play crucial roles and should not be dismissed a prior.

Relevant academic production already exists, even if more marginally, outside Anglo-Saxon countries (see, for example, Gimpel et al., 2021; Vegetti and Mancosu, 2020). However, many of these studies share similar limitations. Some of these limitations are particularly crucial for cross-cultural comparisons. In particular, to the best of my knowledge, no study tests the same drivers across different countries and utilised realistic yet comparable stimuli.

https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking Visited on April 18, 2024.

In numerous cases, studies in other countries test standalone drivers without establishing connections with already studied factors. This limits the possibility of comparison with drivers tested in other countries. Furthermore, each study remains isolated without establishing cross-cultural comparisons. Past works rarely tested the same drivers in multiple national contexts. Only recently, some papers started implementing similar surveys in different countries, as seen in the work of Arechar et al. (2023). However, this study uses artificial statements and focuses solely on COVID-19.

2.4.5 Theoretical Contributions

The present work aims to address the above-mentioned research gaps through five distinct studies. To briefly anticipate, the five major contributions of this project are:

1. The mapping of different drivers within the same surveys,
2. The exploration of social mechanisms,
3. The use of "natural" stimuli to measure participants accuracy,
4. The replication of the same research design in the United States and Italy,
5. The experimental manipulation of one driver: social norms.

Firstly, the present work aims to map the role of different mechanisms. Three comprehensive surveys measure and concurrently test the role of five cognitive and social mechanisms.

Among cognitive drivers, I revisit the role of lack of analytical thinking and motivated reasoning. These two explanations arguably received the most attention in the literature (Pennycook & Rand, 2021b). In addition, they both address underlying cognitive mechanisms that are linked to other factors and explanations. For example, intuitive thinking is thought to be linked to the use of heuristics (Kahneman, 2011). In the case of misinformation, one of those heuristics is the "illusory truth effect" (Pennycook et al., 2018).

Additionally, I measure three underexplored social drivers: anti-elitism, institutional trust, and social norms. These mechanisms have received limited attention in past works. The number of papers exploring these drivers is minimal, and they sometimes present crucial limitations. Nevertheless, there are theoretical reasons to expect that these drivers could be linked to truth discernment. Chapter 3 illustrates those theoretical reasons and the existing evidence supporting them.

Given the emphasis on mapping rather than manipulating these mechanisms, two primary surveys are implemented on similar quota samples of around 1200 participants. Quotas are chosen to represent the general population regarding age, gender, and education. Instead, the third survey involves a convenience sample of around 600 participants. Through these surveys, the objectives were to 1. measure the prevalence of each independent variable in the population, 2. establish correlational evidence regarding the link between these variables and the ability to discern news, and 3. compare the significance and magnitude of these understudied drivers with those already established.

Besides exploring mapping and social drivers, the third contribution of the present work is the use of "natural" stimuli. A new dataset comprising 80 social media news posts is constructed using a replicable and robust procedure. This dataset includes a balanced mix of Italian and US posts featuring fake and reliable news.

Two preliminary studies gather data to validate this newly constructed dataset of stimuli. In particular, these studies address participants' perceptions of the collected social media posts using various dimensions. This validation allows for future replication and the use of real-life stimuli from different cultural contexts.

The fourth contribution is that of expanding misinformation literature outside Anglo-Saxon countries. In particular, the same research design is replicated in the United States and Italy. The United States was chosen because of their centrality in the present literature (see Figure 2.1). This project is not just replicating other designs. It presents differences from other papers in terms of

methodology and content. Thus, it is essential to establish the project's conclusions in the same context explored by past works.

A second country is needed to test the implicit assumption of generalisability for mechanisms at the individual level adopted by most of the literature. In the present project, Italy was chosen. The reasons for this choice are multiple.

First of all, the choice is made on some important practical considerations. Language knowledge is needed to interpret the news content during the stimuli selection procedure. Knowledge of the country and its political context is needed for similar reasons, other than for the correct development of some questionnaire items.

Furthermore, Italy was chosen because it differs from the USA on myriad national-level characteristics. The first and most apparent is spoken language, a potentially critical factor in a language-mediated issue such as misinformation. However, the literature rarely addresses fake news in non-English languages. The potential existence of intrinsic differences between fake news written in different languages is even less addressed.

Regarding the possible role of motivated reasoning, institutional trust and anti-elitism, the USA and Italy differ in their political and welfare systems. Regarding the former, the USA has a clear-cut two-party system and a majoritarian electoral system (Sartori, 2005), while Italy has a fragmented party system (Ignazi, 2017) with a (semi-) proportional electoral system (Passarelli, 2017). Regarding the latter, the USA presents a typical liberal welfare regime, with a limited role of the welfare state in the economy and a strong responsibility left to the market. On the contrary, Italy's welfare regime is often categorised as conservative, with an important role of the State in the economy and the responsibility for individuals' well-being prominently left to the family (Powell et al., 2020).

Finally, Italy and the USA also differ on various national cultural dimensions, such as the ones theorised by Hofstede et al. (1991), which have been found to be correlated with fake news beliefs

(Rampersad & Althiyabi, 2020).

Finally, the project extends beyond mapping the identified mechanisms. The three primary surveys also include an experimental component in which one of the mechanisms (social norms) is manipulated to test its impact on the dependent variable.

This project advances a series of innovations in testing social norm interventions. First, it is the first project to test a social norm intervention on participants' truth discernment. Furthermore, the project explores the influence of a "negative treatment", highlighting the absence of a norm. In addition, this kind of intervention is tested on an unprecedentedly large pool of stimuli and using an existing reference network.

Chapter 3

Theoretical Framework

In this chapter, I will outline the functioning of the five mechanisms at the centre of this project. These mechanisms are:

1. Lack of analytical thinking,
2. Motivated reasoning,
3. Anti-elitism,
4. Institutional (dis)trust,
5. Social norms.

The first two mechanisms have been proposed and explored in the literature, especially in cognitive psychology. These are called *"lack of analytical thinking"* (Section 3.1) and *"motivated reasoning"* (Section 3.2). I will start by briefly outlining their common theoretical ground. In fact, both these mechanisms are based on *"Dual Process Theory"*. I will then illustrate how these explanations work and why they are expected to be linked to misinformation beliefs.

The other three mechanisms focus on social drivers. Other than just relying on their cognitive resources and pre-existing beliefs, I argue that individuals might as well be influenced by their peers and other social entities. The two cognitive drivers I mentioned focus on the role of resources within

individuals. Instead, these three social drivers focus on the relationship between individuals, their peers and the institutions mediating their access to news events and the public discourse.

The first social mechanism, anti-elitism, will be introduced by defining this concept in other works (Section 3.3). I will then illustrate three causal paths that justify the expectation of a link between anti-elitism and misinformation beliefs. These paths are 1. the application of motivated reasoning to anti-elitism, 2. source selection, and 3. reverse causality. The section will conclude by mentioning past works that indirectly support these causal explanations. None of these papers directly address the link between anti-elitism and fake news. Nevertheless, they gather evidence on the proposed causal explanations or some of their parts.

The following section (Section 3.4) will start by defining institutional trust and the various conceptualisations of institutional distrust. I will then outline the mechanisms that could link it to misinformation beliefs and how they relate to those regarding anti-elitism.

The last section of this chapter is dedicated to the theoretical analysis of social norms in this field (Section 3.5). In particular, I start by defining social norms more broadly. I then evaluate how they could be used to describe the interactions between individuals and fake news. I identify three potential social norms that could regulate this interaction. These could be about believing, reading, or sharing misinformation. A subsection clarifies the specific norm this project focuses on (sharing norm) and justifies this choice. After explaining how sanctions and punishments could work in this context, the section concludes by outlining the causal paths which could link this mechanism to misinformation beliefs.

In conclusion, this chapter illustrates the research questions that stem from this project's theoretical framework based on the five mechanisms mentioned above (Section 3.6). In addition, the other research questions address the methodological gaps in the literature outlined in Section 2.4.5.

3.1 The Role of Analytical Thinking

The first mechanism that explains misinformation beliefs is the lack of analytical thinking. To anticipate, it argues that individuals fall for misinformation because they fail to be focused enough to recognise cues of the veracity of what they read. Instead, they rely on intuition and heuristics, which are faster and less cognitively expensive.

Analytical thinking is based on a model of the human mind coming from cognitive psychology. This model is called *"Dual Process Theory"* (see, for example, Kahneman, 2011 or, for an earlier account, Chaiken and Trope, 1999). The second cognitive mechanism this project will explore, motivated reasoning, also builds on this model (Section 3.2).

Dual Process Theory models the human mind with two complementary systems: *System 1*, the *"intuitive system"*, and *System 2*, the *"analytical system"*. On the one hand, System 1 is the default mode often employed for easy tasks. It relies on fast heuristic answers to conserve cognitive energy. On the other hand, System 2 is activated for more complex tasks, requiring focused analytical thinking but demanding more cognitive effort. Within this framework, analytical thinking (System 2) usually leads to more informed decisions than the less effortful System 1 (see, for example, de Neys, 2012; Kahneman, 2011).

Gordon Pennycook and David G. Rand (2019), among others, use this theoretical model to explain why individuals fall for misinformation. According to this branch, individuals believe in fake news because they fail to engage in System 2 thinking. In other words, misinformation beliefs arise from a lack of analytical thinking. Individuals are not focused enough on processing the news they encounter. Consequently, they fail to gather and process the information they receive correctly. Instead, they rely on cognitive shortcuts and heuristics, such as familiarity or ease of reading.

Various papers gather evidence that individuals with a more analytical cognitive style are better at judging news than those with an intuitive style (for a review, see Pennycook and Rand, 2021b). Furthermore, other papers test interventions that foster analytical thinking to increase individuals'

truth discernment (for a review, see Pennycook and Rand, 2022).

However, this theory was born long before the debate on fake news. The idea that reasoning supports good judgment can be traced back to what has been termed the "classical reasoning" approach (Kohlberg, 1969; Piaget, 1932). Before the advent of misinformation in the public and academic discourse, the literature used this theory to explore debates over *"decision-relevant science"* (DRS) (Kahan, 2017). With this therm, authors identify all those scientific facts that can inform critical public decisions and policies. For example, scientific facts and evidence about the danger of nuclear power plants can inform the decision on whether to adopt nuclear energy and build nuclear facilities. Thus, explaining why intense debates and divisions exist over policies for which scientific facts could inform a final decision is difficult. According to authors in the classical reasoning approach, such disputes arise from an over-reliance by some individuals on System 1 (Kahneman, 2003; Stanovich & West, 2000). In other words, some individuals fail to accept relevant scientific facts that could change their opinions because they rely on heuristic and intuitive thinking.

3.2 The Role of Motivated Reasoning

The concept of *"motivated reasoning"* also originated in academic works well before the contemporary "fake news" academic debate. In particular, many papers shaping this theory initially focused on already mentioned political controversies over DRS, or *decision-relevant science*. Published during the late 90s and early 2000s, these studies sought to unravel the puzzle of persistent disputes over policy-relevant topics despite the availability of crucial scientific facts that could potentially resolve conflicts and inform decisions (e.g. Kunda, 1990 or Balcetis and Dunning, 2006). In other words, how is it that, for specific policy-relevant topics, achieving a consensus on "what is objectively true" appears to be an insurmountable challenge?

I have already mentioned nuclear energy as an example of a DRS debate, but others can be

named. For instance, consider the ongoing dispute over human activity's responsibility for climate change. Substantial scientific evidence proves that global warming is happening and that human-based CO2 emissions play a crucial role. Nevertheless, Republican and Democratic voters have widely different positions. The former are usually far more sceptical about admitting that global warming is caused by human activity and CO2 emissions (Kahan et al., 2011).

Similarly, scientific evidence on the effectiveness and safety of human papillomavirus vaccination should be enough to allow for a definitive decision on whether to implement it in schools. On the contrary, the consensus in the public debate is far from settled (Kahan, 2010). Why, then, do these debates continue besides the fact that sufficient scientific facts already exist to make a decision?

As we saw, some scholars have addressed these questions using the classical reasoning approach. However, authors in this second branch suggest different answers. The central proposition of motivated reasoning, as outlined by Kahan (2017), is that individuals do not always engage with information to understand reality accurately. Instead, they sometimes prioritise expressing their political identity, fostering solidarity with like-minded individuals. Individuals use information and knowledge to convey and confirm their identity and allegiance rather than to uncover objective truths.

At the core of this strand is the idea that motivated reasoning leads individuals to utilise their cognitive resources with a different goal: the formation and maintenance of *"identity-consistent beliefs"*. Even analytical thinking can result in errors when employed for motivated reasoning. For example, a study by Kahan (2013) finds that individuals with the highest cognitive reflection scores were more likely to exhibit ideologically motivated reasoning.

To summarise, authors in these two strands propose diametrically different answers to the same questions. According to scholars in the classical reasoning framework, misinformation beliefs are caused by a lack of cognitive effort and an over-reliance on intuition and heuristics. Scholars who favour the motivated reasoning explanation argue the opposite. Analytical thinking can also lead

to misinformed individuals if it is directed at confirming pre-existing political and cultural ideas.

Crucially for this thesis, these two approaches have been tested together in the field of fake news. In a highly cited paper, Pennycook and Rand (2019) conducted a series of studies asking participants to evaluate the accuracy of a series of headlines containing both true and false information. Cognitive style (analytical or intuitive) was measured with widely used Cognitive Reflection Tasks (CRTs). In addition, the authors measured participants' political affiliation and the perceived political leaning of headlines. The results support the classical reasoning approach, indicating that participants with higher CRT scores are better at discerning false from accurate headlines.

Surprisingly, this ability is observed regardless of the headline's political alignment with participants' views. In other words, participants with a more analytical reasoning style demonstrate better judgment even when agreeing with false items and disagreeing with true ones. This evidence suggests that beliefs in fake news stem from a lack of analytical thinking. Contrary to motivated reasoning, participants do not prefer politically concordant headlines.

Furthermore, participants with an analytical cognitive style do not exhibit *increased* belief in politically concordant headlines compared to intuitive participants. This result supports the classical reasoning account more than motivated reasoning. If this were a situation where motivated reasoning was at play, participants with a more analytical cognitive style would believe more in the headlines confirming their political ideas. They would use their higher reflection ability to confirm their ideas rather than discover the truth.

3.3 The Role of Anti-Elitism

In this section, I will explore the idea that individuals could be influenced, other than just by their cognitive abilities, by their relationship with the cultural, mediatic and political elite. Independently of our cognitive ability, our beliefs and media diets are still influenced by how we perceive the actors on the political stage. Crucially, our judgment of news is influenced by our trust in media, i.e. those

who provide us a glimpse into that stage. Furthermore, perceptions of broadly defined *elites* are crucial to misinformation beliefs. I argue that fake news often talks about elites and, more often than not, they do so in strongly negative terms[1].

In the following subsection, I begin by establishing a clear definition of the concept of anti-elitism. Subsequently, I present three mechanisms that justify my expectation of a potential correlational link between anti-elitism and misinformation beliefs.

Determining *which* causal link is actually in place falls beyond this project's scope. This section elucidates *why* I expect those variables to be correlated. However, this project does not directly test the existence of these causal links. Instead, it represents an initial step toward measuring this potential correlation.

3.3.1 Defining Anti-elitism

Anti-elitism is defined as a situation where *"elites are seen as corrupt, betraying and deceiving the people"* (Schulz et al., 2018). This concept denotes a negative attitude toward the political elite, which is not only viewed as harmful to citizens but also as malevolent. In essence, anti-elitism does not stem from the belief that the elite is merely incompetent. Instead, it emphasises the perceived intentional malevolence of the elite.

Anti-elitism is rarely used as a standalone concept. In fact, it is often defined as a sub-dimension of populism. Thus, academic production on anti-elitism is closely tied to the discourse surrounding populism. Similar to this concept, the term "anti-elitism" has gained significant attention in academic literature, along with the development of a proper operational framework, only recently (Schulz et al., 2018).

The literature often defines *populism* as a multidimensional concept comprising various sub-dimensions. The academic debate regarding which sub-dimensions to include remains unresolved. Typically, three main thematic areas are employed: sovereignty of the people, opposition to the

[1]In this project, I am not focusing on government-produced propaganda, which would produce a similarly strong evaluation of the established political power, but in the opposite direction.

elite, and a Manichean division between good and evil (see, for example, Castanho Silva et al., 2020; Schulz et al., 2018, and Akkerman et al., 2014). However, the precise combination of these dimensions and their definitions can vary among authors. Nevertheless, scholars of this field are concordant in assigning a role of central relevance to anti-elitism. In fact, this dimension is the common element to all conceptualisation of populism (Wuttke et al., 2020).

In this project, I have chosen to focus on anti-elitism over populism for several reasons. Firstly, the three mechanisms used to explain misinformation beliefs are not readily applicable to the other sub-dimensions of populism. Secondly, different papers include different sub-dimensions in their populism definitions. As a consequence, the choice of one conceptualisation instead of another would not be easy to justify. Thirdly, anti-elitism has consistently demonstrated a (negative) correlation with institutional trust (Wuttke et al., 2020), whereas populism as a whole does not consistently exhibit a clear link with institutional trust. The core of the mechanisms proposed in this project revolves around the relationship between individuals and political and mediatic institutions. Consequently, a strict link between the concept in consideration and institutions is essential.

3.3.2 Anti-Elitism and Misinformation Believing: One Link, Many Paths

Numerous potential mechanisms could explain a potential link between anti-elitism and misinformation beliefs. This section addresses three: motivated reasoning, source selection, and reverse causality. Observing a correlation between these two variables can suggest a causal link in various directions. One possibility is that anti-elitist beliefs may influence people's ability to discern the truth of what they read (*anti-elitism* $\rightarrow$ *truth discernment*). However, the opposite could also be true. The fact of having believed in some fake news in the past could make a person more anti-elitist (*truth discernment* $\rightarrow$ *anti-elitism*). Additionally, there is the possibility of a self-reinforcing mechanism between the two variables (*anti-elitism* $\leftrightarrow$ *truth discernment*).

This section will initially consider potential mechanisms in the first category. I identify two

explanations for the causal link from anti-elitism to truth discernment: motivated reasoning and source selection. The primary focus of this project is to explore this direct link, where anti-elitism makes individuals more susceptible to believing in fake news. Subsequently, I will investigate the reverse link, which suggests that low truth discernment could make people more anti-elitist. In conclusion, I will demonstrate that these two directional links are not mutually exclusive. Both links are likely to coexist, creating a self-reinforcing cycle.

Motivated Reasoning in Anti-Elitism

The literature on fake news has already found evidence that individuals tend to believe information that aligns with their preexisting (political) beliefs, regardless of its accuracy (Pennycook & Rand, 2021b). Authors often attribute this phenomenon to motivated reasoning, wherein people believe in fake news because they are motivated to protect and reinforce their cultural and political identities. Instead of critically evaluating the truth of what they read, their primary goal is to seek cues that confirm their existing beliefs (Kahan, 2017; Van Bavel & Pereira, 2018) (see Section 3.2).

While the concept of motivated reasoning is well-established, much of the existing literature has primarily focused on a single ideological continuum. In particular, many studies focus on the left-right spectrum, including its various contextual variations, such as the Democratic-Republican divide (e.g., Kahan, 2013). However, other ideological continuums have gained prominence in contemporary politics and public discourse. Notably, the populist-elitist continuum has proven to be increasingly influential (Hadiz and Chryssogelos, 2017; Krastev, 2007).

Motivated reasoning could explain a potential link between anti-elitism and misinformation beliefs. In other words, individuals might believe in fake news because they are motivated to confirm their anti-elitist (or elitist) preexisting beliefs rather than doing so to reinforce their Republican or Democratic views. This possibility has received limited attention in the literature.

In suggesting that anti-elitist individuals believe in fake news because it aligns with their view of elites as corrupt, I am making an assumption about their content. In particular, I am assuming

that fake news consistently presents an anti-elitist depiction of the world. This is not to say that *all* fake news depicts such a narrative. Instead, I argue that fake news more frequently depicts elites as corrupt and malevolent rather than as honest and genuinely concerned with the well-being of "the people". Some evidence in support of this assumption can be found in Mourão and Robertson (2019) and Baptista and Gradim (2022). This assumption could be further tested by conducting a content analysis of fake news. However, this analysis is beyond this project's scope.

Media Diet and Source Selection

Source selection is another mechanism that could elucidate a potential causal link between anti-elitism and misinformation beliefs. In today's media landscape, individuals are not uniformly exposed to the same information flow. Instead, each person has discretion in *selecting* what to read and from which platform.

I contend that anti-elitist individuals, in their selection process, will attempt to avoid media that they perceive as connected to the "elite". Some evidence regarding the perception of a media-elite link can be found in (Fawzi, 2019). This conception of the elite is thus broad, encompassing not only political figures but also cultural and mediatic ones. Individuals lacking trust in the political elite and believing that mainstream media are associated with that corrupt elite will seek new sources that are distant from that cultural sphere or, at the very least, they will try to avoid media perceived as linked to the elite.

In both cases, individuals will be exposed to more "alternative" and "fake news sources", ultimately leading them to place more trust in what the elite labels as fake news. Notably, this mechanism does not rely on the assumption made for the "motivated reasoning" mechanism. The first explanation was based on the assumption that fake news contains anti-elitist content. Instead, this second causal path assumes that individuals, especially anti-elitist ones, perceive mainstream media as connected to the ruling elite. Once again, the empirical testing of this assumption is outside this project's scope. Nevertheless, some papers seem to support it.

While limited, existing evidence suggests that anti-elitist beliefs may influence people's consumption patterns, potentially exposing them to content classified as fake news by elites. This evidence, primarily correlational and derived from surveys and digital trace data, reveals that populist individuals 1. consume more commercial TV news, tabloids and social media (Schulz, 2019); 2. tend to self-select media content that accentuates the divide between the "innocent" people and "culprit" others (Hameleers et al., 2017) and 3. consume more hyperpartisan news and less legacy press (Stier et al., 2020). However, a study from de Rooij et al. (2022) found that, while populists who distrust experts tend to have a more ideologically extreme media diet, those with anti-elitism beliefs maintain a favourable opinion of mainstream media and continue to consume them.

Reverse Causality: Becoming Anti-Elitist

The third and final explanation of a potential correlational link between anti-elitism and misinformation beliefs is reverse causality. Fake news actually shapes people's anti-elitist views. In contrast to the first two mechanisms, where I argued that anti-elitism might lead people to believe in fake news, I am now proposing the opposite. Misinformation could alter specific individuals' attitudes, making them more anti-elitist.

This causal path is based on the same assumption I outlined for the first mechanism. Fake news more frequently portrays elites as corrupt than as honest and concerned about the well-being of the "people". If this holds, an individual who believes a substantial amount of fake news will gradually become more anti-elitist.

Notably, the initial reason for believing in fake news is not crucial in this mechanism. It might be due to a lack of analytical thinking or because the news aligns with individuals' preexisting Republican or Democratic political beliefs. Regardless of the initial reason, believing in such news leads individuals to internalise the content of the fake news. If this content presents an anti-elitist narrative, then that individual will internalise this narrative, regardless of its legitimacy.

Some evidence in the literature corroborates this "becoming anti-elitist" mechanism. For in-

stance, Zimmermann and Kohring (2020) find that beliefs in fake news lead voters away from traditional governing parties toward right-wing populist parties. Similarly, Ognyanova et al. (2020) observe that participants in a panel survey who believe in misinformation exhibit lower trust in mainstream media. Additionally, Hameleers et al. (2022) note that fake news can alter trust in media, prompting changes in media consumption habits, often leading individuals to turn to alternative sources with anti-establishment viewpoints.

A Self-Reinforcing Cycle

The three paragraphs above discussed the mechanisms that could explain a connection between anti-elitism and misinformation beliefs. However, it is essential to emphasise that this separation was done to simplify their presentation. In reality, they could be interconnected and not mutually exclusive.

For example, consider a scenario where a reader initially believes in an "anti-elitist fake news" due to a lack of analytical thinking. Failing to discern the cues indicating its accuracy, this person unintentionally internalises the content of a fake news story about a fabricated corruption scandal. After several such interactions, the individual gradually starts believing that the current elite members are corrupt (the third mechanism). According to motivated reasoning (the first mechanism), this newly held belief makes this person even more susceptible to believing in fake news that confirms this narrative. As this anti-elitist belief becomes increasingly entrenched, the person begins to distrust mainstream sources and actively avoids them, ultimately seeking alternative sources (the second mechanism). Consequently, with increased exposure to fake news due to this shift in media consumption, the likelihood of the initial scenario occurring again rises, closing the cycle.

In this example, I began the cycle from the third mechanism, but the order of mechanisms could vary depending on the specific individual in question. For instance, a person might become anti-elitist because of an "exogenous shock" like a real corruption scandal and subsequently distrust mainstream sources. The cycle could initiate from source selection, but the outcome would remain

the same. In summary, the three proposed mechanisms are not mutually exclusive. Instead, they can work in concert, creating a self-reinforcing cycle.

3.4 The Role of Institutional (Dis)Trust

3.4.1 Defining Institutional (Dis)Trust

Trust in institutions, sometimes referred to as *political trust* (Levi and Stoker, 2000), is essentially a specialised form of trust. Defining trust in a broader context is a complex task, as it has been the subject of extensive academic research (see, for example, Bigley and Pearce, 1998 or the more recent work from Pytlikzillig and Kimbrough, 2016). The OECD provides a comprehensive definition of trust, applicable to both *interpersonal* and *institutional* trust, describing it as *"a person's belief that another person or institution will act consistently with their expectations of positive behaviour"* (OECD, 2017).

At the core of trust is an expectation, which can be positive, as when believing that others will act *in favour* of our expectations. However, it could also be negative, as when believing that others will not act *against* our expectations. This latter conceptualisation of trust is adopted in other definitions like those from Hakhverdian and Mayne (2012), Hudson (2006), and Levi and Stoker (2000). In both cases, trust is defined by an alignment between these expectations and others' behaviour.

Distrust, especially in institutions, can be conceptualised in two ways. The traditional view treats it as the polar opposite of trust on a continuum (Bigley & Pearce, 1998). According to this approach, distrust represents trust's *absence*. When I distrust an institution, I do not expect it to act according to my expectations. Trust and distrust can thus be measured with a single item for each actor, resulting in two possible outcomes: low trust (indicating high distrust) or high trust (indicating low distrust) in that trustee.

According to a recent perspective, distrust in *conceptually different* from trust, with distinct

determinants, characteristics and consequences (Van De Walle & Six, 2014). Distrust is not seen as the *absence of trust* but as the *presence of distrust*, an independent attitude with its distinct contents. Distrust can thus be defined as *"an actor's assured expectation of intended harm from the other"* (Lewicki et al., 1998). According to this conceptualisation, distrust is the presence of negative expectations about an institution. As a result, trust and distrust would require measurement using (at least) two items per trustee, leading to four possible outcomes rather than two.

In addition to the two classical combinations (low trust and high distrust or high trust and low distrust), this conceptualisation allows for two other combinations. Low trust and low distrust coexist when an individual has no expectations regarding an actor. Conversely, high trust and distrust are simultaneously present when an individual holds contrasting expectations toward an actor. Although the latter situation might seem counter-intuitive, it is essential to remember that, in this conceptualisation, trust and distrust are not opposites and can coexist. The literature supports this argument by highlighting real-life situations, such as cases where trust and distrust refer to different aspects of the same relationship (McKnight & Chervany, 2001), or historical examples like the collaboration between Stalin and Roosevelt during WWII, characterised by both trust and suspicion (McKnight & Chervany, 2001).

If one would adopt the last perspective (seeing distrust as a standalone attitude), one might wonder how this concept differs from anti-elitism. In both cases, we deal with situations where individuals *actively hold negative expectations* toward a particular actor. Additionally, the *intended harm* inherent in the distrust definition is semantically akin to that found in the anti-elitism definition. Malevolent intentions play a crucial role in both cases. What sets anti-elitism apart from institutional distrust, particularly in their definitions, is the object of these expectations. In anti-elitism, the object is broadly defined as elites. In contrast, in the case of institutional distrust, the object is usually specific institutional actors, such as the Parliament, the Government, or the judicial system.

This project will measure institutional trust in its traditional formulation, aligning with the approach commonly employed in standard surveys within sociology and political science (e.g. ESS, 2021; EVS, 2020; Haerpfer et al., 2022). Following the conceptualisation of distrust as *absence of trust*, these measurements typically use a single item per institution, often in the form of a question like, *"How much confidence/trust do you have in ...?"*.

The mechanisms proposed to explain misinformation beliefs align more with the second conceptualisation of distrust than the traditional one. However, this choice is also influenced by the relatively recent emergence of the definition of distrust as a conceptually distinct attitude from trust. As standardised measures of institutional distrust have yet to be firmly established, I will continue to use classical measures of institutional trust. This approach ensures that the results of this project can be compared with other research and surveys. Consequently, this project will employ both classical measures of institutional trust to assess *positive expectations* towards institutions and measures of anti-elitism from the literature on populism to account for *negative expectations* as distinct attitudes.

3.4.2 Institutional (Dis)Trust and Misinformation Believing: a Parallel with Anti-Elitism?

As we discussed in the previous subsection, anti-elitism and institutional trust are closely intertwined concepts. Consequently, we can anticipate that institutional distrust might follow similar causal paths as anti-elitism. However, we also noted that they are not perfectly overlapping concepts. Thus, some clarifications are warranted.

The connection between institutional *distrust* and misinformation beliefs varies depending on the definition of distrust we adopt. As mentioned, institutional distrust can be conceptualised as either *absence of trust* (as commonly done in the literature) or as *presence of distrust* (as it is more recently suggested).

Adopting the latter approach, institutional distrust is conceptualised as a *distinct* attitude.

Rather than just referring to the absence of trust, in this approach, distrust represents the presence of negative expectations towards the trustee, i.e. public institutions in this case. In this conceptualization, distrust closely resembles anti-elitism. In both cases, public institutions are perceived as corrupt, led by individuals with malevolent intentions, from whom it is reasonable to anticipate negative actions. Therefore, when conceptualized as *presence of a negative expectation*, I contend that institutional distrust will follow the same causal paths as anti-elitism.

The only thing that differs between mechanisms regarding anti-elitism and institutional trust in this conceptualisation is the subject of these beliefs. While anti-elitism refers to beliefs about elites, we are now addressing beliefs about public institutions. For example, the explanation based on motivated reasoning would be: people believe in fake news because they confirm their preexisting negative beliefs *about public institutions*.

Conversely, if we embrace the perspective of trust and distrust as two extremes on the same continuum, institutional distrust would be conceptualized just as *absence of trust*. Consequently, the causal links between institutional distrust and misinformation beliefs would align with those of anti-elitism only in those occasions where *holding a negative belief* is not necessary for the mechanism to be present.

Motivated reasoning, to be present, requires a belief to be affirmed. If we conceptualize institutional distrust as *absence of trust*, then fake news content will not confirm any negative belief. In other words, in this definition of distrust, I do not anticipate institutions to be malevolent, I *do not expect* them to do "what is right". Consequently, when I read news depicting an institution as corrupt, it will not resonate with any belief I hold, and I will not be inclined to accept it to protect my existing beliefs.

Regarding *source selection*, the expectation is less straightforward. Would I attempt to search for alternative sources if I did not expect public institutions to do "what is right" (e.g., due to incompetence or technical complexity)? Or would I require the *presence* of a negative belief to take

an active decision, such as avoiding mainstream sources or seeking new ones? This is an empirical question beyond this project's scope. However, I argue that the second case is more likely than the first: source selection necessitates a belief's *presence* rather than just the *absence* of another one. In other words, if we conceptualize distrust as *absence of trust*, I contend that the source selection mechanism may not be present.

The last explanation for the correlation between anti-elitism and misinformation beliefs is the notion that fake news, due to their assumed negative portrayal of elites, could lead to individuals becoming anti-elitist. I called this explanation *reverse causality*. This mechanism could also be applied to institutional distrust, even when conceived as *absence of trust*. However, this adaptation would need two considerations. First of all, the assumption would now be that fake news often depicts a negative image of *public institutions* rather then *elites*. Secondly, believing in these narrative would have an effect on *institutional trust*, by lowering it, rather then on *anti-elitist* beliefs. In summary, misinformation beliefs could be correlated with low institutional trust because fake news could reasonably be assumed to depict public institutions as generally *bad* (whether that means *dysfunctional* or *corrupted*).

3.5 The Role of Social Norms

This section is dedicated to exploring the idea that misinformation beliefs could also be influenced by perceptions about how peers interact with them. Independently of our cognitive ability, we always have some variation in the amount of cognitive effort we put into the task we are doing. Different factors determine this amount. I argue that one of these determinants is the beliefs and expectations we hold about how other people interact with fake news.

In simpler terms, all else being equal in cognitive abilities, if I think that the people around me are generally unconcerned about the accuracy of the information they read and share, I am more likely to feel justified in putting minimal effort into my actions. On the contrary, if I believe that:

1. My peers are actively striving to ensure the accuracy of the circulating information,

2. They expect me to do the same and, in some instances,

3. They might even impose sanctions if I fail to comply;

then I will feel a sense of obligation to follow suit. This impression could then lead me, consciously or not, to invest greater cognitive effort in interacting with the news I encounter. More formally, I posit the theoretical feasibility of the existence of a social norm, or multiple social norms, pertaining to fake news[2].

Crucially, I am not arguing that this explanation should be an alternative to cognitive effort or that it should substitute this driving factor. On the contrary, I argue that these two mechanisms could be strictly linked. Specifically, I contend that social norms may serve as one of the determinants of cognitive effort, with individuals potentially varying their cognitive effort based on their perception of this social norm.

Another way social norms could shape the processing of information in news is via meta-reasoning. Beyond enhanced focus, the inclination to adhere to norms might prompt readers to scrutinize their biases more attentively, seeking to minimize them. In other words, the commitment to avoiding norm violations could encourage individuals to be more cognizant of their own cognitive flaws.

In the next paragraphs, I define the concept of general social norms by applying the framework developed by Cristina Bicchieri. I then show how this framework could be applied in the context of fake news and how punishment and sanctions could work in this context. In summarising the theoretical framework, I will elucidate how this mechanism could work from start to finish.

[2]I must add that this idea is not completely new. Although very scarce and recent, some evidence is already present on the role of social norms in misinformation perception (Andı & Akesson, 2020; Gimpel et al., 2021). See Chapter 2.

3.5.1 Concepts and Definitions: What Do We Mean with Social Norms?

Social norms are shared behavioural rules that prescribe actions followed because of reciprocal expectations and potentially social punishment (Bicchieri, 2006). To elaborate on this definition, we are talking about *informal* rules, distinct from formal regulations such as laws. They are considered *shared* norms because they must be known and acknowledged by multiple people, transcending the realm of personal beliefs.

These unwritten rules regulate many behaviours, shaping various aspects of our daily lives. To illustrate with some practical examples explored in the literature, norms can dictate commonplace experiences such as when smoking is allowed and whether one should allow people off the subway before boarding. Moreover, they extend to more profound, life-altering events, as exemplified by the norm surrounding child marriage (for a review, refer to Legros and Cislaghi, 2020).

Other than being *known*, social norms can also be *followed*, as individuals can align their behaviours with normative prescriptions. Conformity happens because people hold *expectations* regarding how others will respond to eventual violations. In this sense, sanctions often come into play, yet they remain *informal*, lacking enforcement by any formal organization or coded procedure. However, social norms are primarily followed because of the above-mentioned *social expectations*.

Individuals possess an *empirical expectation* when they believe others are following the norm. Instead, a *normative expectation* arises when they believe that others not only view adherence as the correct course of action but also anticipate that they themselves will comply. Crucially, according to this theoretical framework, many individuals are *conditional followers* of social norms. They adhere to a norm only under the condition that a. others are following it (empirical expectation) and b. that others consider conformity the morally right choice (normative expectation). In essence, *"a social norm is a collective practice sustained by empirical and normative expectations and by preferences conditional on both these expectations"* (Bicchieri, 2006).

A final noteworthy aspect is that norms can take either a *prescribing* or *proscribing* form (Bic-

chieri, 2006). In prescriptive norms, individuals are guided on what they *should do*, as seen in norms encouraging passengers to give up seats for older people or suggesting that students should engage in drinking during parties. Conversely, proscriptive norms dictate what individuals *should not do*, exemplified by directives like "do not stare at someone you pass by" or "do not touch people you do not know when talking with them".

3.5.2 Norms Governing "Bad" Beliefs and Their Expression

The literature applied the framework of social norms in many contexts. Authors often focused on observable behaviours or events, like drinking (Lewis & Neighbors, 2006), littering (Cialdini et al., 1990), smoking (Nyborg & Rege, 2003), child marriage (Bicchieri et al., 2014), hand washing (Andrighetto et al., 2024) and numerous others.

More recently, this field has expanded its scope to the analysis of more abstract phenomena, including the acceptance of far-right political affiliation (Valentim, 2021), homophobia (Pereira et al., 2009), religiosity (Stavrova et al., 2013), racism (Álvarez-Benjumea & Winter, 2020) and others. I argue that these norms not only regulate *appropriate behaviours* but also influence the acceptability of (expressed) *beliefs*.

Why is this distinction crucial? Why should we distinguish *"social norms concerning behaviours"* from *"social norms concerning beliefs"*? The significance lies in the fact that, in these *"norms of discourse"*, expressing a belief assumes a central role. Moreover, the interpretation assigned to this expression ultimately determines its potential for punishment. In other words, the *"modality"* with which one expresses a belief becomes much more crucial than the modality with which one would perform a behaviour.

To elucidate further with examples, consider a hypothetical society where an anti-littering norm is well established. In such a society, an individual who litters would likely face sanctions, regardless of whether their action was intentional, accidental, or done in jest. Conversely, I contend that socially regulated beliefs may occasionally be expressed without triggering any sanctions. Sanctions

are typically applied to regulated beliefs only when their expression clearly communicates that the sender indeed holds them.

An actor portraying a historical figure or fictional character will typically not face sanctions for expressing Nazi beliefs within the scope of their role. The same is true for a historian citing historical speeches. Joking about Nazism, too, may not necessarily result in sanctions. However, this also depends on the way in which it is done and, notably for this argument, the specific country and social context in which such humour is presented. In other words, the modality of the expression itself is regulated by the norm, a facet which is much less central in norms governing behaviours.

I contend that this principle also holds true in the context of fake news. I argue that "believing in fake news" could be considered socially inappropriate per se. Consequently, expressing support for fake news will only face sanctions when it signifies that the speaker genuinely endorses it. Sharing a piece of fake news for satirical purposes or as part of an analytical discussion is unlikely to incur sanctions (more on this in Section 3.5.4).

3.5.3 Social Norms About What? Believing, Reading and Sharing

When discussing a potential social norm, the first thing to establish is its object: What behaviour is it regulating? When speaking about "traditional" social norms, this question may seem trivial: a norm about littering regulates throwing rubbish outside its appropriate containers.

However, in the context of fake news, the object of a possible norm, if existent, is much less evident. Our interactions with news, particularly with fake news, can take many forms. To simplify the picture, we can identify three main ways of interacting with news: believing, reading, and sharing.

Don't Be Fooled by Fake News!

As already said in the previous paragraph, there are some norms that, instead of regulating a behaviour (like drinking or littering), regulate a belief (e.g. holding homophobic views or racist

ones or having a far-right political positioning). In all of these cases, the real core of the norm is a belief, while the expression of that belief is instrumental in making it evident to observers that the belief is there. I argue that this is the case also for misinformation beliefs.

If we would want to "write" this norm or to explicitly formalise it, it could take the form of something along the lines of *"Don't believe in fake news"* or, to make it more normative, *"Don't be fooled by fake news"*. When this norm is present, the fact of believing in fake news is expected to be, in itself, intrinsically inappropriate and to make the eventual transgressor feel ashamed.

But is not this norm a tautology? If I recognize a news article as fake, it is evident that I do not believe in it. In other words, disagreeing with this norm seems *logically incoherent*. Stating "I believe in this fake news" would inherently contradict itself.

Under these terms, violations seems to be possible only inadvertently: "I thought it was reliable, but I now realize it was fake". In this accidental transgression, an individual is violating the norm just until a sanctioner points out the error to them. Notably, during the transgression, the individual was not aware of it. Consequently, the individual was still not disagreeing with the norm, they just ignored their mistake.

Even deliberate transgression does not result in disagreement with this norm. For example, an individual may still be convinced that the information is true, even after others judged it as false. In this case, they are deliberately violating the norm in the eyes of others, but not in their own. After all, they are still convinced to believe in a true information. Remarkably, the norm is still endorsed by both sides of the picture. Even the transgressor still agree with the norm and does not concede that they are transgressing it.

So is that it? Is this norm intrinsically endorsed by everyone? Not necessarily. I can identify (at least) two scenarios where disagreement with the norm is logically possible. Both scenarios involve situations where the reader "does not recognize" the norm.

First, it is possible for individuals to believe that all news is inherently true. Possibly because

they possess very low media literacy, individuals might assume that all information reaching them is intended for genuine and accurate information dissemination. In this scenario, the concept of fake news itself may be absent from their understanding. Consequently, adherence to the norm is impossible, as subjects do not conceive its existence.

Another situation, distinct in nature, arises when individuals do not care, for different possible reasons, about accuracy or truth. For example, they might believe that terms like "accuracy" and "truth" are meaningless constructs and that a shared, objective reality does not exist. In this case, non-compliance with the norm occurs because they challenge the notion that it is possible to definitively classify news as containing false information, from an ontological perspective. An individual might acknowledge that the norm exists and that others comply with it. However, they still not agree with it because they do not acknowledge its premises, i.e. that a transgression can be even assessed.

Accuracy Motivated Reading: Be Aware of What (and How) You Read

When we think of social norms, reading is perhaps not the first behaviour that springs to mind. Reading is an intimate gesture seemingly devoid of immediate consequences for others, and particularly in modern democracies, it could be seen as a realm where regulations, even informal ones, should not be introduced. In fact, the literature on social norms in the context of reading is remarkably scarce, if not inexistent.

However, one can readily identify numerous instances of reading behaviours that may be deemed inappropriate and liable to evoke sanctions. For instance, reading a pornographic graphic novel in a public setting is likely to be perceived as inappropriate and may prompt bystanders to intervene. Moreover, as in every social norm, the appropriateness of the behaviour is subject to contextual variations. Publicly reading the Communist Manifesto would likely have elicited more disapproving reactions in the McCarthyist-era United States than in Soviet Russia. In support of the existence of some social connotations of the act of reading, one could also mention historical episodes of book

burning (Bosmajian, 2006). During these episodes, political actors, whether institutionalized or not, destroyed "dangerous" or "unacceptable" books to prevent their practical reading and to send a clear message to potential audiences.

Theoretically, I posit that a "reading social norm" could potentially exist in the context of fake news. I argue that reading blatantly fake news could be seen as inherently inappropriate, prompting a normative response. For example, an individual would sanction a fake news-reader by saying something akin to *"you should not read this"*.

One might argue that this sanction would likely be preceded by the question, *"Why are you reading this?"*. Returning to the *belief-expression* argument discussed earlier (Section 3.5.2), it becomes evident that the true essence of these reading norms does not lie solely in the act of reading itself but rather in the beliefs conveyed by this act. For instance, a journalist reading fake news for the purpose of debunking it would likely be viewed as entirely appropriate. This allows me to introduce a second layer of these potential "reading norms": the *modality of reading*.

There are many reasons for reading a document. We can read it to learn something, as when a student reads a school book, or to be entertained, as when somebody reads fiction. Obviously, the same document can also be read with many different motivations.

I contend that the motivation behind our act of reading a document significantly influences the specific dimensions of that document we prioritize. This, in turn, shapes our thought processes during reading and ultimately impacts the meanings we derive from it. Accuracy, or whether the document describes something that really happened, represents merely one of several features, and for many types of documents, it may not even be the most pivotal.

Finally, even if it is true that the same document can be approached with various modalities, I assert that documents often possess *"default modalities"*. These default modalities represent socially regulated and commonly accepted ways in which documents are typically approached. A novel is usually read for entertainment, while a school textbook is typically read for educational purposes.

Turning to the theoretical possibility of a "reading news social norm", I argue that this norm primarily pertains to the *default modality* with which news should be approached. In essence, the norm would regulate the typical purpose and manner in which we engage with news. More explicitly, it would advise us that *"when reading a news to inform yourself, you should be primarily interested in it being true"*.

As a result, a compliant individual would first be pushed to focus on searching cues indicating that the news is accurate and evaluate the news based on these indicators. Importantly, other observers would also judge the reader based on their endorsement of the news, primarily in relation to its accuracy. This emphasis on accuracy in the *default modality* of news reading is why a reader of fake news could potentially attempt to evade sanctions by specifying that their intent is not to gather accurate information.

Don't Share Fake News!

In addition to believing and reading news, I posit that it is also possible to think of a norm that regulates what people share. In this work, I will analyse sharing in the context of online interactions, but most of what I will say could also be applied to offline interactions and sharing in the form of speaking, for example.

In this case, the object of the norm would be the act of sharing fake news, and it could be formulated as a simple directive: *"Do not share fake news"*. In this formulation, the sole act of sharing fake news would violate the norm, independently of the motivation with which the news has been shared.

However, similarly to what is already said for "reading", there are various motivations (or modalities) with which we could be sharing the news: to inform others about an event because we think it is interesting, because it confirms our positions, etc. I argue that it is also possible that the norm would punish sharing only when done with intents linked to its accuracy. Thus, punishment will not be present when the news is shared ironically or to foster a conversation. A more refined

formalization of this norm would then be *"When informing others, do not share fake news"* (more on this in Section 3.5.5).

In this context, the norm was formulated as proscribing, indicating that a certain behaviour should be avoided. However, it is important to note that this norm could also adopt a prescribing form. An alternative formulation could be: "You should only share accurate news". Though only subtly different, these two formulations may yield distinct behavioural outcomes. However, determining whether this norm has a prescribing or proscribing form in a specific context is an empirical question.

So What? "Sharing" Norm as a Proxy of "Believing" Norm

To summarise, it is possible to think of different social norms for each interaction outlined above: believing, reading, and sharing. I argue that the core of an encompassing norm about fake news would be a norm about believing: *"do not be fooled by fake news"*. The other two norms refer to the "believing norm" at their core. In all three cases, the final sanction is "shame for having been fooled by fake news".

In this work, I will focus on the empirical measurement of only one of them: sharing. Why so? Firstly, I think a norm about believing is difficult to explain to eventual respondents, with important consequences on the reliability of any measure trying to capture it. Secondly, a measurement of a social norm about believing would be intrinsically influenced by a paradox. If I say "I am OK with believing this fake news" (the opposite of "I do not want to believe in/be fooled by this fake news"), I am saying that it is fake. Thus, I am not believing it. When I realise I am being fooled, I am not fooled anymore. This makes measuring the "believing norm" very difficult or impossible.

Instead, I propose to focus on the norm of sharing and use it as a proxy for believing. First, it is much easier to understand and explain because it is an observable behaviour. Second, because it is not influenced by the said paradox. I can still decide to share news that I know to be false, even when informing others. Furthermore, I argue that the sharing norm can be used as a proxy of the

believing norm because this could also be the case in people's minds. If I know that believing fake news is socially inappropriate, I infer I should also not share it, given that, by doing so, I could end up making people believe in it. Furthermore, if I know I should not share fake news, I am also forced to understand if it is false before sharing it. That is to say, that, to comply with the sharing norm, I must also comply with the believing norm, although the contrary is not true: I can share without believing.

3.5.4 What is Punished and What Are the Sanctions?

Even if punishments and sanctions will not be directly studied in the present project, the following theoretical discussion further explores the above-mentioned social norm and shows how it could work. I will start by reasoning on what behaviours could be punished and what sanctions could be used in the context of fake news. I will finish by speculating on whether the hypothesized sanctioning behaviours are, in fact, sanctions or just an attempt to stop an unwanted behaviour without any normative content against the transgressor.

Punished behaviours

In the context of a social norm about sharing fake news (e.g. *"Don't share fake news!"*), the punished behaviour would indeed be *the act of sharing fake news*. But is that all? Does the motivation (or modality) with which the news has been shared play a role?

We can envision two scenarios. We can imagine the existence of a *"strict norm"*, in which the sole act of sharing fake news is punished, regardless of the motivations behind the sharing. Alternatively, we can also consider another version of the said norm, a *"motive-driven norm"*, in which sharing fake news is only penalized when specific motivations and methods are involved.

Both of these norms are theoretically possible, and determining which type of norm is in effect in a specific context is an empirical question beyond the scope of this document. Nonetheless, we can speculate about the motivations that might be punished under the "motive-driven norm".

I contend that sharing fake news is unlikely to result in punishment when the sharer explicitly acknowledges the falseness of the shared information. For example, this can occur when the sharing is intended to initiate a discussion about a news item or the concept of "fake news" itself (on a meta-level) or when the sharer is signalling the news as fake (e.g., "Warning: this news is fake").

Additionally, I posit that fake news sharing will rarely be punished when the primary objective is not to inform. For instance, consider a stand-up comedian incorporating a news story into their routine. Unless the performance explicitly aims to provide factual information ("infotainment"), one should not anticipate strict accuracy in the comedic act, and consequently, sanctions against the comedian would be rare.

In summary, in a "motive-driven norm", the sharing act is subject to punishment when two conditions are met. First, the sharer's intention, whether genuine or feigned, is to inform others (rather than employing irony or other motives). Second, the sharer implicitly or explicitly communicates a belief in the news's truth. This combination arises in two scenarios: a. When the sharer genuinely mistakes the fake news as true, or b. When the sharer knowingly spreads the fake news with the intent to deceive recipients.

What are the sanctions?

Having discussed the potential behaviours subject to sanctions within the context of a social norm regulating the sharing of fake news, we now turn our attention to the crucial matter of enforcement. This section delves into the various sanctions that could be employed to uphold such a norm.

First and foremost, it is important to note that formal sanctions targeting fake news already exist or have been proposed in various legal systems around the world[3]. However, in this paragraph and this work, my primary focus is on informal sanctions, typically a distinctive feature of social norms.

The most basic response to shared fake news is a *correction*. For example, an individual could

[3]https://www.poynter.org/ifcn/anti-misinformation-actions/. Visited on 4 October 2023.

try to correct the sharing of fake news by stating something on the line of *"this is fake news"*. This intervention communicates to the sharer and the audience that the information is false, in an attempt to stop it and correct potential mistaken beliefs that result from it. However, someone may argue that this does not qualify as a true sanction since it lacks a normative or moralizing component and does not necessarily aim to "punish" the sharer. If we believe a sanction should have reputational consequences for the individual being sanctioned, we can consider two potential sanctions within the context of fake news.

The sanctioner may attempt to damage the transgressor's reputation by *portraying them as either "dumb" or "evil"* (or potentially both). In the first scenario, this occurs when bystanders perceive the transgressor as sharing a fake news item while being "dumb" enough to believe it is true. The effectiveness of this sanction increases when the fake news is particularly blatant. In the second scenario, bystanders must believe that the sharer knowingly spread fake news with the intent to deceive others. Alternatively, bystanders may perceive the sharer as uninterested in the accuracy of the news (and thus not necessarily believing in it) but sharing it for personal, partisan (political, economic, reputational, etc.), or ulterior motives.

Finally, I argue that another negative consequence of violating the sharing social norm would be of *shame*. Whether we want to categorize this as a "self-inflicted punishment", thus identifying it as a sanction, or not, is debatable. Nonetheless, the experience of shame in response to sharing fake news may indicate the social norm's existence, considering the central role attributed to this emotion in the social norm literature (Elster, 2007).

The most plausible situation in which one could experience shame after sharing fake news is the "dumb case", where they initially believed it to be true, only to realize later it was fake. Again, the intensity of this shame would likely be greater when the fakeness of the news was relatively easy to discern. Importantly, the extent of the "audience" for this mistake matters: if "nobody saw" or "nobody cared", the individual may find it easier to overcome the shame. This observation suggests

a possible social dimension to this shame, closely tied to our normative expectations for others.

3.5.5 How the Social Norm Mechanism Works

How might a norm that regulates news sharing shape people's interactions with fake news? Could it influence whether they ultimately believe in misinformation?

Let us begin with a scenario in which an individual is aware of the "sharing accurate news" social norm and adheres to it. In this context, they hold certain expectations regarding the behaviours of others. They expect others to exert their utmost effort to avoid sharing fake news (an empirical expectation). Moreover, they also anticipate that others believe that *"putting all our effort into not sharing fake news"* is the correct course of action, something that everyone *"should"* do. Essentially, they think others expect this course of action (normative expectation).

Let us now suppose that that individual is faced with a piece of news, the accuracy of which (whether it is fake or reliable) is unknown. They find the news interesting and are contemplating whether to share it. Because they adhere to the norm, they will, before sharing it, make an effort to determine the news' accuracy level. This increased cognitive effort will enhance the individual's ability to evaluate the news, assuming all other factors are constant.

Even if the individual did not intend to share the news, they contend that the norm communicates something about the significance of "accuracy" within their reference network, including what they should believe. If no one wants them to share fake news, then this "fake news" must carry some inherent negative value. Otherwise, why would everyone be so concerned? In essence, I argue that an individual can deduce the existence of the "believing norm" from the presence of the "sharing norm". The reasoning could go as follows: *"If nobody wants me to share fake news, perhaps it is also inappropriate to believe in them and, in general, I should prioritize the accuracy of my beliefs".*

Paradoxically, knowledge of the sharing norm might have a reverse effect. If an individual anticipates that everyone diligently refrains from sharing fake news, they may assume that the

quality of shared content is generally high or carefully assessed. As a result, they might be tempted to reduce their cognitive effort (which is intrinsically tiring) when evaluating news since they expect circulating news to have been scrutinized by others.

But what if an individual thinks the norm is not there? Again, they come across an interesting news article and contemplate sharing it. This time, the individual assumes that people in their social network do not prioritize the accuracy of what they share. The individual knows that others do not expect them to exert any effort in verifying the truthfulness of the news. The individual is also aware of the inherent difficulty of this task. In this context, sharing fake news seems just as acceptable as sharing reliable information. Consequently, they may question why they should invest the effort to discern between the two, leading them to be less diligent in evaluating what they share.

To summarise, I argue that a social norm about sharing fake news will influence the extent to which individuals believe it is appropriate to exert cognitive effort in assessing news, even when they are not directly considering sharing them. This increased cognitive effort, in turn, enhances their ability to discern the truth.

3.6 Research Questions

This chapter will conclude by summarising the research questions that stem from the research gaps highlighted in Chapter 2 and by the theoretical framework outlined in this chapter. All the research questions and hypotheses have been pre-registered on OSF at https://osf.io/hvdjz.

The first research question regards the absence of mapping of different, especially social, mechanisms:

Research Question 1: What are the key mechanisms influencing how people (mis)perceive the credibility of fake news? In particular, are there any *social and political* mechanisms?

Thus, the first aim of this project is to map the existence of five mechanisms thought to be linked to misinformation beliefs. These are the lack of analytical thinking, motivated reasoning, anti-elitism, institutional trust, and social norms of news sharing.

However, this work is not limited to mapping, as I manipulate the mechanism of social norms. The aim is to measure if perceptions of a social norm about the importance of sharing only accurate news can be shaped and with what results. Thus, the second research question of this work is the following.

Research Question 2: Can a social norm information treatment highlighting either the presence or the absence of a social norm regulating news sharing shape fake news perception?

Finally, regarding the cross-cultural comparison:

Research question 3: Do the findings from Research Questions 1 and 2 vary between Italy and the US?

Chapter 4

Research Design

The empirical evidence forming the core of this project is gathered through five distinct studies in Italy and the United States (see Table 4.1). These studies can be categorized into two validation studies and three main studies.

The main hypotheses are tested using the three main studies. The study named *"Convenience-ITA"* aims at testing the survey structure and replicating the other two main studies. It involved a convenience sample of approximately 600 online respondents, recruited by the company PollStar in Italy. Results from this study did not highlight the need for changes in the survey structure and items. Consequently, the other two main studies implement the exact same questions and structure as *Convenience-ITA*.

The fourth study, named *"Main-ITA"* is also implemented in Italian, targeting a quota sample of approximately 1200 Italian online respondents, representative of the Italian population on age, gender and education. The fifth study, named *"Main-USA"* and implemented in English, surveys a similar quota sample of around 1200 online respondents from the United States. Quota are chosen to be representative of the US population on the same demographic variables. Both studies, *Main-ITA* and *Main-USA*, are distributed via Qualtrics.

As anticipated, studies *Convenience-ITA*, *Main-ITA* and *Main-USA* share almost identical structures and items. The only differences between the two Italian main studies and the US one regard

contextual differences between the two countries. For example, parties reported in political affiliation questions differ between the two countries. Additionally, the social media posts used for truth discernment elicitation (see below) also differ, being taken from the two respective national contexts.

Participants are tasked with evaluating a randomized selection of social media posts containing both false and reliable news. Since these stimuli are real social media posts, two validation studies are conducted to validate them. These are called *"Validation-ITA"* and *"Validation-USA"*. Furthermore, these two studies also collect data on participants' motivations and personal normative beliefs regarding the importance of accuracy in online news sharing. This data is used to inform the treatment messages utilized in the three main studies. *Validation-ITA* involves a convenience sample of approximately 600 Italian online respondents recruited by PollStar, while *Validation-USA* is conducted on a similar convenience sample of around 600 US online respondents via Qualtrics.

Name of the study	**Type of sample**	**Country**	**Observations**
Validation-ITA	Convenience	Italy	617
Validation-USA	Convenience	United States	633
Convenience-ITA	Convenience	Italy	610
Main-ITA	Quota	Italy	1250
Main-USA	Quota	United States	1260

Table 4.1: Overview of the studies.

In this chapter, I will outline the methodological details of the three main studies (Section 4.1) and the two validation studies (Section 4.2). Following that, I will describe the process I employed to select and process the dataset of 80 real social media posts (Section 4.3), which participants are asked to evaluate. The chapter will then conclude by providing an overview of the analytical methods utilized to explore the data (Section 4.4) and the pre-registered hypotheses (Section 4.5).

4.1 Structure of the Survey Experiments

The main theoretical conclusions of this project are based on the data collected from three surveys. Each of these studies is developed to be as identical as possible to the others, to allow for comparing their results. In this section, I will detail how I measure all the explanatory and control variables, grouped by the mechanisms they belong to. Additionally, I will explain the measurement of the dependent variable. Unless stated otherwise, all measures are adopted from existing studies in the literature to avoid creating new and unnecessary measures, and to ensure consistency and comparability with prior research.

Moving on to the experimental component of these studies, I will present the treatment groups' structure and the treatment message's content. Finally, I will provide other methodological details of these three main studies. For the complete wording of all items, please refer to Appendix C.

4.1.1 Measuring the Explanatory Variables

Cognitive style is assessed using three Cognitive Reflection Tasks, which involve participants solving (primarily numerical) puzzles designed to gauge their ability to overcome impulsive yet incorrect responses. An example of such a task is as follows: *"The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?"*. The classic CRT items are well known and used in many studies so I adapted a recent variant to ensure that participants did not already know the answers. In particular, all CRT tasks in the present study are adapted from Arechar et al. (2023), although similar tasks can be found in related literature (e.g. Bronstein et al., 2019; Pennycook and Rand, 2019).

The surveys include a total of six questions on *politics*. Party affiliation is assessed using an item adapted from Enders and Uscinski (2021): *Which party do you align with?* [1]. Participants in the US version could select from seven options ranging from *Strong Democrat* to *Strong Republican*,

[1] I opted for this formulation because other papers in the fake news literature used more US-specific formulations (see for example Pennycook and Rand, 2019), while this version is closer to the one used in standard sociological and political surveys (like the EVS, 2020; Haerpfer et al., 2022).

while those in the Italian version could choose from the nine largest parties[2]. Additionally, both versions allow participants to select *"No party appeals to me"* and *"I prefer not to answer"*.

Political orientation is measured using the standard item: *"In political matters, people talk of "the left" and "the right". How would you place your views on this scale, generally speaking?"* (Haerpfer et al., 2022). Additionally, four scales used to gauge political ideologies from the European Values Study (EVS, 2020), which have been utilized in prior literature on fake news (Arechar et al., 2023), are included.

To measure *trust in institutions*, I utilise an item adapted from the European Values Study (EVS, 2020): *"Please indicate how much confidence you have in the following institutions"*. The list of institutions includes Parliament, the Government, the press and social media, mirroring the original survey. Additionally, trust in science is measured using two items from the World Values Survey (Haerpfer et al., 2022): *"How much do you agree or disagree with the statement that science and technology are making our lives healthier, easier, and more comfortable?"* and *"All things considered, would you say that the world is better off, or worse off, because of science and technology?"*. As these two questions measure trust in science as a process rather than trust in any specific scientific institution, I also decided to include *"universities"* in the list of institutions. This choice, although less common in the literature, aims to capture trust in academic institutions.

To gauge *anti-elitism*, I employ three items sourced from studies on populism (Wuttke et al., 2020). Participants are asked to express their level of agreement or disagreement with the following statements:

1. *"Politicians talk too much and take too little action"*.
2. *"The differences between the people and the so-called elite are greater than within the people"*.
3. *"Politicians care about what ordinary people think"*.

[2]Because of a mistake during the implementation of the Italian main survey, Forza Italia is not present in the list. However, it is present in *Convenience-ITA*, and some robustness checks indicate that its exclusion did not influence responses for other parties and the reliability of this question, which, in any case, is not central to the analysis.

To capture the perception of a *social norm* regarding the importance of sharing accurate news, I combine items from the literature on social norms and items from the literature on fake news. *Individual motivations* for sharing, akin to factual beliefs (Bicchieri et al., 2014) within this norm's context, are assessed by adapting a measure utilized in both Arechar et al. (2023) and Pennycook et al. (2021): *"When deciding whether to share a news on social media, how important is it to you that the content is..."*. Participants are asked to choose between five factors: 1. *Surprising*, 2. *Funny*, 3. *Accurate*, 4. *Interesting*, and 5. *Aligned with your politics*. In past studies, participants rated the importance of each factor using a Likert scale. Instead, participants are asked to rank these factors from most to least important. This adjustment aims to discourage participants from labelling multiple factors as very important and instead compel them to choose *between* them.

The survey then asks participants to guess how many other respondents selected "accuracy" as their first priority (*"In your opinion, how many other participants indicated accuracy of the content as their top priority when deciding whether to share a news on social media?"*). This item serves to gauge beliefs about the behaviours of others (i.e. *empirical expectations*, Bicchieri et al., 2014). Participants are incentivized with 0.3$/€ (as in Bicchieri et al., 2014) to ensure their focus and encourage their best guess. To aid participant comprehension of the question and incentive, they are presented with the following options: 1. *Almost no one*, 2. *Few participants (around a quarter)*, 3. *Half of them*, 4. *Most of them (around three quarters)*, and 5. *Almost all of them*. This allowed for a clear correct/wrong answer. Conversely, employing a 0-100% slider for selecting a precise percentile would have complicated the reward computation and its understanding by the respondents.

Personal normative beliefs, which pertain to beliefs about what should be done (Bicchieri et al., 2014), are assessed by rephrasing the previous question on individual motivation. Participants are asked to rank the same motivations but with a focus on what they believe should be done, rather than their own actions: *"When deciding whether to share a news on social media, how important*

should it be to people that the content is...".

The final question concerning social norms addresses participants' beliefs about what others believe (i.e. *normative expectations*, Bicchieri et al., 2014). Specifically, participants are asked to estimate how many other respondents selected "accuracy" as the top priority that should motivate people's news sharing on social media (*"In your opinion, how many other participants indicated that accuracy of the content should be the top priority when deciding whether to share a news on social media?"*). This item mirrors the one used for empirical expectations, including the incentivisation.

In addition to the variables related to the explored mechanisms, several control variables regarding demographics and media consumption are measured. Age is assessed using an item adapted from the World Values Survey (Haerpfer et al., 2022), re-coded to focus on age, instead of year of birth (as requested by Qualtrics). Gender is measured with a revised question from the Pew Research Center (Amaya, 2020) to account for non-binary individuals. Educational level and occupational group are measured with items from the European Values Study (EVS, 2020). Media consumption is measured using a question adapted from the World Values Survey (Haerpfer et al., 2022). Participants are asked to indicate the frequency with which they obtained information from various sources. The sources include social media, newspapers, the internet, TV news, conversations with friends or colleagues, email, radio news, and mobile phones. Responses are categorized as *daily*, *weekly*, *monthly*, *less than monthly*, or *never*.

The questionnaire also includes an attention check and a manipulation check. The attention check tests whether participants read the questions and are not answering randomly. Participants are asked to choose a pre-defined answer: *"Help us keep track of who is paying attention, please select - 'Somewhat disagree' in the options below."*. This question is placed just before the accuracy task. The manipulation check tests participants' recall of the treatment message by asking: *"Some minutes ago, we showed you a message containing data about Americans' priorities in news sharing. Can you recall what it said?"*. Participants can then choose among three possible answers. This

question is asked during the accuracy task, between round seven and round eight.

4.1.2 Measuring the Dependent Variable

The main dependent variable, *perceived accuracy*, is measured using a repeated accuracy task. Participants are presented with a screenshot of a social media news post and asked: *"To the best of your knowledge, how accurate is the above post?"*. Responses are collected on a 6-point Likert scale, ranging from *"Extremely inaccurate"* to *"Extremely accurate"*. Additionally, participants are provided with an *"I don't know"* option (DK option). This inclusion addresses a potential limitation of previous studies (for example, in Arechar et al., 2023 and in Pennycook and Rand, 2019). As noted in Chapter 2, the absence of such an option could artificially inflate the incidence of misinformation beliefs (Altay et al., 2023). Besides this difference (the presence of a DK option), the item was identical to the one used in many other papers in the literature (e.g. Arechar et al., 2023).

The *structure* of this accuracy task involves participants evaluating 10 social media news posts, randomly selected from a larger dataset of 40 posts per country. These datasets contain a balanced amount of reliable and fake news (for more on how stimuli are selected, see Section 4.3). The randomization process ensures that each participant 1. does not encounter the same post more than once, and 2. does not encounter both the reliable and the fake news posts referring to the same news event. The second constraint aims to prevent participants from gaining information about a news event from one post and using it to judge another post about the same event.

4.1.3 The Treatment and the Treatment Groups

During the above-mentioned accuracy task, the survey exposed participants to a treatment message that informed them about their peers' social expectations. As anticipated, this was done to test whether the eventual perception of social norms regarding the importance of accuracy in news sharing could be manipulated and with what results. Two distinct messages are utilized for this

purpose: one intended to strengthen the perception of the norm's presence, while the other aimed to weaken it.

Following the completion of the initial five rounds of the accuracy task (after round five and before round six), participants are randomly assigned to one of two treatment groups. Participants in the first treatment group saw a message highlighting the presence of the social norm, while participants in the second group saw a message highlighting the absence of it.

Consequently, the study adopts a mixed design, incorporating both between-subjects (comparing the two treated groups) and within-subjects (comparing the first five to the last five rounds) structures. In other words, the first five rounds of both groups serve as a control group, while the subsequent five rounds represent the "presence of a norm" treatment group and the "absence of a norm" treatment group, respectively. In this way, half of the observations (the first five answers of all participants) constitute the control group while the two treatment groups account for one-quarter of observations each (the last five answers of participants in the two respective treatment groups).

An alternative setting would have been utilizing three distinct groups in an exclusively between-subjects structure. This would have resulted in having just one-third of the observations in the control group. However, the main objectives of this project pertain to the mapping of mechanisms without interventions, also to allow comparison with other papers' results. Consequently, the decision was to prioritise the dimension of the control group over the presence of a strictly defined between-participants treatment structure.

After two other iterations of the accuracy task (after round seven and before round eight) the survey exposes participants to a manipulation check. In this check, participants are asked to remember the content of the treatment message. This allows for testing the understanding and recall of the treatment. Independently of their answers, participants are shown the correct answers. In other words, participants are exposed to the treatment message another time after the

manipulation check, as a re-treatment. Figure 4.1 presents a diagram summarising the treatment structure.

The two treatment messages are written to inform respondents about other participants' motivations to share news online (impacting respondents' empirical expectations) and their beliefs on what should motivate news sharing (impacting respondents' normative expectations). Both messages are prefaced by a statement introducing the context of the message:

"Before answering the following questions, we want to share the following information with you: We asked American respondents in an earlier survey session the same questions that you just answered about the importance of different factors when deciding whether to share a news article or not.".

The treatment message emphasizing the *presence* of an accurate sharing social norms is structured as follows:

"A large majority of these Americans stated that accuracy is their first priority and that it should be the top priority of people when deciding whether to share news on social media."

The treatment message highlighting the absence of a norm is identical, except for the number of people having the said social expectations:

"Only a small minority of these Americans stated that accuracy is their first priority and that it should be the top priority of people when deciding whether to share news on social media."

The message remains the same in the Italian version, with the only difference being that it refers

to the Italian population rather than the American one.

In order to avoid deception, all treatment messages are derived from real data collected during the validation studies (see Section 4.2). To create contrasting messages based on data from the same population, I divided the data from these surveys into hourly sessions. Within each hourly session, I calculated the proportion of participants who prioritized "accuracy" in their responses to questions about motivations and personal normative beliefs. This segmentation allowed for the creation of distinct subsets of participants with varying levels of emphasis on accuracy. Subsequently, I identified two sessions from each country, comprising at least 30 participants. One session, characterized by a higher proportion prioritizing accuracy, informed the "presence" message, while the other session, with a lower proportion prioritizing accuracy, informed the "absence" message.

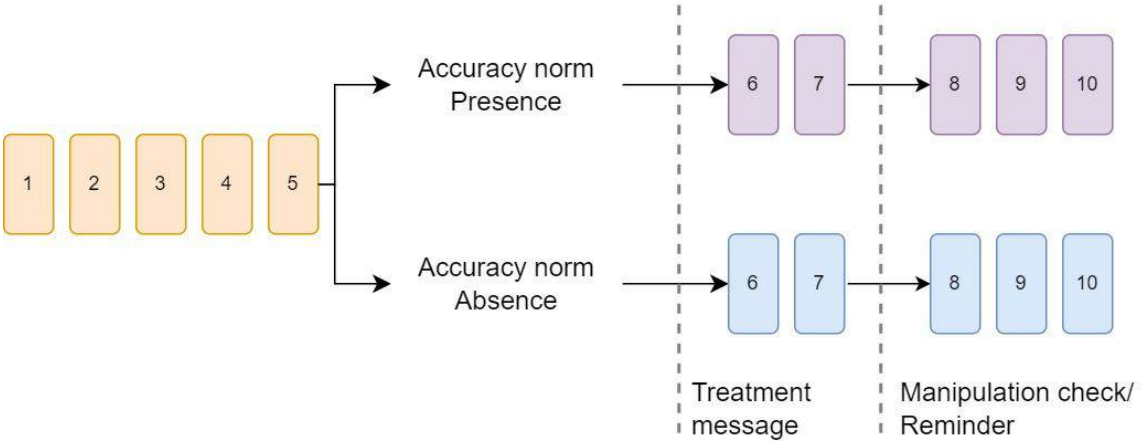


Figure 4.1: Structure of the treatment across the accuracy task. Each rectangle represents an iteration of the accuracy task.

4.1.4 Other details of the survey

Translation and Back-Translation

In cross-cultural studies, like the present one, a common concern is the development of non-equivalent versions of the same survey, which can inadvertently confound artificially created differences in survey comprehension with genuine cultural variations (Sperber et al., 1994).

Among others, one widely adopted method for addressing this concern is back-translation (Ep-

stein et al., 2015). This process involves translating the questionnaire from the original language to the target language by one translator and then translating it back to the original language by a second, independent translator. The back-translation is then compared to the original text. If the translation closely matches the original text semantically, without introducing new meanings, the translations are sufficiently accurate and maintain the intended content.

To ensure cross-cultural adaptation, the questionnaire used in *Convenience-ITA*, *Main-ITA* and *Main-USA* was initially developed in English and then translated and back-translated into Italian. Two independent Italian doctoral students, fluent in Italian and English conducted this procedure. Additionally, two other translators performed the same process for supplementary sections of the questionnaire, which were added later. Any discrepancies between the original and back-translated texts were resolved through collaboration with the respective translators.

Ethics

The questionnaire of *Convenience-ITA*, *Main-ITA* and *Main-USA* received approval from the ethical committee of Collegio Carlo Alberto in Turin (see Appendix D)[3].

At the conclusion of the questionnaire, participants are provided with a debriefing page. This page clarifies that the treatment messages are derived from data on a *subset* of participants from another survey. In addition, it presents information on the figures of the entire sample.

The debriefing page also informs participants that some social media posts they judged contain false information. The same page also includes a list of all the posts that the participant saw, reporting whether each post contained false information. Furthermore, each post is accompanied by an interactive link to sources where participants can access more information about the post and its veracity. For fake news posts, these links direct participants to articles from PolitiFact or

[3]The two validation studies did not undergo ethical approval. However, they present a structure and types of questions that are very similar to those of the main studies (for which approval was obtained). The two validation studies were less intrusive than the main ones, as they presented far fewer questions. Furthermore, the studies presented the same text for informed consent as the one used in the main studies. Validation studies were implemented on the same platforms and addressed participants selected in the same way as in the main studies. Additionally, the two validation studies also presented a debriefing page that warned participants about the reliability of the posts they judged and gave them links to additional information on their sources.

Facta (see Section 4.3) that assessed these posts as false.

Participants who wish to understand how their data are utilized can access a link at the end of the survey. This link directs them to a web page featuring the study's initial descriptive results. Additionally, the webpage will eventually host a paper version of the project once it is finalized.

Before reaching the debriefing page, participants have the opportunity to provide comments, suggestions and critiques regarding the survey through an open text box.

4.2 Structure of the Validation Studies

As outlined in earlier chapters, one of the main contributions of the present project is using real-life stimuli from different cultural contexts. Specifically, the main surveys (*Convenience-ITA*, *Main-ITA* and *Main-USA*) incorporate screenshots of real social media posts from Italy and the United States in their accuracy tasks. To ensure the comparability of these stimuli across both countries and to eliminate any potential outliers from the datasets, the project also include two validation studies (*Validation-ITA* and *Validation-USA*).

These studies follow identical structures and were administered to convenience samples from Italy and the United States. Their objectives are as follows:

1. Measuring the perceived *quality* of the social media posts across various dimensions (e.g. readability, relevance, etc.), to ensure the comparability of items.

2. Measuring the perceived *partisanship* of the social media posts to gauge the impact of motivated reasoning.

3. Collecting information about participants' motivations to share news online, which would be incorporated in the treatment messages of the main surveys (see Section 4.1).

The structure of *Validation-USA* is the following. Each participant is asked to judge 3 out of 40 randomly selected social media posts (see Section 4.3 on how the dataset was built). This collection

of 40 social media posts is the same then used in *Main-USA*.

For each screenshot, participants are asked to assess: 1. perceived quality, 2. perceived accuracy, and 3. perceived partisanship. The randomization process adheres to the same constraints used for *Convenience-ITA*, *Main-ITA* and *Main-USA*. These restrictions ensure that participants a. do not encounter the same post more than once, and b. do not assess fake and reliable posts relating to the same event.

In the validation studies, one of the aims is to ensure the comparability of social media posts across countries and to verify that they do not represent vastly different informational formats. The perceived quality of social media posts is assessed using an adapted item from Graefe et al. (2018), who developed a shortened battery of questions about news quality based on Sundar (1999). For each post, participants are shown the following text:

"How would you describe the post you just saw?" For each of the following adjectives, please indicate how well the word describes the post you just read (for example, "1" means that the post is not precise, while "10" means that the post is extremely precise). The post is...".

Participants were then asked to rate the post using a list of 12 adjectives (*Accurate, Trustworthy, Fair, Reliable, Entertaining, Interesting, Vivid, Well written, Coherent, Concise, Comprehensive,* and *Descriptive*) using a 10-point Likert scale.

One of the mechanisms re-tested by this work is motivated reasoning, i.e. the idea that individuals tend to believe information that aligns with their political views (see Section 2.2.2). To empirically test this mechanism, it is necessary to measure the perceived political leaning of news. In other words, it is necessary to measure what political ideas, if any, these news are confirming. With this goal in mind, participants are asked to assess the perceived partisanship of the news by adapting items from similar studies in the literature (e.g. see Pennycook and Rand, 2019). The

question is phrased as follows:

"Regardless of their veracity, do you think that the facts described in the post would be more favourable for Democrats or Republicans?"

This formulation differs from those used in other papers (*"Assuming the above headline is entirely accurate, how favourable would it be to Democrats versus Republicans"*, Pennycook and Rand, 2019) in two main ways. Firstly, participants are not pushed to believe that potential fake news is accurate and instead are asked to judge the partisanship of posts *independently* of their veracity. Secondly, the focus is placed on the facts (*"...do you think that the facts described..."*) described in the posts, clarifying that the question pertains to the content rather than the framing of the news.

To facilitate a comparison between perceived quality as measured in the validation studies and perceived *accuracy* as measured in the main studies (see Section 4.2), participants are asked the same question (*"To the best of your knowledge, how accurate is the above headline?"*) in both sets of studies.

After judging three randomly selected posts, participants are asked about their motivations to share news online (*"When deciding whether to share a news on social media, how important is it to you that the content is..."*) and what should motivate news sharing (*"When deciding whether to share a news on social media, how important should it be to people that the content is..."*), using the same items as in the main studies. This is done to inform the treatment messages of *Convenience-ITA*, *Main-ITA* and *Main-USA* (see previous subsection).

The survey concludes with an open text box for comments and suggestions and a debriefing page similar to the one used for the main surveys (see previous subsection), which informed participants about the veracity of the posts they saw and provided sources for gathering more information about

them.

The structure of *Validation-ITA* is identical, except that the items are randomly chosen from the Italian dataset of screenshots, instead of the US one. In addition, perceived partisanship of the posts is measured by using "left" and "right" instead of Republicans and Democrats.

4.3 Selection of the Social Media News Posts

The methodological decision to utilize authentic social media posts as the primary stimuli to be submitted to participants is one of the main contributions of the present project, as it allows the measurement of truth discernment within a realistic information environment, thereby enhancing the external validity of the gathered evidence. However, a consequence of this choice is that the conclusions of this study could vary considerably based on the chosen posts. To address this concern, I will outline the rigorous and replicable procedure to select fake and reliable social media news posts.

I tried to fit together different aims in creating and implementing this procedure. The most important criterion is avoiding arbitrariness as much as possible to make this procedure replicable and not based on any choice on the researcher's side. A second aim is to have a comparable set of fake and reliable posts. This involves avoiding scenarios where fake news in one format (e.g. social media posts) is compared with a reliable piece of information in a structurally different format (e.g. news articles from mainstream sources). A final aim is to isolate the impact of features of the posts that are relevant in real life but not included in the present analysis and, if included, could confound the impact of every mechanism studied.

4.3.1 Choosing the Fake News Posts

The process of stimuli selection begins for each event by identifying social media posts containing false information. This is accomplished by consulting debunking sites dedicated to labelling disputed pieces of information with varying levels of veracity, from completely false to completely true. This

approach is commonly employed in similar studies (e.g. Pennycook and Rand, 2019). To determine the specific debunking site for each country, I referred to the list of fact-checking sources affiliated with the International Fact-Checking Network.

In the United States, I opted for PolitiFact, given that 1. its archive contains a diverse collection of articles debunking fake news in various formats (news articles, Facebook posts, Instagram posts, Twitter posts, and so on) 2. it is easily browsable through the use of categories and hashtags, and 3. it has already been used by other papers in the literature (e.g. Allcott and Gentzkow, 2017 or Arechar et al., 2023). In Italy, I chose Facta for the following reasons: 1. along with Open, it is the only debunking site analysing misinformation in the format analysed by this work, while other sources focus on other types (like politicians' declarations, in the case of Lavoce.info and Pagella Politica); 2. it employs a consistent and navigable categorization system for posts.

In both cases, I selected fake news posts by reviewing the most recent articles and applying a series of exclusion criteria. Posts are excluded under the following circumstances:

1. The post does not match the one referenced in the debunking articles. While identical posts (e.g., from different users with the exact same text) are accepted, *similar* posts are not considered.

2. The post consists solely of a video unless the content is described in the caption or text overlay.

3. The debunking article link leads to a journalistic article, blog post, or political statement rather than a social media post.

4. For the Italian dataset, the article does not fall under the "Notizia falsa" tag (e.g. it fell under "fuori contesto", out-of-context). Within the "notizia falsa" tag, only articles with the red "notizia falsa" label in the image are included.

5. The linked post comprises an extensive wall of text without any accompanying image unless it could be presented in the Facebook default "cut" format (e.g. "...See more") without omitting

crucial information.

4.3.2 Choosing the Reliable News Posts

After gathering fake news posts according to the outlined procedure, I selected corresponding reliable news posts using the following steps:

1. Search news on the same topic on Google News, by using keywords taken from the fake news item, usually from its caption or headline. In the Italian case, restrict the results from Google News to Italian pages if necessary.

2. Adjust the keywords until finding at least one news article that meets the exclusion criteria (see below) on the result page of Google News.

3. If multiple suitable news articles are present, randomize a number from one to ten (e.g., using Google's random number generator) and choose the corresponding article.

4. Open the selected news and review its content. If it adheres to the exclusion criteria, annotate the link.

5. Look for a social media post referencing the chosen news event, preferably on the same platform as the corresponding fake news. This can be done by searching the news title's relevant keywords. Notes:

   (a) The chosen social media posts must directly mention the chosen reliable news with a link or in the caption, without adding any other fact or misleading interpretation.

   (b) For Instagram, where text searches are not possible, employ hashtags on the Instagram website or use Google with the syntax

   ```
   {"site: Instagram intext: [keywords]"}
   ```

6. If multiple suitable posts are found, randomize a number from 1 to 10 and choose the corresponding article.

These are the exclusion criteria used to select reliable news articles from Google News (Step 2). The randomly selected post was excluded when it included:

1. Editorials.
2. Debunking articles.
3. Rankings (for example "Top 50 albums of 2022").
4. Tutorials and summaries (for example "Everything you need to know on the 2022 elections").
5. Reports from public institutions.
6. Multiple news events in one article (if unavoidable, fact-check each event separately).

4.3.3 Processing and Final Datasets

The posts chosen following this selection procedure are then captured via screenshots and processed to eliminate influencing and unwanted features. In particular, the selected posts are processed to remove comments, likes (only when numbered) and number of visualization (a feature that Twitter, now named X, introduced during the data collection), via visual blurring.

The initial dataset resulting from this procedure consisted of 80 posts from Italy and 80 posts in the United States, spanning the period from September 2022 to January 2023. These posts were sourced from platforms such as Facebook, Twitter (now X) and Instagram. To create the final selection of 40 items for each country, efforts were made to cover various topics. The only restriction imposed during this selection process was to include at least three events (equivalent to 6 posts) related to the Russian invasion of Ukraine per country [4]. For the complete list of selected posts, refer to the Appendix B.

[4]This was done to allow for a possible focus on this topic, which dominated the public debates on both countries during the studied time frame. However, this analysis is not included in the present work.

4.4 Analytical Methods

4.4.1 Analytical Methods of Validation Studies

The data analysis from the validation studies mainly relies on descriptive analysis techniques. In particular, the core of the analysis relies on one-sample and two-sample t-tests. Observations are grouped in various ways, and their scores are compared using averages and confidence intervals at a 95% confidence level.

The presence of outliers within each country is assessed by computing interquartile ranges (IRQ). For example, it is important to test that no post is perceived as unusually interesting or uninteresting. Firstly, I computed IQR by subtracting the third quartile's score from the first quartile's score. I then computed the threshold for mild outliers. The upper boundary was calculated by summing the third quartile's score with the IQR multiplied by 1.5. The lower boundary was calculated by subtracting the IQR multiplied by 1.5 from the first quartile's score. Extreme outliers' boundaries were calculated similarly but using a coefficient of 3 instead of 1.5. This procedure is summarised by the following formulas:

- $IQR = Q_{75} - Q_{25}$
- $Mild_{lower} = Q_{25} - 1.5 \times IQR$
- $Mild_{upper} = Q_{75} + 1.5 \times IQR$
- $Extreme_{lower} = Q_{25} - 3 \times IQR$
- $Extreme_{upper} = Q_{75} + 3 \times IQR$

The score of a post for the adjective *"interesting"* was then compared with those boundaries to assess whether it was an outlier. The same procedure was repeated for all adjectives and the two validation studies.

The assessment of cross-country differences in the perceived quality of posts was based on repeated one-sample t-tests. First, for each adjective, the average score of all Italian posts was compared to the average score of all US posts. However, similar averages could also be caused by different distributions. For example, it could mean that all posts have the same score in one context, while in the other, reliable news has higher scores and fake news has lower scores. The same analysis was repeated to test for this eventuality by dividing fake news posts from reliable posts. In conclusion, I used a visual analysis through histograms by countries and veracity to check whether similar averages were caused by different distributions.

4.4.2 Analytical Methods of Main Studies

Answers from *Convenience-ITA*, *Main-ITA* and *Main-USA* are analysed using the same procedures and models. To anticipate, hypotheses are tested using Ordinary Least Square (OLS) and multi-level regression models. These models include different combinations of explanatory and control variables. Additional analysis includes descriptive statistics, factor analysis and various robustness checks.

For each of the three studies, the unit of analysis comprises answers to the accuracy tasks. As previously mentioned, each participant judged ten randomly selected posts. This resulted in approximately 12.000 observations for *Main-ITA* and *Main-USA* each.

These observations are not independent. Instead, they are clustered by participants and posts. This means that answers given by the same participants are likely to be correlated. In addition, answers given by different participants to the same posts are likely to be correlated. Consequently, the analytical methods employed must address this lack of independence.

To account for this absence of independence, the core models for *Convenience-ITA*, *Main-ITA* and *Main-USA* entail a series of OLS linear regressions with robust standard errors clustered on participants and fixed effects for posts. The use of clustered standard errors and fixed effects allows us to account for the cross-classified structure of the data. These models were included in the

pre-registration, which can be found OSF at https://osf.io/hvdjz and at Appendix E.

These OLS models include the following list of variables. The dependent variable is the perceived accuracy of social media posts. The list of independent variables includes a dummy variable for item veracity (0 = false, 1 = reliable), separate coefficients for each explanatory variable, and interaction coefficients between the veracity dummy and each explanatory variable. Additionally, models include socio-demographic variables and media consumption as controls.

The main dependent variable of these models is the perceived accuracy of social media posts, assessed with a 6-points Likert scale. However, the primary focus of the analysis is not the *"perceived accuracy"* of posts per se but rather *"truth discernment"* (as defined in various papers, like Pennycook and Rand, 2020 or Arechar et al., 2023). This variable is computed as the difference between the perceived accuracy of reliable posts minus the perceived accuracy of fake posts. This difference in means is calculated on a different basis depending on the specific type of analysis. Using truth discernment allows for examining misinformation beliefs as the ability to distinguish false from reliable information. Conversely, analysing perceived accuracy alone would only enable the study of the propensity to believe information *in general*.

In regression models, the interaction between veracity and the explanatory variables allows measuring those variables' effect on truth discernment (Arechar et al., 2023). Consequently, these interactions are the primary focus of these studies' regression analysis and hypothesis testing.

For example, consider a regression model that includes veracity, CRT scores (measuring cognitive styles), and their interaction. The coefficient for veracity describes the effect of veracity of posts on their perceived accuracy, for participants with an intuitive cognitive style. The coefficient for CRT describes the effect of cognitive styles on the perceived accuracy of posts, for posts with a low veracity (for fake news). The interaction coefficient describes the effect of cognitive styles on the difference between the perceived accuracy of reliable posts and that of fake news posts. In other words, it describes how the truth discernment of participants with an analytical cognitive

style differs from that of participants with an intuitive cognitive style. Consequently, it measures cognitive styles' effect on truth discernment. This analytical approach mirrors similar studies in the literature (Arechar et al., 2023).

Several variations of each model were computed, including models with or without fixed effects at the post level. Additionally, when introducing controls, I computed versions with interactions between controls and veracity and versions without such interactions. The analysis also tests various mixes of variables. Initially, hypotheses are tested using a "reduced" model containing only the primary explanatory variables of each mechanism. Subsequently, the analysis is reiterated by substituting *a priori*-built indexes with those resulting from factor analysis. Additionally, another formulation included secondary variables linked to the pre-registered hypotheses, such as trust in science.

To summarise, the main model specification used for hypothesis testing consists of OLS regression models whose main focus is the interaction effect between posts' veracity and respondents' characteristics. This is a rather complex model specification, requiring numerous interaction effects, other than clustered standard errors and fixed effects. However, this is the model specification used in most recent papers investigating the same research question (see for example Arechar et al., 2023). I thus decided to replicate this model choice to allow comparability of the results with the literature. Additionally, I computed models with a more straightforward specification. Instead of using interaction effects to measure truth discernment, I created a new dependent variable incorporating the ability to discern news as the difference between participants' judgments and posts' veracity. I called this new specification *distance from the truth* (see below for more details).

The analysis also includes a series of robustness checks to test the consistency of regression models' results. These checks focuses on the operationalisation of variables and the specification of statistical models.

As previously mentioned, a primary robustness check involves aggregating variables into in-

dexes. In particular, the aim was to test whether the results changed when the two *a-priori* built unweighted indexes for anti-elitism and institutional trust were replaced with indexes derived from factor analysis. The *a-priori* built index for institutional trust includes all five variables of trust in different institutions (Parliament, the Government, the press, social media and universities). The *a-priori* built index for anti-elitism includes the three variables measuring anti-elitism.

The indexes derived from factor analysis are built using the following procedure. In the three main studies, I initially conducted Principal Component Analysis (PCA) and Principal Factor Analysis (PFA) on all numerical variables resulting from Likert scales. This is done to determine *how many* indexes to build and *which* variables to aggregate in them. Subsequently, the resulting indexes are computed using both weighted and unweighted averages of the re-scaled scores of the aggregated variables. The correlation between these indexes' weighted and unweighted versions highlights their substantial overlap.

Consequently, I computed an alternative version of the pre-registered OLS models, where the two unweighted *a priori* indexes are replaced by the unweighted index resulting from the factor analysis. This index comprises three items for institutional trust (trust in Parliament, Government and the Press) and the reverse-coded version of one item for anti-elitism (*"Politicians care about what ordinary people think"*).

A second robustness check focuses on operationalising expectations about the sharing social norm. In the original approach, these items are treated as standard categorical variables, with a level for each possible answer (*"Almost no one", "Around a quarter", and so on*). An alternative version of these variables is computed by transforming them into two numerical variables ranging from 0 (*"Almost no one"*) to 1 (*"Almost everybody").* Subsequently, the OLS pre-registered models are recomputed by replacing the original categorical variables for social expectations with these numerical versions. This allows for assessing whether the results remain consistent across different operationalizations of social expectations.

Another robustness check involves the operationalization of the dependent variable. In particular, I computed a new variable termed *"distance from the truth"*. Instead of simply measuring the score given by participants to each post, this variable measures how far the participant was from the "correct answer", giving the minimum score (0) to fake news and a maximum score (5) to reliable news. For instance, if a participant rates a piece of fake news with a score of 3, the distance from the truth is 3 (difference between 0 and 3). If the same score is given to a piece of reliable news, the distance is 2 (the difference between 5 and 3).

After computing this variable, I re-ran the OLS pre-registered models, using distance from the truth as the dependent variable. This formulation allows for a direct focus on the coefficients of the explanatory variables, as it already encapsulates the difference between fake and reliable posts in its definition, rather than concentrating on the interaction between veracity and the explanatory variables (as in the case of truth discernment).

A final robustness check involves computing the pre-registered models and their alternative versions using multilevel regressions (see Supplementary Materials A for more). This approach allows for incorporating nested data structures, such as clustering responses within participants and posts. In particular, for each model analysing the bivariate relation between an explanatory variable and the dependent variable, I computed three additional versions: 1. a model including random intercept for participants, 2. a model including random intercepts for posts, and 3. a model including random intercepts for both participants and posts. The model incorporating all mechanisms simultaneously is computed only in the third version, including random intercepts for participants and posts. It is worth noting that versions utilizing random slopes for participants often result in singularity or failures to converge and are thus excluded. These multilevel regression models are also included in the pre-registration.

Evidence resulting from regression models will be presented via coefficient plots and regression tables. Additional regression tables are detailed in Appendix A. In these coefficient plots, each

dot represents the regression coefficient of the interaction between an explanatory variable and the veracity of posts. As anticipated, this interaction measures the variable's impact on the scores given to reliable posts compared to those given to fake news (i.e., truth discernment, a measure of participants' ability to discern news). As such, it represents the core of the correlational analysis. Each dot is depicted with the 95% confidence interval of its estimate, coloured black if significantly different from zero and grey if not.

The data analysis also involves the use of additional methodologies. Before computing the pre-registered regression models, each variable undergoes descriptive analysis through univariate analysis. Similarly, the interaction between these variables and misinformation believing is examined through bivariate descriptive analysis, initially focusing on perceived accuracy and then on truth discernment.

In addition, the analysis also includes the above-mentioned factor analysis. Variables composing each mechanism are analysed individually and grouped in statistical indexes. In the main formulation, only anti-elitism and institutional trust variables are aggregated into indexes based on theoretical considerations. Specifically, the three items for anti-elitism and the five for institutional trust are grouped into two distinct indexes. Firstly, I re-scaled each variable to have a range from 0 to 1. Secondly, I constructed the two indexes by computing an unweighted average of the variables' scores. I then conducted factor analysis using both PCA and PFA on these re-scaled variables to explore the presence of unexpected clusters of variables. The resulting indexes (calculated through both weighted and unweighted averages) are discussed in the main studies' results and are used in robustness checks (see Appendix A).

4.5 Hypotheses

The Research Design chapter concludes by summarising the hypotheses tested in this project. These hypotheses address the research gaps mentioned in Chapter 2 through the methodology

outlined in this chapter. All hypotheses are informed by the theory outlined in Chapter 3. As already mentioned, all research questions and hypotheses have been pre-registered on OSF at https://osf.io/hvdjz.

Analytical Thinking

The literature already gathered evidence on the influence of cognitive styles in determining individuals' truth discernment. In particular, individuals with a more analytical type of thinking are usually better at judging news accuracy. I retest this finding with a new collection of real and validated social media posts. Additionally, I prove this mechanism against other relevant mechanisms. The resulting hypothesis is formulated as follows:

Hypothesis 1 (confirmatory): Participants with high scores in the CRT tasks will be more accurate in judging news.

Motivated Reasoning

Past papers explained misinformation beliefs with motivated reasoning. According to this idea, individuals are worse at judging information when it confirms their prior political beliefs. I retest this theory by asking participants to judge a selection of either neutral, politically congruent or incongruent social media posts. Consequently, I expect that:

Hypothesis 2 (confirmatory): Participants will be less accurate in judging news with a political leaning which is concordant with their political affiliation, than when judging politically discordant news.

Anti-elitism and Institutional (dis)trust

I argue that individuals might also be influenced by their relationship with institutions strictly linked to the public discourse, such as politicians, media and the scientific community. This is a rather underexplored facet of truth discernment. In particular, I expect individuals with bad consideration of institutions to be more prone to believe in fake news. I propose three explanations that lead me to expect a correlation between anti-elitism and truth discernment: motivated reasoning, source selection and reverse causality. The same applies, with some considerations, to institutional distrust. Consequently, I expect that:

Hypothesis 3a: Participants with high anti-elitism will be less accurate in judging news.

Hypothesis 3b: Participants with low institutional trust will be less accurate in judging news.

Social Norms about the Importance of Accuracy

Another understudied aspect of misinformation perception is the potential role of social norms. Specifically, I argue that individuals might be influenced by the importance they give to the accuracy of circulating information. I expect participants who think that the accuracy of shared news is important (factual belief) and that this should be the case (personal normative belief) to be better at judging news. In other words, I anticipate that:

Hypothesis 4a: Participants with a high factual belief about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Hypothesis 4b: Participants with a high personal normative belief about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

However, the core of the proposed social norm explanation lies in the social expectations linked to it. I argue that participants might be influenced by how much they perceive that others are interested in the accuracy of circulating information. Specifically, I expect participants who think that others are prioritizing accuracy in deciding what to share (empirical expectation) to be better at judging news. Additionally, I expect participants who believe that others also value this as the appropriate thing to do (normative expectation) to be better in their assessments. To summarise, I hypothesise that:

Hypothesis 5a: Participants with a high empirical expectation about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Hypothesis 5b: Participants with a high normative expectation about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

These two latter hypotheses will also be tested experimentally. If the social norm mechanism influences participants' truth discernment, a message highlighting others' compliance with such norms should have a positive impact on participants' ability to judge news. On the contrary, a message highlighting how others are not interested in the accuracy of circulating information should make participants feel allowed to put less effort into their evaluations. Consequently, I expect that:

Hypothesis 6a: A priming message highlighting the presence of an accurate sharing social norm is expected to increase participants' truth discernment.

Hypothesis 6b: A priming message highlighting the absence of an accurate sharing social norm is expected to decrease participants' truth discernment.

Replication Outside Anglo-Saxon Countries

The current literature implicitly assumes the generalizability of its explanations outside the explored contexts, given that they focus on individual mechanisms. However, this assumption is rarely tested. Instead, I will replicate the same studies in the US (the most studied context) and Italy, to allow the test of these conclusions outside the Anglo-Saxon context.

Hypothesis 7: The mechanisms driving fake news beliefs will be similar across Italy and the United States.

Chapter 5

Results of Validation Studies

As anticipated, the primary objective of the two validation studies is to validate the stimuli assessed by participants in the accuracy tasks of the three main studies. Specifically, it is crucial to determine that *within* each country-specific dataset, there are no outliers. It is important to assess that no social media posts have unusually high or low scores across the 12 measured dimensions of perceived quality. For example, the presence of a post which is *unusually* interesting, (or entertaining, reliable and so on) could bias the analysis.

Additionally, it is important to assess to what extent fake news is qualitatively distinguishable from reliable news. This assessment is important to understand the *difficulty* of the accuracy task. Are people asked to distinguish very distinct social media posts? Or are they evaluating reliable news which is qualitatively identical to fake news? This estimation is also valuable for the replicability of the studies. Suppose that another study collects social media posts to be used in accuracy tasks. Without an estimation of how fake news is distinguishable from reliable news, we would not know if participants are asked to perform a similar task or a more difficult (or easier) one, compared to those in the present work.

The third aim of the two validation studies is to evaluate the comparability of the collected real social media posts across the two countries. Are Italian fake news similar to those originating in the USA? Or are we comparing pieces of information which are *structurally* dissimilar? The

same inquiry applies to reliable news sources. Determining the nature of potential cross-country differences is contingent upon addressing these questions. These differences could be caused by the fact that participants are asked to judge different posts rather than by actual contextual differences in how they evaluate information.

The validation studies are also used to gather relevant information for the three main studies. In particular, the validation survey asks participants to evaluate the perceived partisanship of the various social media posts. This allows the testing of motivated reasoning in the main surveys. A participant in the main surveys is considered to be evaluating a politically congruent social media post when that post was evaluated as favourable to their political side by participants in the validation study.

Ideally, we would expect to find a balanced set of neutral, left- and right-leaning fake and reliable posts in the two countries. However, perceived partisanship in the validation studies is not a crucial exclusion criterion for posts. For example, even if the resulting sample included a large majority of posts leaning toward one side, I would not exclude posts to make the selection balanced. The rationale of this decision is to not force a situation within the survey experiments, without it being supported by the empirically implemented collection of data. The stimuli selection procedure is designed to have a collection of items which is as probabilistic and least arbitrary as possible. Excluding items based on their perceived partisanship distribution would partially invalidate these principles. Furthermore, it would mean prioritizing the test of one mechanism (motivated reasoning) on an artificially selected sample of posts, rather than the test of all other mechanisms on a realistic one.

Finally, the validation studies are used to inform the social norm treatment messages (see Chapter 4). Using fictitious statements about other participants' factual and personal normative beliefs would mean giving them potentially false information, thus deceiving them. Instead, questions about participants' factual and personal normative beliefs in the validation studies are used to

inform the content of the treatment messages. To allow opposing messages for the two treatment groups (about the presence or absence of a social norm), I subset observations in sessions of one hour each and then choose data from different sessions for the two messages.

5.1 Perceived Quality of the Social Media Posts

5.1.1 Presence of Outliers

The presence of outliers was evaluated by calculating boundaries of scores using inter-quartile ranges (see Section 4.4). Figure 5.1 shows the results for posts in the Italian dataset, using data from *Validation-ITA*. Figure 5.2 replicates the same analysis for the USA context, using data from *Validation-USA*. In summary, these graphs illustrate the average scores of perceived quality of the 80 social media posts, compared to the two sets of boundaries. The first set is for mild outliers and relates to a broad definition of outliers. The second set is for extreme outliers, which relates to a more stringent definition. Social media posts with average scores outside the two sets of boundaries are considered respectively mild or extreme outliers on those adjectives.

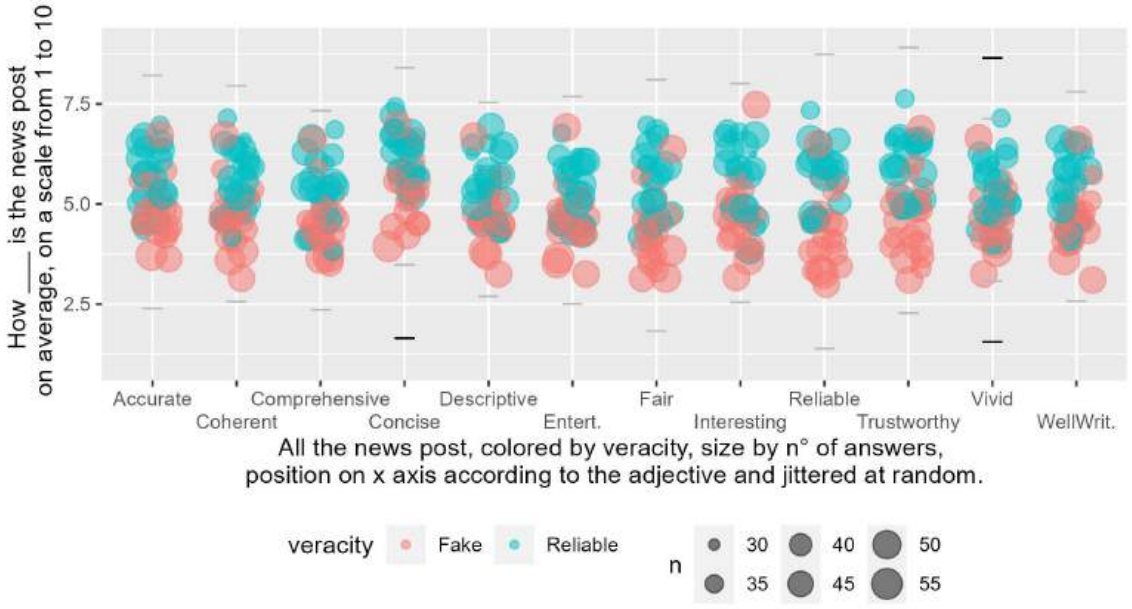


Figure 5.1: Perceived quality of social media posts in the Italian sample (*Validation-ITA*).

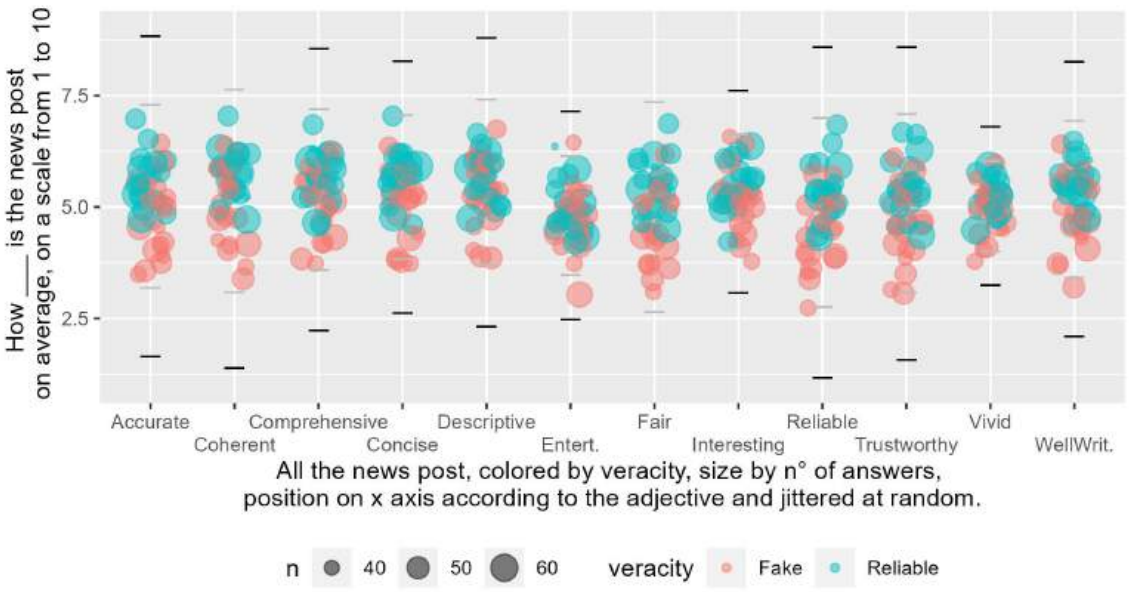


Figure 5.2: Perceived quality of social media posts in the US sample (*Validation-USA*).

Figure 5.1, referring to data from *Validation-ITA*, should be interpreted as follows. On the horizontal axis, scores of posts are grouped according to the adjective they refer to. For example, the first cloud of dots refers to scores for the adjective *accurate*. For each adjective, each dot represents a social media post collected for Italy, coloured in red if it is a fake news and blue if it is a reliable post. The position on the y-axis indicates the average score participants gave to that post, in that adjective, on a scale from one to ten. Additionally, the size of each dot corresponds to the number of answers collected for a certain post. In fact, each participant saw a random sample of posts. As a consequence, each stimulus received a slightly different number of evaluations. Finally, the small horizontal lines at the top and bottom of each adjective's scale depict the boundaries beyond which a post is defined as an outlier. The grey dashes represent boundaries for mild outliers, while the black dashes represent those for extreme outliers.

Validation-ITA does not present any extreme outlier. All 40 social media posts have an average score between the boundaries for extreme outliers. Regarding mild outliers, there is only one post

crossing the corresponding boundaries. The average vividness score (7.142 points) of the post *8-Reliable*[1] is slightly above the boundary for mild outliers of that adjective (7.13 points). The difference between the two is around 0.012 points (on a scale from 1 to 10).

Figure 5.2, referring to *Validation-USA*, follows the same visual structure. *Validation-USA* does not present any extreme outlier. Again, all 40 social media posts have an average score between the boundaries for extreme outliers. However, the number of mild outliers is higher in this study, compared to *Validation-ITA*. In particular, ten posts cross these boundaries (see Table 5.1 for a summary). The two posts that cross upper boundaries are *10-Reliable* and *14-Fake*, which are relatively more entertaining than the other posts (with differences respectively of 0.219 and 0.308 points on a 1-10 scale). The other eight posts have scores relatively lower than the rest in some dimensions. In particular, post *5-Fake* is below the threshold on interest (difference of -0.262 points), vividness (-0.222 p.), conciseness (-0.096 p.) and reliability (-0.021 p.). Post *4-Fake* is relatively less well written (-0216 p.) and trustworthy (-0.012 p.) than other posts. To conclude, *9-Fake* is relatively less entertaining (-0.436) and *11-Fake* is less concise (-0.057) than the rest.

Despite the presence of some mild outliers on single dimensions, all posts were kept in the dataset of stimuli to be used for the three main studies. In *Validation-ITA*, there is only one mild outlier, and its distance from the boundary is minimal. In *Validation-USA* there are six posts which are outliers in at least one of the 12 quality dimensions. However, even in these instances, they are mild, not extreme, outliers and remain within the boundaries of all other dimensions. Furthermore, in some cases, the crossing of thresholds seems to be caused more by particularly concentrated distributions of average post scores, rather than by isolated cases. For example, the range of average scores for vividness in *Validation-USA* is particularly concentrated. As a consequence, *extreme* boundaries for vividness are stricter than *mild* boundaries for other dimensions (e.g. reliability).

[1]Refer to Appendix B for the complete list of stimuli. Each stimulus is identified by a numerical label indicating the event/topic they mention and a label indicating whether they are fake or reliable.

Event	Veracity	Adjective	Mean	Low Bound.	Upper Bound.	Difference
9	Fake	Entert.	3.036	3.472	6.142	-0.436
5	Fake	Interesting	3.78	4.043	6.637	-0.262
5	Fake	Vivid	3.78	4.002	6.038	-0.222
4	Fake	Well Writ.	3.204	3.42	6.94	-0.216
5	Fake	Concise	3.732	3.828	7.058	-0.096
11	Fake	Concise	3.771	3.828	7.058	-0.057
5	Fake	Reliable	2.732	2.753	6.997	-0.021
4	Fake	Trustworthy	3.061	3.073	7.087	-0.012
10	Reliable	Entert.	6.361	3.472	6.142	0.219
14	Fake	Entert.	6.45	3.472	6.142	0.308

Table 5.1: Summary of mild outliers in *Validation-USA*.

5.1.2 Qualitative Differences Between Fake News and Reliable Posts

To address whether reliable news is distinguishable from fake news, I conducted repeated unpaired two-sample t-tests on their average differences in scores across the 12 quality dimensions. The results of this analysis are reported in Figure 5.3 and Figure 5.4. These two figures have the same graphical interpretation. On the x-axis, they present the 12 quality dimensions. On the y-axis, they present the difference between scores given to reliable posts and fake news on each dimension, as estimated by the t-tests. In addition to the point estimate, each dot is presented with its confidence interval at the 95% level. The dashed horizontal line represents a difference of zero. In other words, if the confidence interval of an estimated difference crosses this line, fake and reliable posts are statistically indistinguishable on that dimension. In addition, the plots show the number of observations on which each estimate is calculated with numbers in rectangles in the lower part of the graph.

Results indicate that reliable posts are perceived as "better" than fake news across all adjectives and in both countries. The p-value of t-tests on the difference between reliable and fake news scores is always below the critical threshold of 0.05. The adjective on which posts are less distinguishable is *entertaining*, for which the p-value is 0.047. However, this is true only in *Validation-USA*. In *Validation-ITA*, the p-value for the same adjective is far from the threshold (3.42×10^{-11}).

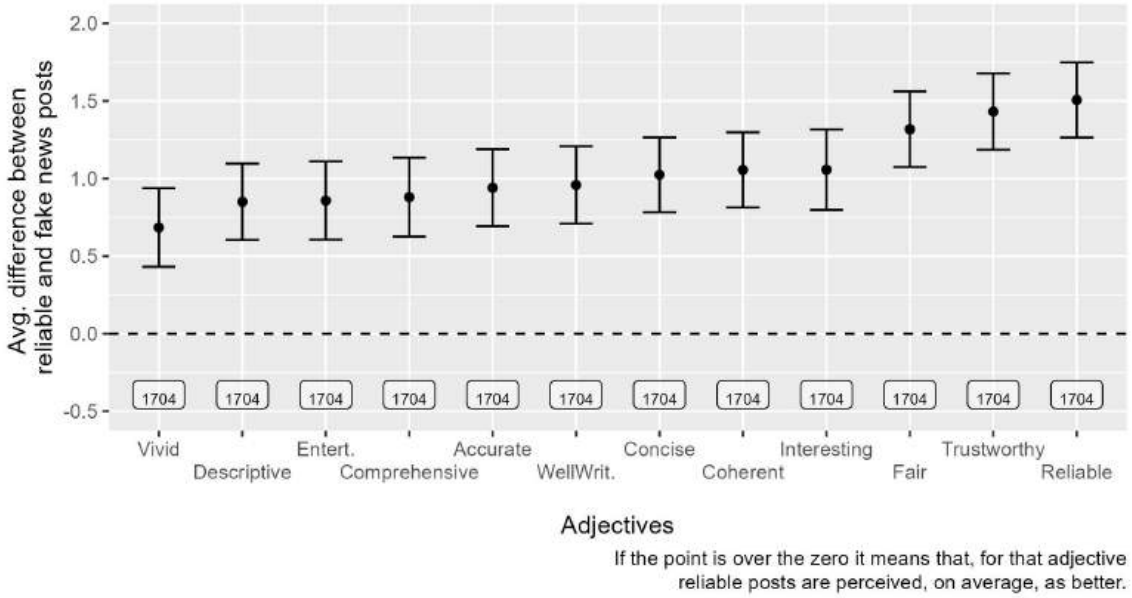


Figure 5.3: Difference between the perceived quality of reliable and fake social media posts in the Italian sample (*Validation-ITA*) with 95% confidence intervals. Difference calculated as *scores of reliable news - scores of fake news.*

Generally speaking, reliable posts are rated higher by one point on a scale from one to ten. The estimation of this difference goes from a minimum of 0.27 points in the case of entertainment in *Validation-USA* to a maximum of 1.5 points in the case of reliability in *Validation-ITA*.

Notably, the discrepancy between reliable and fake news is typically more pronounced for dimensions crucial to journalistic quality, such as reliability, fairness (difference of 1.32 points for *Validation-ITA* and 1.02 points for *Validation-USA*), and trustworthiness (diff. of 1.43 p. for *Val.-ITA* and 0.88 p. for *Val.-USA*). For example, reliability is the adjective with the widest the gap between reliable posts and fake news both in *Val.-ITA* (1.5 p.) and in *Val.-USA* (1.02 p.). Conversely, the difference is much smaller for dimensions less related to journalistic quality. For example, as anticipated, fake news is nearly as entertaining as reliable news in *Val.-USA* (difference of 0.27 p.). In *Val.-ITA*, the adjective with the narrowest difference is *vivid* (0.68 p.).

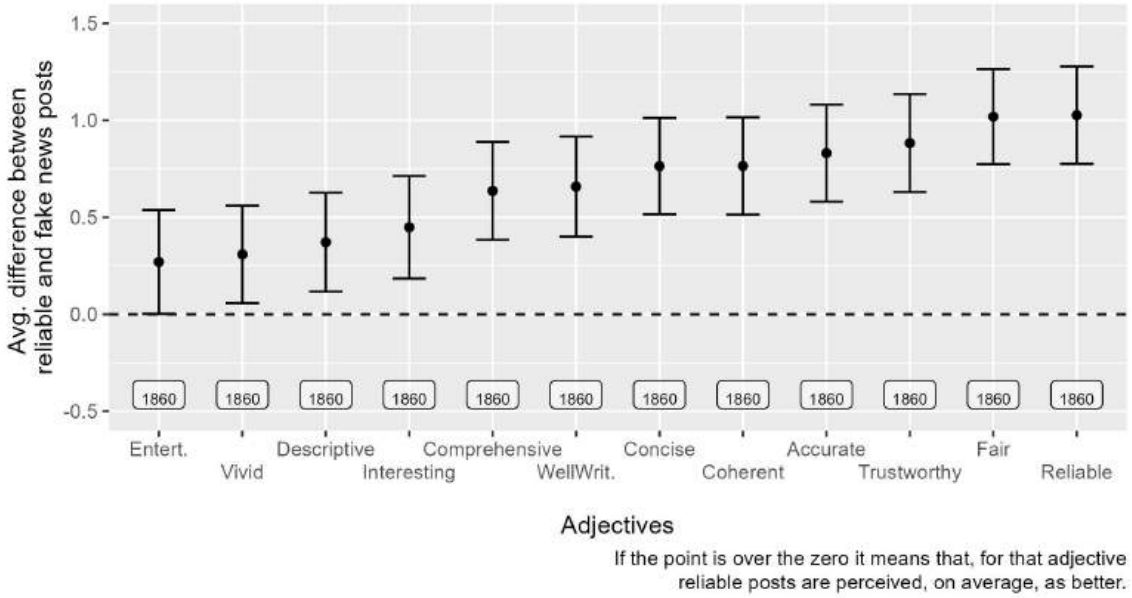


Figure 5.4: Difference between the perceived quality of reliable and fake social media posts in the US sample (*Validation-USA*) with 95% confidence intervals. Difference calculated as *scores of reliable news - scores of fake news.*

5.1.3 Cross-Country Comparability

Moving on to the comparability *between* Italy and the US, Figure 5.5 and Figure 5.6 compares the average scores given to posts in the two countries, across the 12 dimensions. Each subplot is dedicated to one adjective. For example, the first subplot at the top left shows how the average score for accuracy changes in the two countries. In each subplot, the average score given to posts in *Validation-ITA* is compared to that in *Validation-USA*. Each dot represents the point estimate resulting from a one-sample t-test, coupled with its relative 95% confidence interval. Figure 5.5 is dedicated to the comparison of reliable posts, while Figure 5.6 is dedicated to fake news posts.

Starting from analysing reliable posts, Figure 5.5 shows how, in general, reliable posts are perceived very similarly across the two countries. The confidence intervals of the two national estimates overlap in eight out of 12 adjectives: *accurate*, *coherent*, *comprehensive*, *descriptive*, *fair*, *interesting*, *vivid* and *well-written*. Statistically speaking, reliable posts in the Italian dataset are as

Country	Adjective	Average	Low Bound.	Upper Bound.
ITA	Concise	6.34	6.18	6.49
USA	Concise	5.73	5.56	5.90
ITA	Entern.	5.42	5.25	5.59
USA	Entern.	4.95	4.77	5.14
ITA	Reliable	5.78	5.62	5.93
USA	Reliable	5.39	5.22	5.56
ITA	Trustworthy	6.04	5.88	6.20
USA	Trustworthy	5.47	5.30	5.64

Table 5.2: Average scores of posts in four dimensions for which the estimates differ across *Validation-ITA* and *Validation-USA*. Focus on reliable posts.

accurate, coherent, and comprehensive (and so on) as their counterparts in the US-based dataset.

However, some country-specific differences exist. Reliable posts in the two countries received different average scores in the other four dimensions: *conciseness*, *entertainment*, *reliability* and *trustworthiness*. Italian posts are perceived as slightly better than their US-based counterparts in all of these cases. Table 5.2 below summarises all the dimensions which have statistically different estimates across the two contexts.

Figure 5.6 shows that, even for fake news, in general, posts are perceived very similarly across the two countries. The confidence intervals of the two national estimates overlap in nine out of 12 adjectives: *accurate*, *coherent*, *concise*, *entertaining*, *fair*, *reliable*, *trustworthy*, *vivid* and *well-written*. On all of these dimensions, fake news in the Italian dataset is statistically indistinguishable from their counterparts in the US-based dataset.

However, even in this case, some country-specific differences exist. Posts in the two countries received different average scores in the other three dimensions: *comprehensiveness*, *descriptiveness*, and *interest* (with a very narrow distance from the overlap). Contrary to what I said for reliable posts, US-based fake posts are perceived as slightly better than their Italian counterparts in all these cases. Table 5.3 below summarises all the dimensions with statistically different estimates across the two contexts.

To summarise, posts received statistically indistinguishable scores in *Validation-ITA* as in *Validation-*

Country	Adjective	Average	Low Bound.	Upper Bound.
ITA	Comprehensive	4.51	4.33	4.70
USA	Comprehensive	5.02	4.83	5.20
ITA	Descriptive	4.68	4.50	4.86
USA	Descriptive	5.31	5.12	5.49
ITA	Interesting	4.69	4.49	4.88
USA	Interesting	5.08	4.88	5.27

Table 5.3: Average scores of posts in three dimensions for which the estimates differ across *Validation-ITA* and *Validation-USA*. Focus on fake news posts.

USA in most dimensions. In dimensions where this is not the case, the estimates in the two countries differ by around half a point on a scale from one to ten, at best. In conclusion, besides some cross-country differences in how posts are qualitatively perceived, results indicate that collected social media posts are largely comparable across the two countries and the 12 measured dimensions.

5.2 Perceived Political Leaning of the Social Media Posts

The stimuli used to investigate motivated reasoning, which posits that individuals are more inclined to believe information aligning with their preexisting beliefs, always necessitate a non-trivial requirement: we need to know *which* beliefs they are affirming. In the context of the present work, we need to measure the perceived partisanship of the collected social media posts: are they right-leaning? or left-leaning? The two validation studies addressed precisely this issue, leveraging items from existing literature (see Section 4.2).

Figure 5.7 and 5.8 illustrate the average perceived partisanship of the 80 social media posts, rated on a scale from 0 (representing "Left" in the Italian context and "Democratic" in the US context) to 10 (indicating "Right" in the Italian context and "Republican" in the US context). Each point denotes the average score given to a post, estimated with a t-test, along with the 95% confidence interval of the estimate, colour-coded based on its veracity. The dashed line at the centre of each plot represents the midpoint of the scale, denoted as *"Equally favourable for Democrats and Republicans"* (in *Validation-USA*).

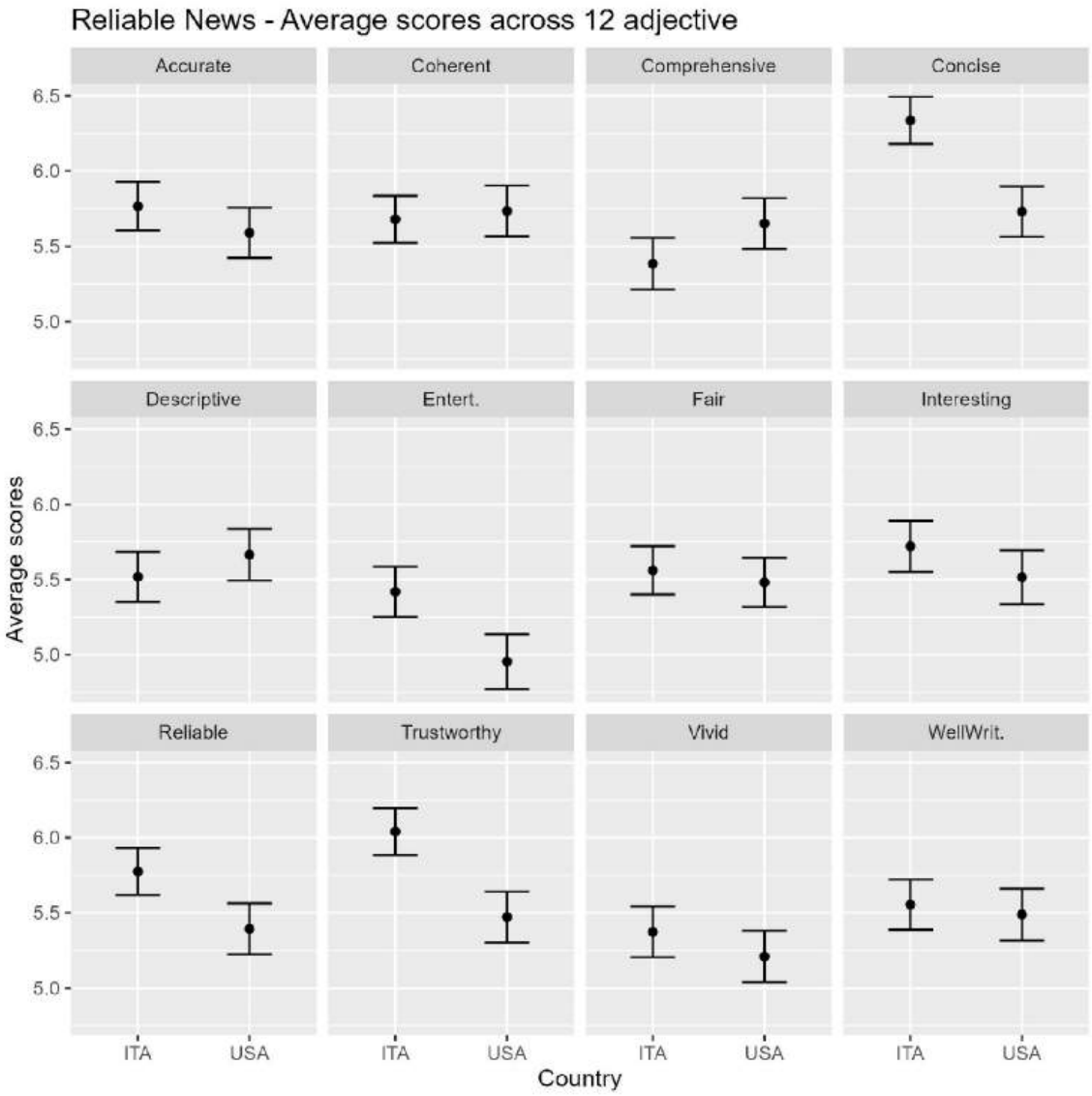


Figure 5.5: Average scores given to reliable posts in *Validation-ITA* (ITA), compared to those in *Validation-USA* (USA). Each subplot focuses on a different adjective. Point estimates resulting from one-sample t-tests, coupled with 95% confidence intervals.

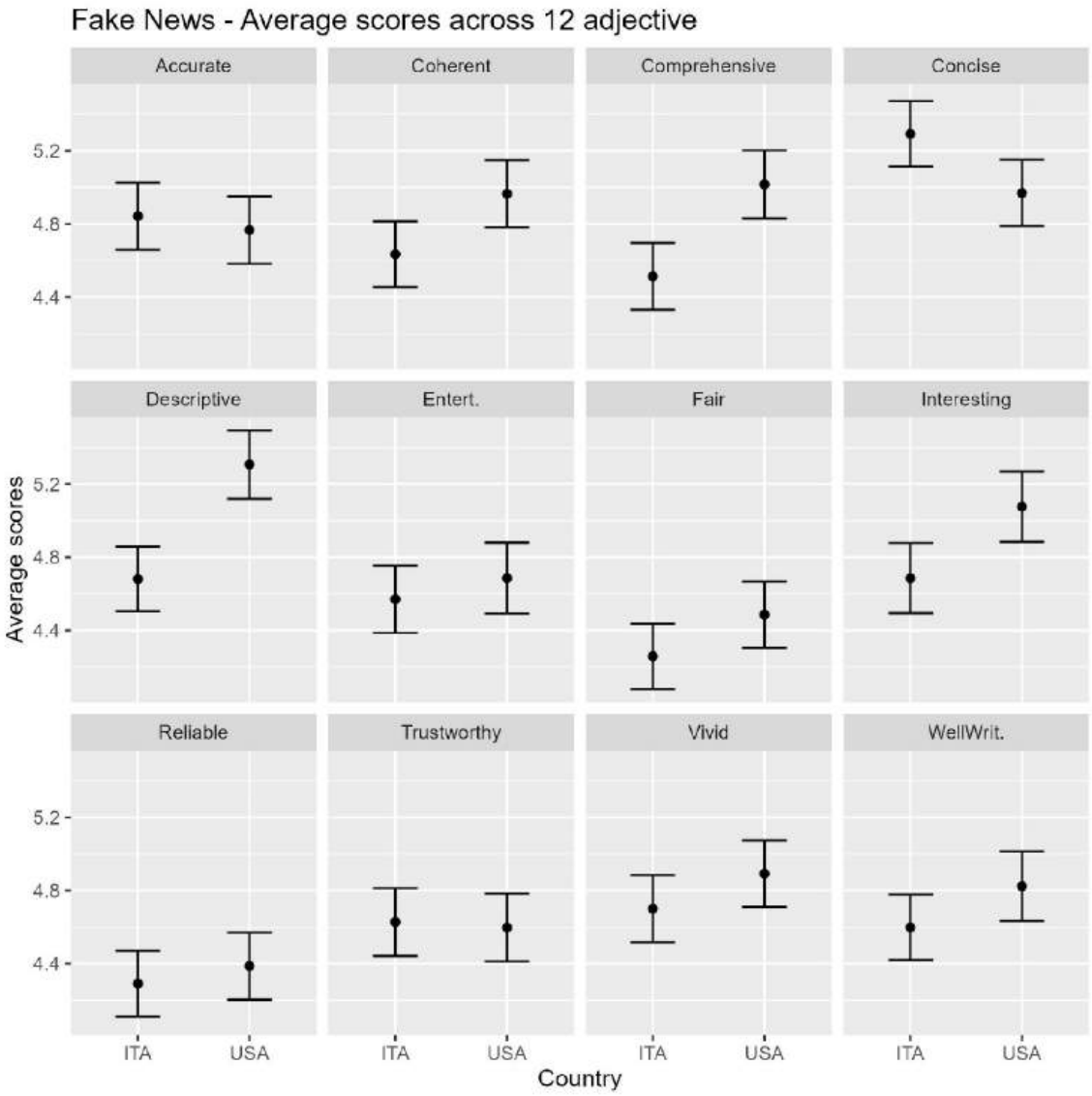


Figure 5.6: Average scores given to fake news in *Validation-ITA* (ITA), compared to those in *Validation-USA* (USA). Each subplot focuses on a different adjective. Point estimates resulting from one-sample t-tests, coupled with 95% confidence intervals.

As can be seen in Figure 5.8, *Validation-USA* portrays a relatively balanced representation of Republican-leaning, Democratic-leaning and neutral posts. Instead, Figure 5.7 shows how the Italian posts tend to skew more toward the right-leaning spectrum. Additionally, in *Validation-USA*, fake news posts are more frequently perceived as Republican-leaning than vice-versa, whereas reliable posts exhibit the opposite trend. On the contrary, such a pattern between partisanship and veracity is less pronounced in the Italian dataset.

However, if we do not stop at observing the average scores and we go further in evaluating the *significance* of these differences, the situation is much more consistent across both countries. Generally, the 80 social media posts are perceived as marginally partisan and, more frequently, as neutral. In the vast majority of cases, the confidence intervals intersect with the dashed line representing neutrality. Among the 40 US posts, only 4 (all fake news) exhibit a significant departure from neutrality, while this count rises to 14 in the Italian case (comprising 9 fake news and 5 reliable posts).

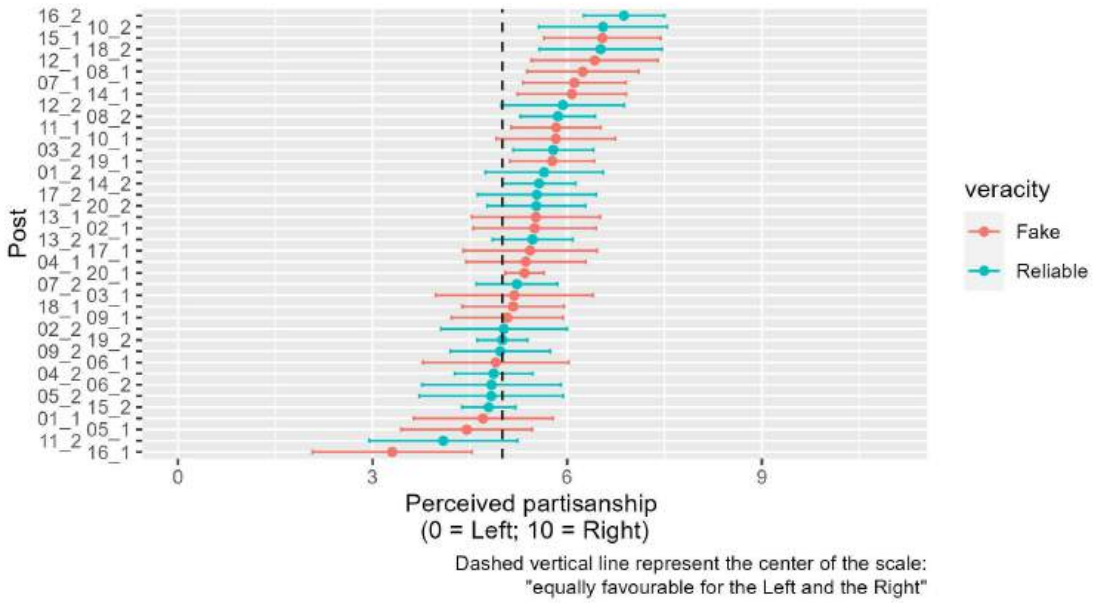


Figure 5.7: Perceived political leaning of the social media posts in the Italian case (*Validation-ITA*)

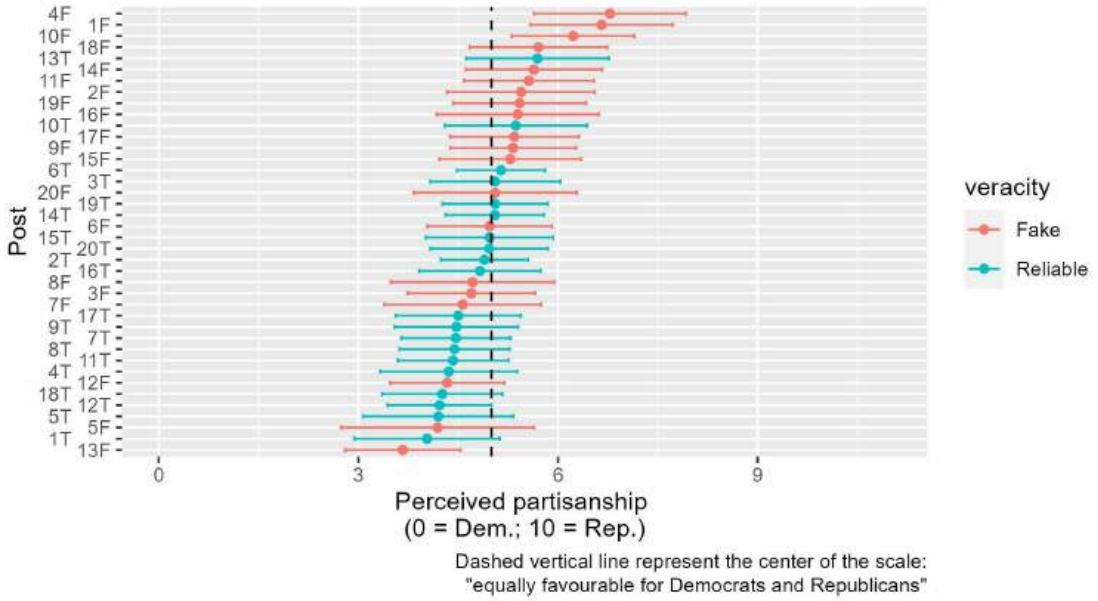


Figure 5.8: Perceived political leaning of the social media posts in the US case (*Validation-USA*)

The above-mentioned neutrality of posts in both studies can be elucidated in two ways: either the posts are genuinely perceived as non-partisan, perhaps because they do not broach political topics or events, or they are deemed simultaneously leaning towards both ends of the political spectrum. In a polarized environment, it is plausible that both factions perceive the media or public discourse as hostile[2]. Consequently, participants from one side might interpret posts as favourable to other factions, and vice-versa. This dynamic could result in an average score that appears neutral, while the distribution of individual scores is, in reality, bi-modal.

To investigate this possibility, I created a new variable named "intensity of partisanship", which quantifies the distance from the scale's midpoint, regardless of the direction of this distance. This approach ensures that scores reflecting opposite partisan leanings for the same post are not counterbalanced but averaged, thus capturing the potential bi-modality of individual perceptions[3]. This

[2]Evidence confirming this hypothesis can be found in van der Linden et al. (2020). They found that liberals associate traditionally right-wing media sources with the term fake news and conservatives do the same with left-wing media.

[3]In mathematical terms, this second variable is just a collapsed version of perceived partisanship: $I = |P - 5|$ where I represent the intensity of partisanship, P represent the score of perceived partisanship from 0 to 10 and 5

alternative formulation enables an assessment of whether a neutral score arises from low perceived partisanship or from two divergent, and thus mutually negating, high perceived partisanship scores.

The analysis reveals that, particularly in the Italian context, posts are neutral because they are rated with scores very close to the centre of the scale. In most cases, the average distance of scores from the scale's midpoint is below two (on a 6-point scale). While the situation in *Validation-USA* mirrors this trend, the intensity of partisanship is slightly higher (roughly by 1 point). US posts elicit more polarized perceptions than Italian ones, albeit they appear more neutral. The polarization is hidden exactly by diverging scores compensating for each other.

It is noteworthy that fake news tends to be perceived as more partisan than reliable posts, especially in *Validation-USA*. This reaffirms the observation, noted in the analysis of the perceived quality of posts, that reliable posts embody some aspects of traditional journalism, particularly fairness.

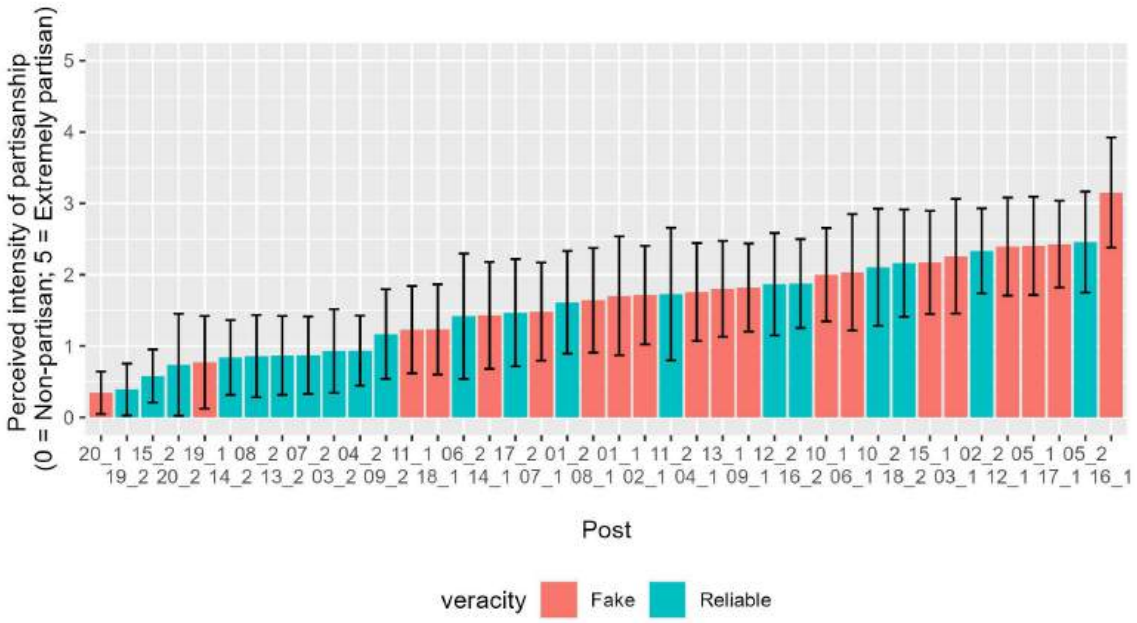


Figure 5.9: Perceived intensity of partisanship of the social media posts in the Italian case (*Validation-ITA*)

is the midpoint of the scale. In this way, two extreme scores P, for example, 0 and 10, are represented as the same value of the intensity of partisanship I: 5.

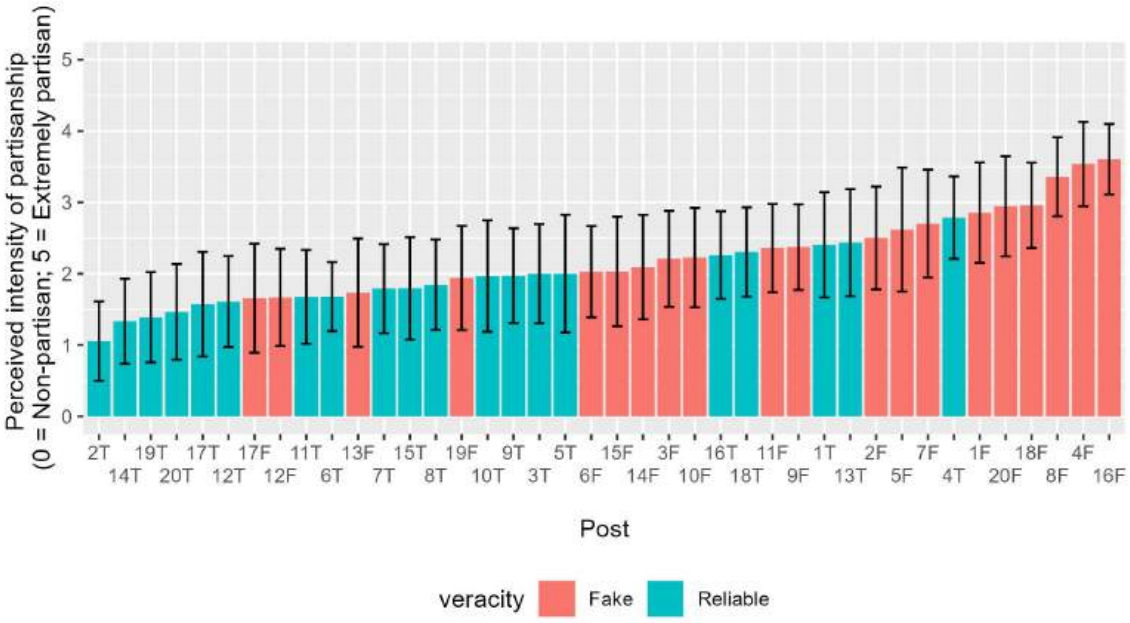


Figure 5.10: Perceived intensity of partisanship of the social media posts in the US case (*Validation-USA*)

To summarise, while the Italian sample of news presents all the possible combinations of facticity and perceived partisanship, this is not the case in the US sample. In particular, none of the reliable news in the US sample is perceived as significantly favouring one specific faction. The confidence intervals of their scores always intercept the midpoint of the scale, which identifies posts that equally favour Democrats and Republicans.

The absence of partisan posts prevents the test of motivated reasoning in *Main-US*, as will be explained in Chapter 8. It is not possible to test whether participants are less accurate when judging news that favours their political faction if the news are not favouring any faction. More precisely, the absence of partisan reliable news prevented the computation of truth discernment at different levels of political congruence.

The absence of partisan reliable posts is not an issue that can be solved by excluding items from the US pool. Instead, it would require collecting and validating additional posts, an operation outside the resources available for this project. Additionally, I have no indications that integrating

new stimuli would automatically result in the presence of partisan reliable news, as the collection already involves 20 reliable items and none is considered partisan. Also in light of the considerations made at the beginning of this chapter regarding perceived partisanship not being a crucial exclusion criterion for posts, the final decision was to retain the initial news set and limit the analysis to the influence of political ideology and partisanship for *Main-USA*. Instead of continuing the social media posts sampling until a set of partisan reliable posts emerged, the decision was to acknowledge that the perceived neutrality of reliable posts in *Validation-USA* prevented the exploration of motivated reasoning in this *Main-USA*.

Chapter 6

Results of the Convenience Sample

As anticipated (refer to Chapter 4), *Convenience-ITA* is the first of three survey experiments aimed at testing multiple competing mechanisms that could explain truth discernment. In addition, it also experimentally manipulates the perceptions of social norms through a treatment message. This initial study engaged a convenience sample comprising around 600 Italian internet users, with an average duration of approximately 15 minutes. Compared to the general Italian population, this sample presents an over-representation of central age groups, particularly individuals aged between 35 and 54, as well as a higher proportion of women and individuals with advanced educational backgrounds. Regarding political affiliation, representation from all major parties is present in the sample, albeit with certain disparities: "Sinistra Italiana", "Movimento 5 Stelle" and "Lega" are over-represented, while "Forza Italia" and "Fratelli d'Italia" are under-represented.

This chapter aims to provide a concise summary of the study results, starting with the re-evaluation of previously explored cognitive drivers (see Section 6.1 and Section 6.2), followed by an examination of the role of anti-elitism and institutional trust (Section 6.3 and Section 6.4). Section 6.5 will focus on the outcomes concerning the potential presence of an "accurate sharing" social norm and its association with truth discernment. The chapter will then conclude with some considerations on the treatment effect of the social norm priming message in Section 6.6.

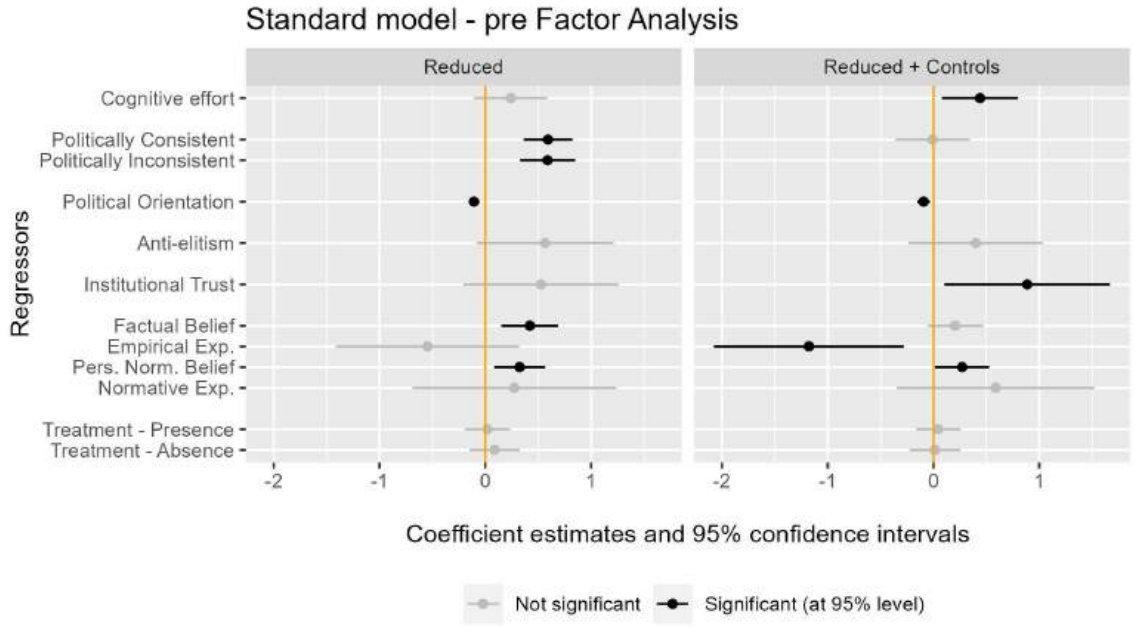


Figure 6.1: Reduced OLS regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

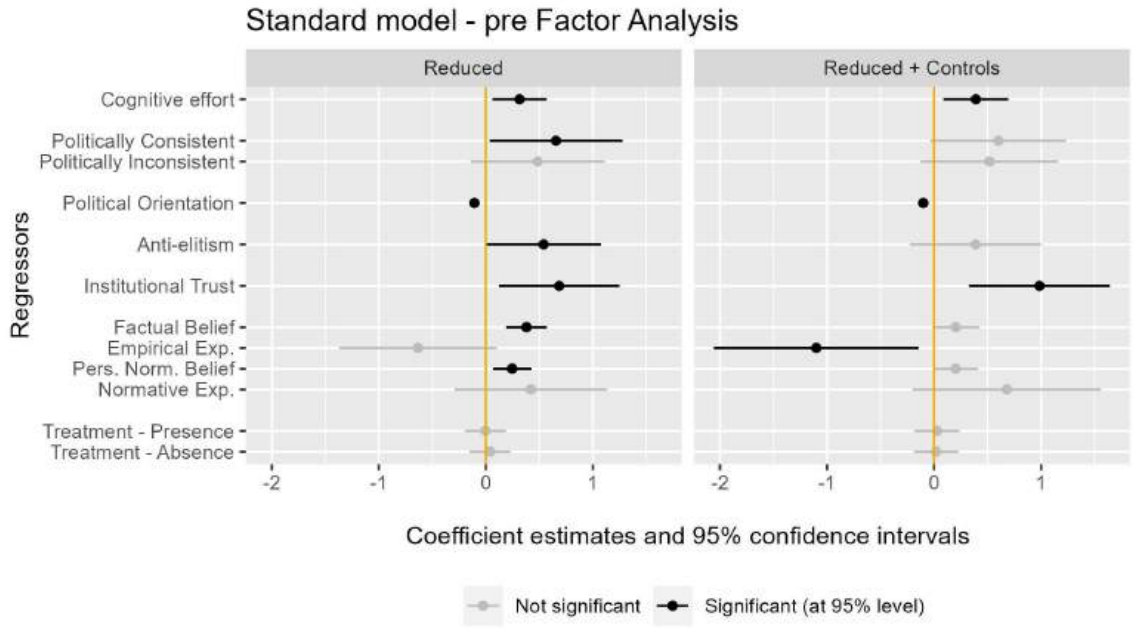


Figure 6.2: Reduced MLM regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

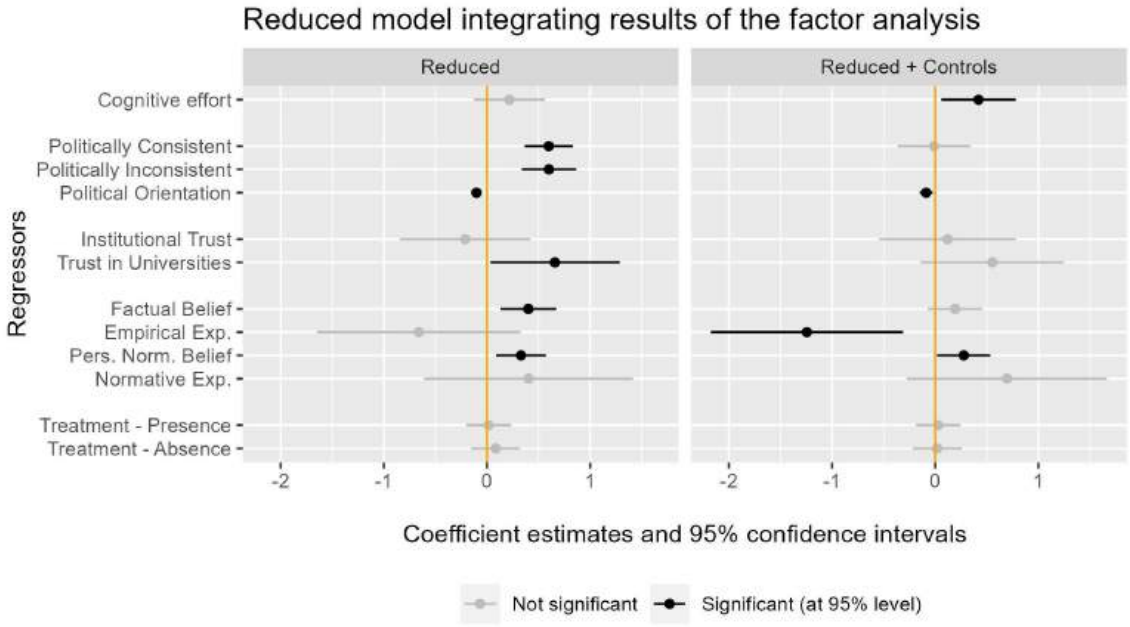


Figure 6.3: OLS regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

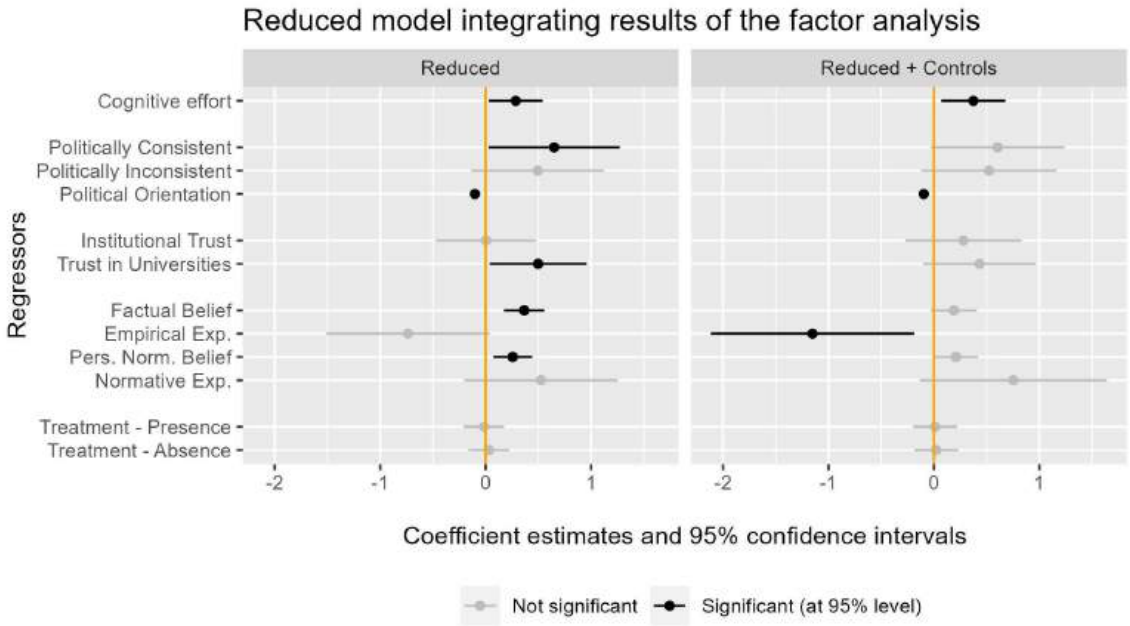


Figure 6.4: MLM regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

6.1 Cognitive Styles

The first analysed mechanism is the role of cognitive style. To reference back to the theoretical framework of this project, participants with a more analytical cognitive style are expected to have better truth discernment (Chapter 3). Cognitive styles are measured with Cognitive Reflection Tasks. Higher scores in these tasks are thought to be linked with more analytical thinking (Chapter 4).

The coefficient measuring the effect of cognitive styles on truth discernment is significant in bivariate regression models. This is true when using a formulation that includes clustered standard errors on the level of participants ($\beta = 0.601$, $p < 0.001$), and a bivariate formulation which also includes fixed effects at the level of stimuli ($\beta = 0.618$, $p < 0.001$). In other words, the gap between scores given to reliable *versus* fake posts by participants with perfect scores in the CRT is around 0.6 points wider (on a scale from zero to five), compared for the same gap for participants with a CRT score of zero.

However, when this variable is tested in multivariate regression models, the significance of its estimate is less stable (Figure 6.1 to 6.4, and Table 6.1). The effect of a perfect CRT score on truth discernment is not significant in the *Reduced* model which includes only the major explanatory variables and clustered standard errors at the level of participants ($p = 0.174$). The same is true for the *Reduced* model where some explanatory variables have been substituted with their corresponding indexes resulting from factor analysis ($p = 0.216$).

Instead, the effect of an analytical cognitive style is significant in models which include controls and fixed effects at the level of stimuli. This is true both when using the model with the original formulation of indexes ($\beta = 0.44$, $p = 0.017$) and when using indexes emerging from factor analysis ($\beta = 0.42$, $p = 0.023$).

In addition, the coefficient for CRT scores is significantly associated with a higher truth discernment in all multilevel models (Figure 6.2 and 6.4; for point estimates and regression tables, refer to

Appendix A). This is true in all bivariate formulations ($p < 0.001$) and in multivariate models. The coefficient is positive and significantly different from zero in the *Reduced* model with and without controls (Figure 6.2). The same is true in formulations which use indexes resulting from factor analysis (Figure 6.4), with and without controls.

To summarise, results in *Convenience-ITA* seem to confirm that participants with a more analytical cognitive style as measured in CRT have higher truth discernment. This association is present in bivariate models, reduced OLS models with controls (*Reduced + Controls*) and all multilevel models, in support of Hypotheses 1. However, the null hypothesis is not refuted for multivariate OLS models without controls and stimuli fixed effects. In these formulations, the coefficient for CRT scores is not statistically significant.

6.2 Motivated Reasoning

To measure motivated reasoning, each post was categorized as either consistent, inconsistent, or neutral to the respondent's political orientation. For example, a post judged as left-leaning in *Validation-ITA* is categorised as politically consistent for participants with a leftist political positioning. The same post is considered inconsistent for rightist participants and neutral for participants who place themselves in the middle of the political spectrum. Posts with a statistically indistinguishable score from the perceived partisanship scale's centre are all labelled as neutral.

Compared to cognitive styles' results, those concerning the role of motivated reasoning present a less definitive picture. Broadly speaking, participants exhibit enhanced news judgment capabilities when evaluating both consistent ($\beta = 0.593$, $p < 0.001$) and inconsistent ($\beta = 0.589$, $p < 0.001$) news compared to neutral news (see Figure 6.1 and Table 6.1). This result is confirmed in bivariate models ($p < 0.001$ for all coefficients) and the multivariate OLS model with factor analysis indexes ($p < 0.001$ for all coefficients).

The OLS models with controls and stimuli fixed effects present a problem in computing the two

coefficients. The models compute only the coefficient for consistent posts, while the estimate for inconsistent posts is not calculated. This is due to the introduction of stimuli fixed effects, which are calculated on the same level as political consistency (the level of stimuli). The estimated coefficient is not significant ($p = 0.956$ in *Reduced + Controls* and $p = 0.949$ in *Indexes + Controls*).

The computation of multilevel regression models does not present the same problem (see Figure 6.2 and Figure 6.4, tables in Appendix A). These models allow for accounting for the cross-classified data structure without introducing items' fixed effects. In these models, coefficients for consistency are significant. This is true when using the *Reduced* formulation ($\beta = 0.655$, $p = 0.038$) and in the formulation including factor analysis indexes ($\beta = 0.65$, $p = 0.04$). However, in models with controls, the significance persists only at a 90% level ($p = 0.063$ and $p = 0.062$). The coefficient for inconsistency, on the contrary, is never significant (for example, $p = 0.130$ in the *Reduced* model, and more or less on the same level for other formulations).

No statistically significant findings emerged regarding the comparison between consistency and inconsistency. Participants demonstrated similar proficiency in judging consistent and inconsistent news, as highlighted by the coefficients' magnitude (presented above). This contradicts the expectations of motivated reasoning encapsulated in Hypothesis 2. This hypothesis posits that individuals would be more inclined to believe consistent information over inconsistent information. This observation is also confirmed by simplified models (not presented here) comprising solely the explanatory variable and truth discernment.

To further explore the comparison between consistency and inconsistency, I computed an alternative formulation of this variable (called *Dichotomous-Pol.Cons.*). In this formulation, answers were categorized based on the perceived partisanship of posts as a dichotomous category. Posts were divided into just two (instead of three) categories - left-leaning and right-leaning - solely based on their average score, without considering whether this score significantly differed from neutrality (in *Validation-ITA*). Political consistency was then constructed by combining the post's perceived

partisanship with the respondent's declared political orientation. The resulting variable only has two levels: politically consistent and politically inconsistent. This variable, tested only in bivariate models, exhibited significance and validated Hypothesis 2. Participants are worse at judging consistent posts when compared to inconsistent ones (β = -0.278, p = 0.006 in the model with items fixed effects). However, it must be noted that this formulation is a rather strong stretch of the results. In fact, it forces the neglect of political neutrality of posts (which is very common), categorizing them into either consistency or inconsistency.

6.3 Anti-elitism

The analysis of potential associations between anti-elitism and participants' truth discernment begins with bivariate regression models. In these models, the three items for anti-elitism are included one at a time in models that just include them, veracity of items and their interaction with it, predicting perceived accuracy. The three items are called *Talk-too-much*, *Difference-with-people*, and *Don't-care-about-people*. For the complete wording of these questions, see Section 4.2. The last item (*Don't-care-about-people*) is a reverse-coded version of the original question.

Taken singularly in bivariate models, only *Don't-care-about-people* (reverse-coded version of *"Politicians care about what ordinary people think"*) significantly associates with participants' truth discernment (β = 0.415, p = 0.021 in the model with items fixed effects). The direction of this association is positive, contrary to expectations. Participants who believe politicians do not care about what ordinary people think are better at discerning news. The other two variables measuring anti-elitism are not associated with truth discernment (p = 0.273 for *Talk-too-much* and p = 0.331 for *Difference-with-people*, in models with fixed effects).

The index constructed as a simple average of scores for the three anti-elitism items is never significant, except in the MLM *Reduced* formulation without controls (p = 0.047, see Appendix A.1 and Figure 6.2). Surprisingly, even in this instance, the relationship between anti-elitism and truth

discernment contradicts the hypothesized direction ($\beta = 0.54$).

To summarise, participants with higher scores in anti-elitism questions typically exhibit similar news discernment capabilities as other participants, and in cases of divergence, they outperform those with lower anti-elitism scores. Overall, this study fails to support Hypotheses 3a as anti-elitism is never negatively associated with truth discernment.

A factor analysis (not presented here) tested whether anti-elitism is a standalone concept or could be incorporated with institutional trust in a *distrust-trust* comprehensive index. The analysis was conducted on all the numerical variables, to reveal other potential clusters of variables. Results highlight how the three anti-elitism items have only slight correlations. Furthermore, descriptive bivariate analysis (not presented here) and bivariate regression models demonstrated how the apparent positive association with truth discernment is primarily driven by a single item: *Don't-care-about-people*. Finally, this item shared a high and negative correlation with trust in political institutions. In other words, factor analysis revealed how *Don't-care-about-people* could be used as a measure of institutional distrust as "presence of negative expectations toward institutions" (see Chapter 3), rather than as a separate indicator. Consequently, *Don't-care-about-people* was aggregated with items measuring trust in political institutions, with which it also shared a similar relation with truth discernment. The resulting index for institutional trust was included in the *Indexes* formulation of regression models, presented in Figure 6.3 and 6.4 and commented in Section 6.4.

6.4 Institutional Trust

The association between trust in various institutions and truth discernment is first analysed through separate OLS bivariate models with participants clustered SE (standard errors), with and without items FE (fixed effects). Trust in only two out of the five measured institutions is significantly associated with truth discernment. Participants who trust social media have a significantly lower

truth discernment (β = -0.488, p = 0.01 in model with FE). On the contrary, those who trust universities are better at discerning news (β = 0.805, $p < 0.001$). Trust in the press (p = 0.054), the Parliament (p = 0.847) and the Government (p = 0.152) are not significantly associated with truth discernment.

Trust in institutions, quantified by an index averaging scores from all institutions, is consistently significant and positively linked with truth discernment across all multivariate formulations (p = 0.027 in OLS *Reduced + Controls*, see Figure 6.1 and 6.2, Table 6.1 and A.1 for other estimates), except in the *Reduced* model without control, where it is non-significant (p = 0.16). Initially, this appears to support Hypothesis 3b. Participants with higher trust in institutions are better at discerning news. However, this aggregate result hides the different directions of bivariate associations highlighted above. The positive effect of institutional trust is mainly driven by trust in universities.

I thus ran factor analysis to aggregate the different dimensions of institutional trust based on their correlations rather than on *a-priori* considerations. Factor analysis underscores the tight correlation between trust in Parliament, Government and the press, contrasting with the weaker correlations of trust in universities and social media with other institutions. In addition, these are the two variables with significant associations with truth discernment. Consequently, I opted to separate these latter variables and construct an index solely incorporating trust in Parliament, the Government, and the press. The third anti-elitism item is also incorporated into this index due to its high correlation with these variables and a similar relationship with the dependent variable.

Trust in universities is added as a separate variable in *Indexes* multivariate formulations. Instead, trust in social media is too correlated with other variables to include it in the analysis without introducing multicollinearity. In addition, theoretical mechanisms highlighted in Chapter 3 are more difficult to apply on social media than other institutions. Consequently, I decided to exclude it from the multivariate analysis.

Models incorporating this new index alongside trust in universities as a separate variable demon-

strate that trust in institutions is not associated with truth discernment (see Figure 6.3 and 6.1 and *Indexes* and *Indexes + Controls* models in Table 6.1 and A.1). The apparent positive relation is likely driven by trust in universities, which exhibits significance and a positive link with truth discernment when isolated from other institutions ($\beta = 0.659$, $p = 0.04$). However, even the significance of trust in universities is not robust to including controls ($p = 0.119$ in the OLS *Indexes + Controls* formulation).

In conclusion, Hypothesis 3b is not supported in *Convenience-ITA* as the relationship between trust in institutions and truth discernment is contingent on the institution under consideration. In most instances, this relationship is non-existent and sometimes even negative, as in the case of social media. Instead, the influence of trust in universities is positive and occasionally significant. Thus, HP3 is supported only for trust in universities.

6.5 Social Norm on the Importance of Accuracy

Before exploring the potential connection between participants' truth discernment and their perception of a social norm regarding news sharing, it is important to establish the existence of such a norm within the sample.

To begin, I will outline what motivates participants to share news online (factual beliefs) and what they believe *should* motivate them and their peers (personal normative beliefs). Subsequently, I will delineate participants' expectations concerning what motivates others to share news online (empirical expectations) and what they believe others stated *should* motivate news sharing (normative expectations). Finally, I will provide a summary of correlational evidence regarding the influence of these beliefs and expectations on truth discernment.

6.5.1 Does an "Accurate Sharing" Social Norm Exist?

Figure 6.5 depicts the distribution of responses to four questions aimed at gauging the existence of an "accurate sharing" social norm within the Italian convenience sample. Notably, when asked about

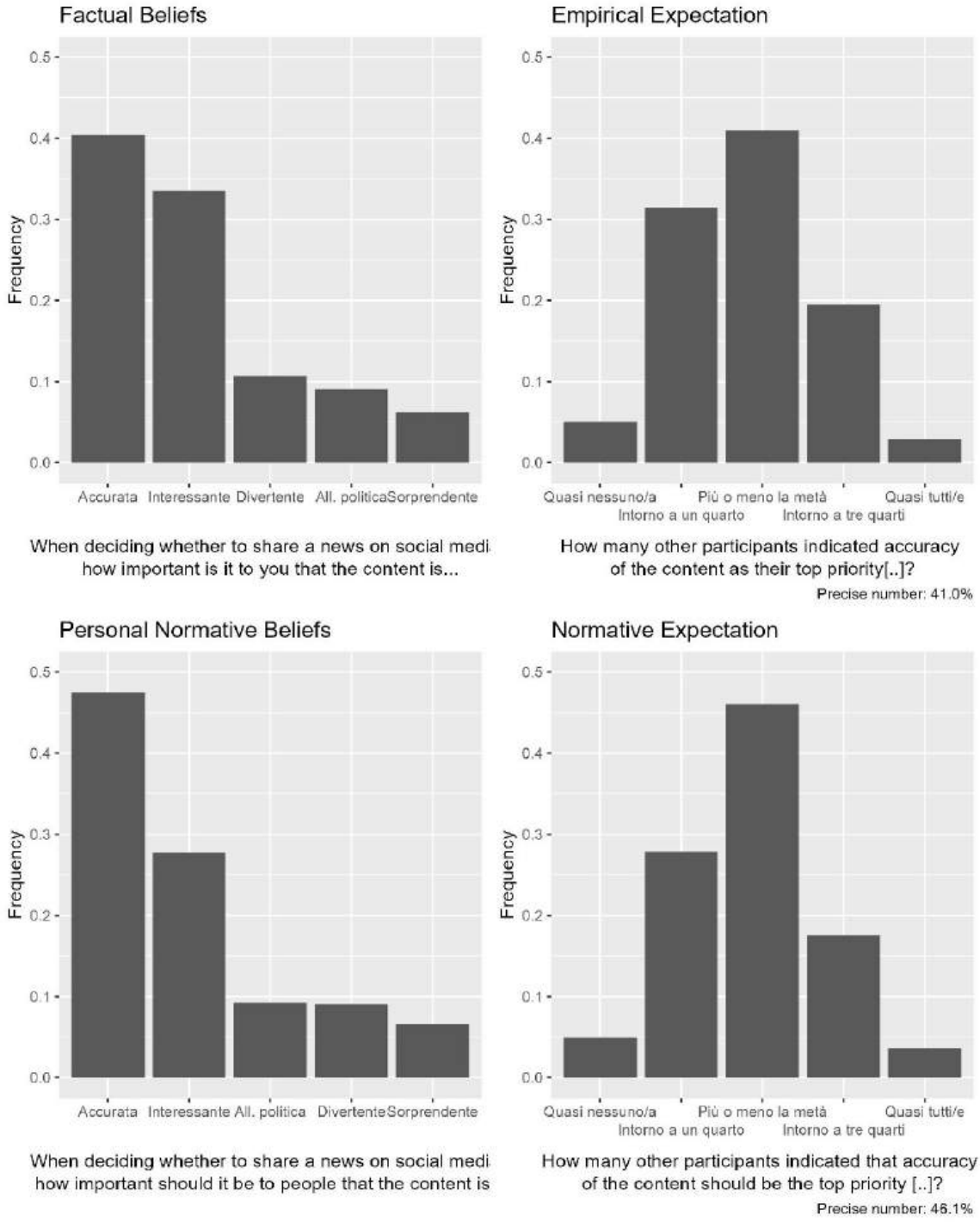


Figure 6.5: Descriptive univariate analysis of the presence of beliefs and expectations regarding the presence of an "accurate sharing" social norm.

the characteristics motivating their news-sharing decisions, 40.46% of participants cited accuracy as their primary priority, with another 33.55% indicating that the most important factor is for the news to be interesting. Similarly, personal normative beliefs (what *should* motivate news sharing) reflect a comparable distribution, although with a more pronounced gap between "accurate" (47.45%) and "interesting" (27.75%).

In essence, the mode suggests that, according to participants, accuracy is, and *should be*, their first priority when deciding what to share. However, more than half of the sample prioritizes other motivations in both cases. Another consistently chosen factor is relevance, with around 30% of participants indicating that a piece of news must be interesting before they share it and that this should be the case for everyone.

Concerning expectations, participants are asked to guess an estimate of how many of their peers chose *accuracy* as their priority in the questions about factual and personal normative beliefs. Considering what I just said about answers to these questions, the "correct" answer is "Half of them".

The mode of the distribution of answers indicates that participants expect only half of their peers to choose accuracy as their first priority. This holds true for both empirical (40.98% of the sample choosing "Half of them") and normative expectations (46.06%), even if this figure is slightly higher in the second case. In other words, the most chosen answer is correct for both the questions about empirical expectations and normative ones.

Furthermore, both distributions are right-skewed. In the question about empirical expectations, more participants select "Few participants" (31.47%) and "Almost no one" (5.08%), than "Most of them" (19.5%) and "Almost all of them" (2.95%). A similar situation is reflected in the question about normative expectations (respectively 27.87% and 4.91% compared to 17.54% and 3.8%).

In summary, participants exhibit mixed expectations regarding their peers' sharing behaviour. While they recognize the importance of accuracy in shared news for a significant portion of the

population, most participants doubt its universality.

6.5.2 Correlational Evidence on the Relation with Truth Discernment

Beyond merely measuring the *presence* of the outlined norm, the primary objective of the analysis is to gather correlational evidence on whether the perception of this "accurate sharing" norm influences truth discernment. This inquiry is undertaken similarly to the approach applied to the other variables and mechanisms.

Starting with beliefs, the analysis reveals that participants who prioritize accuracy in news-sharing decisions demonstrate better news discernment than their peers who prioritize other factors. This result holds true in bivariate regression models ($\beta = 0.388$, $p < 0.001$ in model with items FE) and in some multivariate models (Figure 6.1 and 6.2, Table 6.1 and A.1). For example, the coefficient is significant in OLS multivariate model ($\beta = 0.421$, $p = 0.002$) and in its multilevel corresponding version ($\beta = 0.38$, $p < 0.001$). However, the coefficient is not significant when adding controls ($p = 0.122$ and $p = 0.067$). Thus, data support Hypothesis 4a, but not in all model formulations.

The positive association with truth discernment holds true for personal normative beliefs as well. This effect is significant in bivariate models ($\beta = 0.473$, $p < 0.001$ in model with items FE) and in all multivariate models (p ranging from 0.008 to 0.037) except for the multilevel formulation with controls, where it is slightly above the threshold ($p = 0.058$). Thus, data also support Hypothesis 4b, and with a higher number of significant results.

Conversely, coefficients for normative expectations never achieve significance. Participants who believe that all their peers think accuracy should be prioritized are no more proficient at judging posts than those who believe only a small minority do so. This is true for bivariate regression models ($p = 0.781$ in the model with items FE) and in all multivariate regression models (p ranging from 0.097 in MLM *Indexes + Controls* to 0.576 in OLS *Reduced*). Consequently, the null hypothesis is never refuted for Hypothesis 5b.

On the contrary, empirical expectations are sometimes significant, but the observed direction of this link contradicts expectations and hypotheses. Participants who anticipate their peers prioritizing accuracy exhibit poorer discernment between fake news from reliable posts (for example, β = -1.178 in OLS *Reduced + Controls*). In particular, this is true in all multivariate models with controls (p ranging from 0.009 in *Indexes + Controls* to 0.024 in *Reduced + Controls* in the multilevel formulation, see Table 6.1 and A.1). Therefore, Hypothesis 5a is not supported by data from *Convenience-ITA* as empirical expectations, when significant, are negatively linked with truth discernment.

6.6 Treatment Effect

Convenience-ITA also includes an experimental component to assess whether informing participants about their peers' social expectations could alter their ability to discern news. After illustrating the results of the manipulation check, I will show results from a series of analyses to assess the effectiveness of this treatment. The first part is dedicated to descriptive analysis, mainly run with two sample t-tests (average perceived accuracy of reliable posts - avg. perc. acc. of fake news) to calculate truth discernment across different rounds and treatment groups. The second part shows results from regression analysis.

To begin with, participants seem to understand and recall the treatment message. Answers from the manipulation check show that 76% of participants responded correctly when asked to choose the treatment message in a closed-ended question.

A first descriptive analysis compares answers in the two groups before (first five rounds) and after the treatment (last five rounds). Participants in the *Presence* group have an average truth discernment of 0.791 points with a 95% confidence interval of [0.618, 0.963] before the treatment, which lowers to 0.721 p. [0.555, 0.889] after the treatment. Participants in the *Absence* group have an average truth discernment of 0.665 points [0.497, 0.833] before the treatment, which increases

to 0.81 p. [0.641, 0.979] after the treatment.

Instead, Figure 6.6 displays the truth discernment of the two treatment groups in each round. On the one hand, participants' performance in the *Presence* group does not follow the expected increased trend. Instead, after an initial rise *before* the treatment, truth discernment has a stable trend. On the other hand, participants' performance in the *Absence* group has a stable trend until round seven of the accuracy task. After the manipulation check and the following re-treatment, truth discernment increases. However, this increase is not marked enough to create significant differences with truth discernment in other rounds.

The treatment message is substantially ineffective in *Convenience-ITA*. Truth discernment remained statistically constant across all the groups and indistinguishable before and after the treatment. Moreover, when considering the estimates without accounting for confidence intervals, the direction of the changes contradicts the hypotheses: participants exposed to a message emphasizing the absence of a social norm experienced a (non-significant) increase in their truth discernment.

This result underwent rigorous testing through a series of robustness checks. The treatment effect was estimated as a regression coefficient in OLS and MLM models (using the same formulations outlined in previous sections, some of which are depicted in Figure 6.1 to Figure 6.4). Additionally, both the t-test analysis and the regression analysis were repeated using "distance from the truth", an alternative formulation of the dependent variables. Across all analyses, the outcomes consistently yielded a null result regarding the effectiveness of the treatment message (with *p-values* ranging from 0.438 in the bivariate model without FE to 0.965 in the MLM *Reduced* model). Consequently, data from *Convenience-ITA* fails to support Hypothesis 6a and 6b.

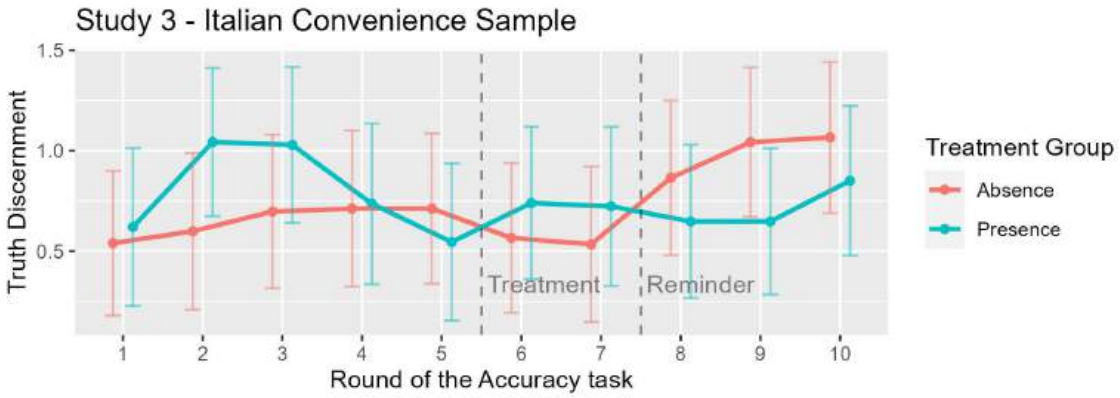


Figure 6.6: Truth Discernment in the Treatment Groups (with 95% CI).

Table 6.1: OLS Regression Models - Convenience-ITA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.241	0.440	0.217	0.420
	0.177 (0.174)	0.184 (0.017) *	0.175 (0.216)	0.184 (0.023) *
Pol. Consistency	0.593	−0.010	0.599	−0.011
	0.118 (<0.001) ***	0.179 (0.956)	0.118 (<0.001) ***	0.180 (0.949)
Pol. Inconsistency	0.589		0.601	
	0.133 (<0.001) ***		0.135 (<0.001) ***	
Pol. Orientation	−0.107	−0.093	−0.102	−0.088
	0.026 (<0.001) ***	0.031 (0.003) **	0.026 (<0.001) ***	0.032 (0.006) **
Anti-elitism	0.566	0.400		
	0.331 (0.087) +	0.324 (0.217)		
Inst. Trust (original)	0.527	0.886		
	0.375 (0.160)	0.399 (0.027) *		
Fact. Bel. - Accurata	0.421	0.206	0.401	0.193
	0.137 (0.002) **	0.133 (0.122)	0.139 (0.004) **	0.134 (0.151)
Emp. Exp. - Intorno a un quarto	0.092	0.290	0.053	0.241
	0.285 (0.747)	0.282 (0.304)	0.289 (0.856)	0.292 (0.409)
Emp. Exp. - Più o meno la metà	0.087	0.250	0.041	0.207
	0.316 (0.782)	0.305 (0.413)	0.323 (0.898)	0.318 (0.516)
Emp. Exp. - Intorno a tre quarti	−0.052	0.101	−0.111	0.057
	0.322 (0.871)	0.320 (0.752)	0.327 (0.734)	0.333 (0.863)
Emp. Exp. - Quasi tutti/e	−0.546	−1.178	−0.662	−1.244
	0.442 (0.217)	0.458 (0.010) *	0.503 (0.188)	0.475 (0.009) **
Pers. Norm. Bel. - Accurata	0.324	0.271	0.330	0.276
	0.123 (0.008) **	0.130 (0.037) *	0.124 (0.008) **	0.131 (0.035) *
Norm. Exp. - Intorno a un quarto	0.144	0.069	0.180	0.095
	0.364 (0.693)	0.335 (0.837)	0.370 (0.626)	0.352 (0.788)
Norm. Exp. - Più o meno la metà	−0.060	−0.167	−0.012	−0.136
	0.390 (0.879)	0.354 (0.637)	0.396 (0.976)	0.369 (0.713)
Norm. Exp. - Intorno a tre quarti	0.251	0.209	0.306	0.255
	0.403 (0.533)	0.373 (0.575)	0.411 (0.456)	0.387 (0.510)
Norm. Exp. - Quasi tutti/e	0.275	0.588	0.403	0.695
	0.492 (0.576)	0.477 (0.218)	0.517 (0.436)	0.495 (0.160)
Treatment - Presence	0.019	0.045	0.016	0.029
	0.109 (0.861)	0.108 (0.678)	0.110 (0.887)	0.109 (0.792)
Treatment - Absence	0.089	0.015	0.084	0.019
	0.121 (0.462)	0.122 (0.901)	0.121 (0.489)	0.123 (0.876)
Trust in Universities			0.659	0.553
			0.320 (0.040) *	0.355 (0.119)
Inst.Trust (factor analysis)			−0.211	0.120
			0.323 (0.512)	0.338 (0.721)
Num.Obs.	4229	3721	4205	3707
AIC	23 723.9	19 981.3	23 604.2	19 915.7
BIC	50 221.3	42 000.1	49 925.6	41 834.1

Chapter 7

Results of the Italian Quota Sample

Main-ITA is the second main study of this project and it has the same structure and hypotheses as *Convenience-ITA* and *Main-USA*. *Main-ITA* was conducted on a quota sample comprising approximately 1200 Italian internet users. Quotas were set to replicate the distribution of age, gender, and education observed in the general Italian population. During data collection, answers were collected by Qualtrics until each quota was filled.

Even if Qualtrics communicated that all the quotas had been successfully filled, the final sample only partially represents the population across the above-mentioned variables. Descriptive univariate analysis shows that while the gender distribution matches the general population, the sample is younger than expected, under-representing individuals aged 55 or older. However, the most notable disparity is observed in education. The sample is markedly more educated than the general population.

Regarding political affiliation (for which quotas were not set), the sample closely mirrored the proportions seen in polling data for most parties, albeit with slight over-representation of some, such as "Movimento 5 Stelle", "più Europa", "Sinistra Italiana/Europa Verde", and "Italia Viva". Notably, as mentioned before, Forza Italia is not represented due to an error during the coding phase. Nevertheless, robustness checks (not presented here) indicate that Forza Italia voters likely

opted not to answer rather than switching to other parties, thus ending up excluded from the study[1].

This chapter will summarise the results from *Main-ITA* utilizing the same analytical approach and visualizations as presented in Chapter 6 for *Convenience-ITA*. For a detailed explanation of the plots and analysis methodology, please refer to Chapter 4.

7.1 Cognitive Styles

The positive link between an analytical cognitive style and truth discernment observed in *Convenience-ITA* is similarly confirmed for *Main-ITA*, as highlighted in Figure 7.1 to 7.4 and in Table 7.1 and A.2. Participants with higher CRT scores consistently demonstrate better judgment of news posts than those with lower scores. This trend is evident across all types of analysis and model formulations.

First, this positive association is confirmed in bivariate regression models with ($\beta = 0.618$, $p < 0.001$) and without items fixed effects ($\beta = 0.601$, $p < 0.001$). In addition, it is present in all OLS multivariate formulations (β ranging from 0.457 to 0.555, p always below 0.001). Finally, this result is consistent in all multilevel regression models (β ranging from 0.417 to 0.514, p always below 0.001). Therefore, *Main-ITA* finds support for expectations of Hypothesis 1.

7.2 Motivated Reasoning

Results for motivated reasoning in Study 4 exhibit similarities to those observed in Study 3, albeit with generally lower significance. Overall, participants tend to perform better at judging partisan posts, both politically consistent and inconsistent, compared to neutral ones. However, this finding only achieves significance in certain formulations.

In bivariate regression models, participants have a better truth discernment when judging both consistent ($\beta = 0.436$, $p < 0.001$) and inconsistent posts ($\beta = 0.870$, $p < 0.001$).

[1] In this study, as in the others, participants were not compelled to answers every question to progress through the survey. Instead, if a question went unanswered, a reminder appeared to prompt participants to provide a response. If participants still chose not to answer, the questions were recorded as unanswered. Subsequently, Qualtrics excluded all participants with unanswered questions, as this was one of its exclusion criteria.

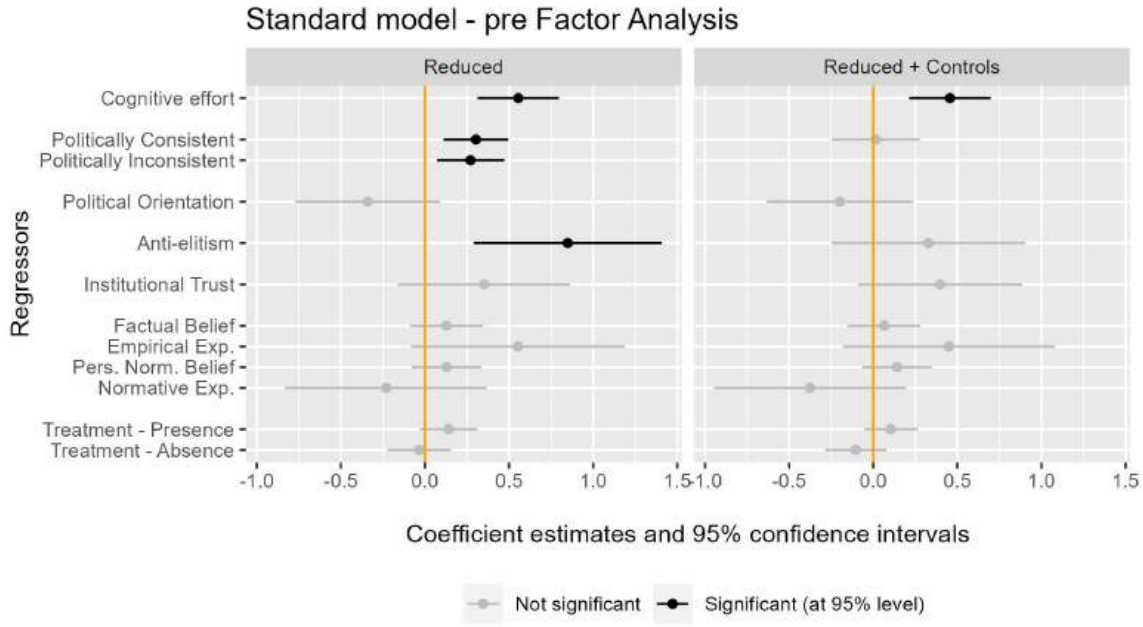


Figure 7.1: Reduced OLS regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

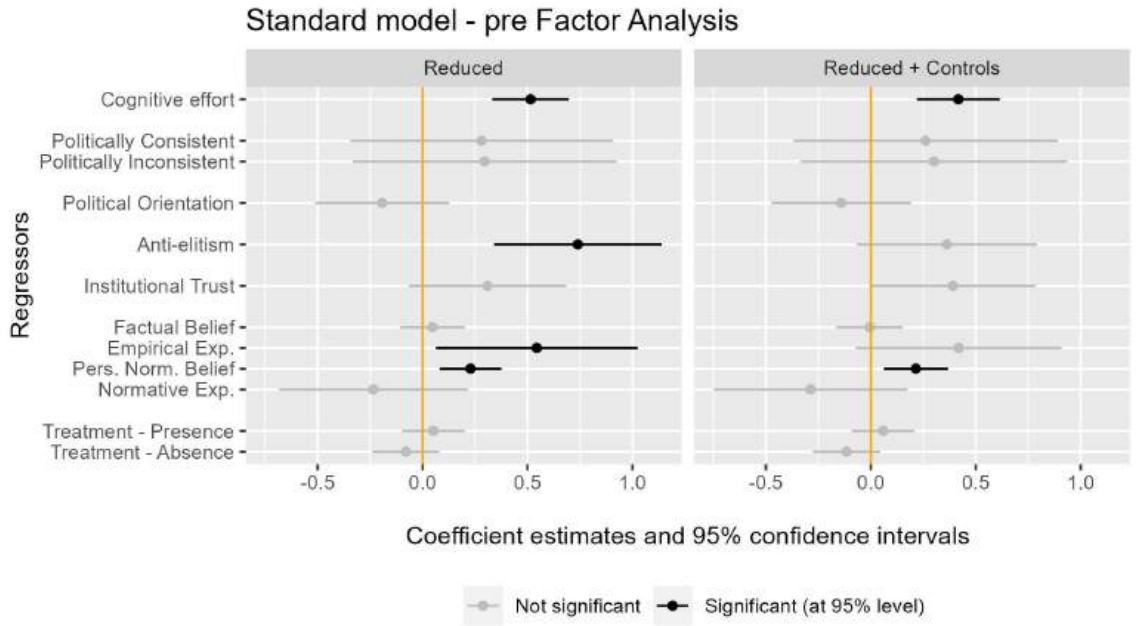


Figure 7.2: Reduced MLM regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

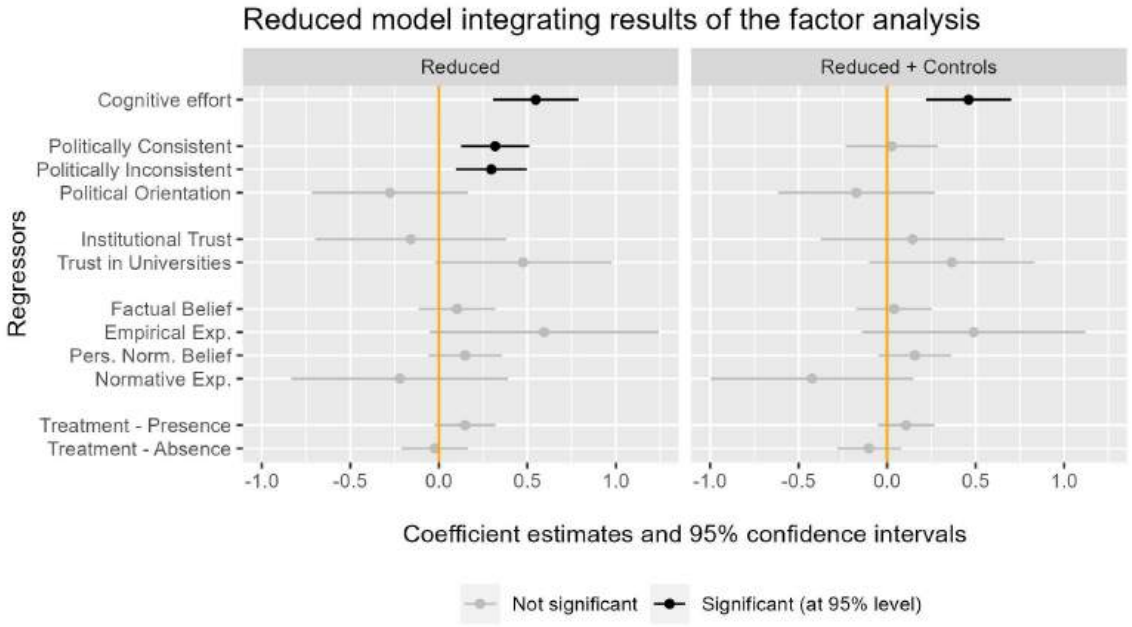


Figure 7.3: OLS regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

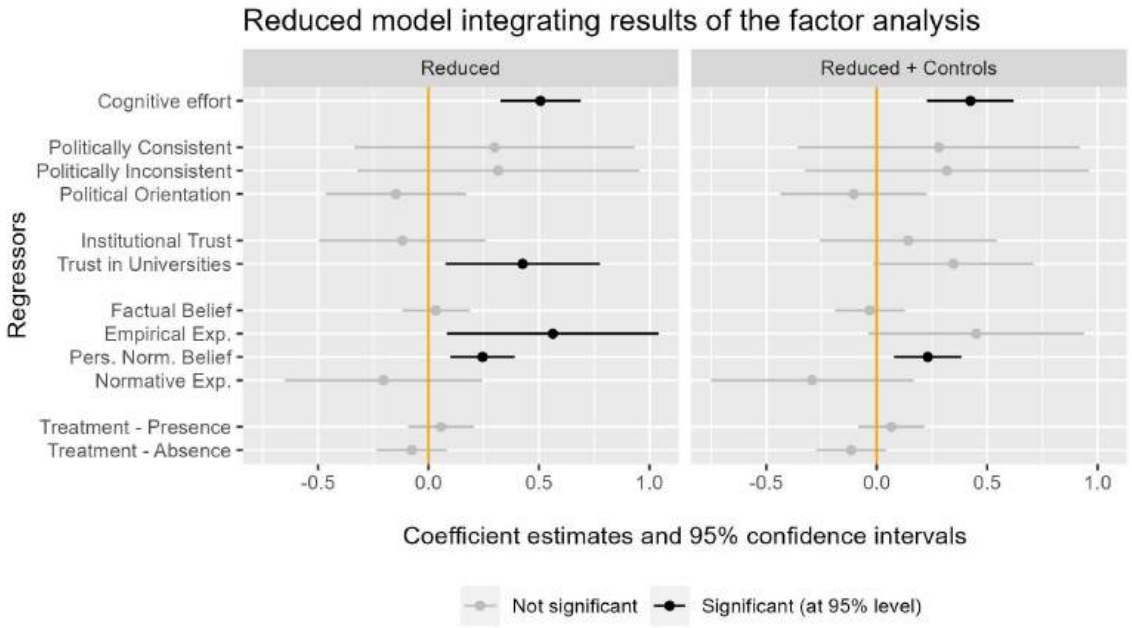


Figure 7.4: MLM regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

In addition, the two coefficients are significant in OLS *Reduced* (β = 0.303, p = 0.002 and β = 0.271, p = 0.008 respectively) and *Indexes* models (β = 0.318, p = 0.002 and β = 0.296, p = 0.008 respectively) (see Table 7.1). However, as I said for *Convenience-ITA*, no OLS model can compute the two coefficients when including items' fixed effects.

Instead, estimates from multilevel regression models succeed in computing both coefficients even when employing random intercepts at the level of posts. Nevertheless, these estimates are not statistically significant (*p-value* ranging from 0.332 to 0.417, see Table A.2).

Bivariate regression models focused on the variable *Dichotomous-Pol.Cons.*, which forces political neutrality into one of the two levels, demonstrate significantly lower truth discernment when judging consistent posts compared to inconsistent ones (β = -0.278, p = 0.006 in the formulation with items FE). However, when employing the three-level version of the variable (including neutrality, denoted as *Pol. Consistency*) as depicted in Figure 7.1 and 7.2, no significant differences are observed in the judgment of consistent posts as compared to inconsistent ones. In fact, the confidence interval of the estimate for the effect of political consistency (β = 0.141, [-0.047, 0.328]) overlaps with that of political inconsistency (β = 0.425, [0.227, 0.623]) (results from bivariate regression model without FE).

In conclusion, *Main-ITA* fails to support Hypothesis 2 as political consistency's effect is rarely significant. In addition, when the effect is significant, the association is positive, contrary the expected direction.

7.3 Anti-elitism

The three items measuring anti-elitism (*Talk-too-much*, *Difference-with-people*, and *Don't-care-about-people*) are firstly analysed separately in bivariate regression models. *Talk-too-much* (β = 0.416, p = 0.025) and *Difference-with-people* (β = 0.558, $p < 0.001$) show a significant and positive association with truth discernment. Instead, the association for *Don't-care-about-people* is signifi-

cant only at an *alpha* level of 0.1 ($p = 0.081$). Furthermore, *Don't-care-about-people* is significant in bivariate multilevel models ($\beta = 0.186$, $p = 0.036$ in the formulation with random intercepts for participants and items).

This overview is diametrically different from the one pictured in *Convenience-ITA*, where the only significant coefficient was the last one. However, the two studies share the direction of these links. Items measuring anti-elitism are *positively* associated with truth discernment.

The index derived from aggregating anti-elitism questions as a simple average demonstrates significance in both OLS ($\beta = 0.849$, $p = 0.003$) and MLM ($\beta = 0.740$, $p < 0.001$) multivariate models without controls. However, the coefficients are not significant when introducing controls ($p = 0.262$ and $p = 0.098$ respectively). Contrary to expectations, the direction of the relationship indicates that participants with higher anti-elitism scores exhibit better judgment of news. In conclusion, data from *Main-ITA* does not support Hypothesis 3a as anti-elitism fails to show a significant link to truth discernment, and when it does, it is not in the expected direction.

A factor analysis conducted on all the numerical variables revealed that *Don't-care-about-people* correlates more strongly with variables measuring institutional trust than with other indicators of anti-elitism. Consequently, *Don't-care-about-people* was aggregated with the three variables assessing trust in Parliament, the Government and the press. This decision was made, among other reasons, to ensure consistency of analysis with *Convenience-ITA* and *Main-USA*. The resulting index replaced the initial institutional trust index in multivariate regression models presented in Figure 7.3 and 7.4. The significance of this index in explaining participants' truth discernment will be discussed in the subsequent paragraph.

7.4 Institutional Trust

Bivariate descriptive analysis and bivariate regression models reveal a nuanced scenario underlying the relationship between trust in various institutions and truth discernment. Broadly speaking, the

resulting picture closely resembles the one outlined in *Convenience-ITA*.

Trust in Parliament ($p = 0.114$ in model with items FE), the Government ($p = 0.949$) and the press ($p = 0.218$) are not significantly associated with truth discernment in bivariate regression models. On the contrary, trust in universities ($p = 0.005$) and social media ($p < 0.001$) demonstrate significant links. Nevertheless, these associations have opposite directions: individuals who trust universities demonstrate better news judgment ($\beta = 0.491$), while those who trust social media exhibit poorer discernment (β = -0.835).

The original index for institutional trust, calculated as a simple average of all variables for individual institutions, does not exhibit significance across any of the formulations. This is true in OLS multivariate models with ($p = 0.110$) and without ($p = 0.179$) controls (see Table 7.1). In addition, the same null result is replicated in multilevel models with (where, however, the estimate is close to the significant threshold: $p = 0.051$) and without ($p = 0.106$) controls (see Table A.2). The non-significant coefficients for the institutional trust index in multivariate regression models are always positive (β ranging from 0.309 to 0.398 points).

A potential explanation for this lack of significance is that the original formulation of the index aggregates variables with opposing effects (trust in social media and trust in universities). I built a new index to test this eventuality, combining results from factor analysis and bivariate analysis.

Factor analysis yields similar results to those of *Convenience-ITA*. Trust in Parliament, Government and the press are aggregated into a new index, along with *Don't-care-about-people*. Trust in universities was included as a separate variable in multivariate models, given its relation with truth discernment and its weaker correlation with trust in other institutions. Instead, trust in social media is excluded from the multivariate analysis. This decision is based on trust in social media's collinearity with other variables and its distinct relationship with the dependent variable, compared to other institutional trust variables (negative instead of null).

Multivariate regression models incorporating the new index for institutional trust, along with

trust in universities as a separate variable, are presented in Figure 7.3 and 7.4. Institutional trust is confirmed to have a null relationship with truth discernment ($p = 0.563$ in OLS model without controls, $p = 0.586$ in OLS model with controls, $p = 0.537$ in MLM model without controls, $p = 0.486$ in MLM model with controls). In contrast, trust in universities exhibits a positive association with it. However, the significance of this association varies according to the model's formulation. It is significant in MLM model without controls ($\beta = 0.426$, $p = 0.017$), mildly significant in the OLS model without controls ($\beta = 0.476$, $p = 0.061$) and in the MLM model with controls ($\beta = 0.347$, $p = 0.062$), and not-significant in the OLS model with controls ($\beta = 0.366$, $p = 0.124$).

In summary, the relationship between trust in institutions and truth discernment varies depending on the institution. In most cases, this relationship is non-existent, and trust in social media demonstrates a negative influence, while the link with trust in universities is positive and occasionally significant. In conclusion, data from *Main-ITA* do not support Hypothesis 3b, except in limited cases.

7.5 Social Norm on the Importance of Accuracy

This section is dedicated to the results regarding the presence of social norms about sharing accurate news and its potential link with truth discernment. The structure will be the same as that followed in Chapter 6.

7.5.1 Does an "Accurate Sharing" Social Norm Exist?

The presence of beliefs and expectations regarding the "accurate sharing" social norm in *Main-ITA* is generally similar to the one already outlined for *Convenience-ITA* (Figure 7.5). Accuracy emerges as an important factor determining participants' stated sharing behaviour, with 35.6% of the sample indicating it as their first priority when sharing news. Additionally, almost 43.1% of participants expressed that accuracy *should* be the primary consideration for everyone.

Similarly to what was said for *Convenience-ITA*, the relevance of the news remains another

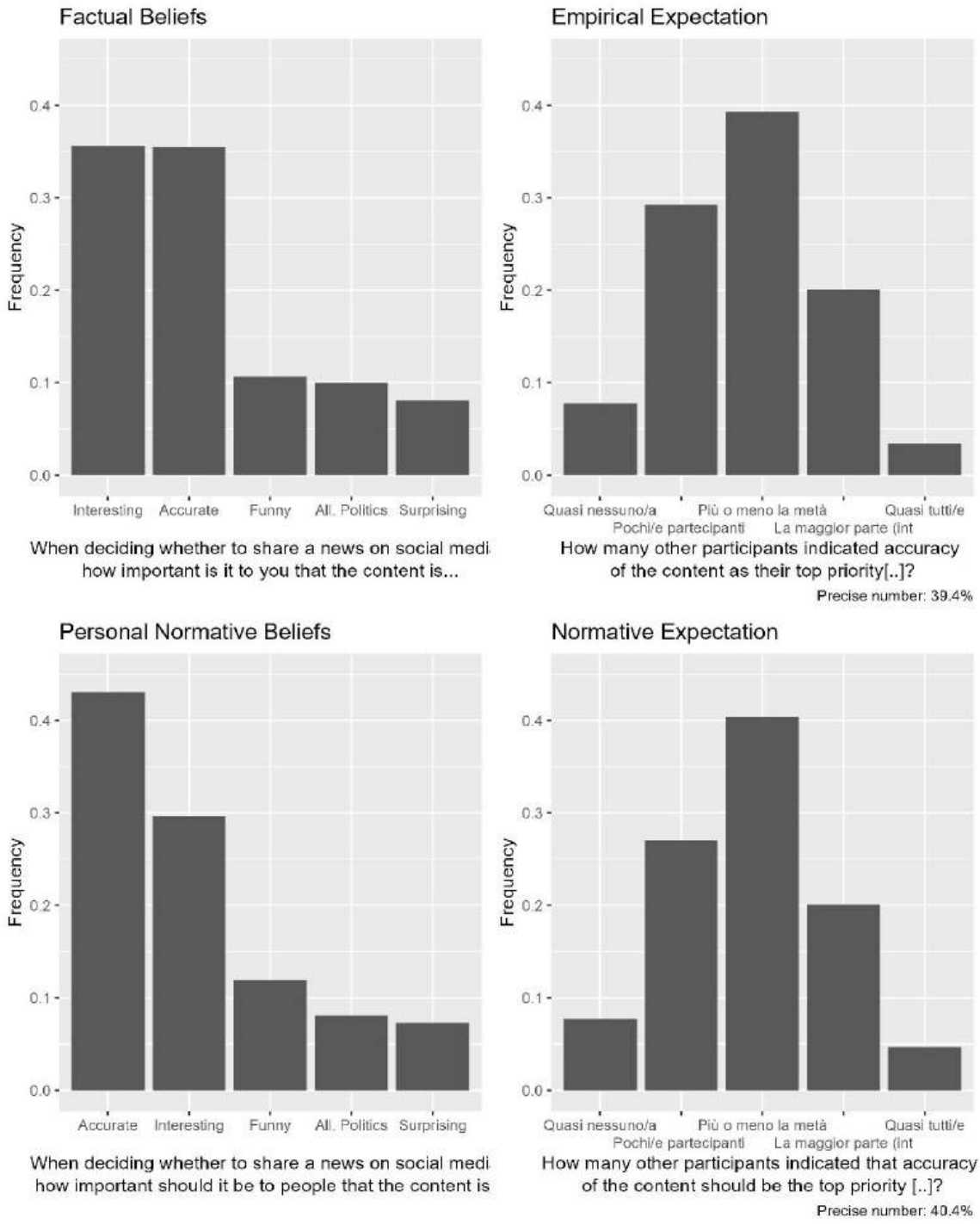


Figure 7.5: Descriptive univariate analysis of the presence of beliefs and expectations regarding the presence of an "accurate sharing" social norm.

crucial factor for participants, with 35.6% of the sample prioritising it in the factual beliefs question. This is the same percentage as participants prioritising accuracy. Another 29.7% of participants report relevance as the motivation that *should* be prioritized the most, in the question on personal normative beliefs.

The mixed situation of *Convenience-ITA*, where accuracy and relevance were the two most frequently chosen priorities, is replicated in *Main-ITA* and it is even more pronounced. In fact, while accuracy is the mode in the factual beliefs question of *Convenience-ITA*, participants of the Italian quota sample are evenly split between choosing accuracy and relevance when asked what their priorities are. Furthermore, when asked what people *should* prioritize, accuracy still stands out as the most important factor, albeit with a smaller gap with relevance (13 percentage points, compared to almost 20 pp in *Convenience-ITA*).

Regarding expectations, the distributions generally resemble those presented for *Convenience-ITA*. The mode indicates that participants believe around half of their peers prioritize accuracy in factual (39.4% of participants) and personal normative beliefs questions (40.4%), with a slightly right-skewed distribution. Interestingly, compared to *Convenience-ITA*, the skewness of both empirical and normative expectations is higher in *Main-ITA*. In particular, there are relatively more participants choosing the two answers of the scale (7.8% of participants choosing "Almost no one" in both questions, 27% and 29.3% of participants choosing "Few participants" in emp. and norm. expectation questions respectively) indicating that they believe that accuracy is not prioritised by others.

To summarise, while participants recognise the importance of accuracy in their peers' sharing decisions, most are very dubious that accuracy is prioritised by the entirety of the population, and some believe that, in fact, nobody is prioritising it. Interestingly, this picture roughly matches the actual factual and normative beliefs distribution. In other words, participants are generally correct in their (absence of) social expectations, given that their participants often do not prioritise

accuracy.

Overall, this ambiguity suggests an absence of a clear and established norm, which appears even more pronounced in *Main-ITA* compared to *Convenience-ITA*.

7.5.2 Correlational Evidence on the Relation with Truth Discernment

The analysis followed the same procedure used for other variables to explore the potential correlation between beliefs about an "accurate sharing" norm and truth discernment. Figures 7.1 to 7.4 present some of the results obtained.

In bivariate regression models, factual beliefs regarding the social norm are significantly and positively linked with truth discernment, as expected ($\beta = 0.251$, $p = 0.002$ in the model with items FE). The same is observed for personal normative beliefs as well ($\beta = 0.226$, $p = 0.003$). However, the significance of both coefficients is nullified in OLS multivariate models (p ranging from 0.250 to 0.716 for factual beliefs and from 0.135 to 0.227 for personal normative beliefs). Factual beliefs are also not significant in MLM multivariate models (p ranging from 0.548 to 0.951) On the contrary, personal normative beliefs are always significant and positively linked with truth discernment in this latter regression formulation (β ranging from 0.215 to 0.245, p ranging from 0.001 to 0.006).

To summarise, data from *Main-ITA* do not support Hypothesis 4a, while they support Hypothesis 4b, albeit with some null results.

Similar to the findings in *Convenience-ITA*, empirical expectations did not achieve significance in most models. This is true in bivariate regression models ($p = 0.166$ for the level "Almost all of them" in the model with items FE), as well as in OLS (p ranging from 0.072 to 0.162) multivariate models. The coefficient achieves significance in MLM multivariate models without controls ($p = 0.026$ in the *Reduced* model and $p = 0.021$ in the *Indexes* model). In these formulations, the direction is as expected. Participants who think that almost all their peers prioritise accuracy are better at judging posts than participants who think that almost none of them do so ($\beta = 0.544$ and

$\beta = 0.563$). However, when including controls, these coefficients are significant only at an *alpha* level of 0.1 ($p = 0.094$ and $p = 0.071$). To summarise, *Main-ITA* supports Hypothesis 5a only partially.

Coefficients for normative expectations, instead, never achieve significance. This is true in bivariate regression models ($p = 0.878$ for the level "Almost all of them" in the model with items FE), OLS multivariate models (p ranging from 0.146 to 0.478, Table 7.1) and MLM multivariate models (p ranging from 0.212 to 0.370, Table A.2). Consequently, *Main-ITA* fails to support Hypothesis 5b.

7.6 Treatment Effect

Main-ITA also tested the effect of a social norm information treatment highlighting the presence or absence of an "accurate sharing" social norm, as in the other studies, using a combination of two-sample t-tests, bivariate, and multivariate regression models.

Regarding the understanding of this message, results from the manipulation check indicate that 74% of participants responded correctly when asked to recall the content of the message.

A first descriptive analysis compares answers in the two groups before (first five rounds) and after the treatment (last five rounds). The treatment increased the truth discernment of participants in the *Presence* group from 0.837 [0.719, 0.954] to 0.924 [0.807, 1.040]. Thus, the increase is not strong enough to significantly increase participants' truth discernment, as the confidence intervals overlap. The change is even smaller in the *Absence* group, where truth discernment decreases from 0.802 [0.683, 0.921] to 0.786 [0.666, 0.907].

Another analysis divides the computation of truth discernment across the ten rounds of the accuracy task to observe its trend before and after the treatment and the re-treatment (Figure 7.6). Results indicated that truth discernment, calculated as the average difference of scores given to reliable posts minus those given to fake news, remained statistically constant before and after

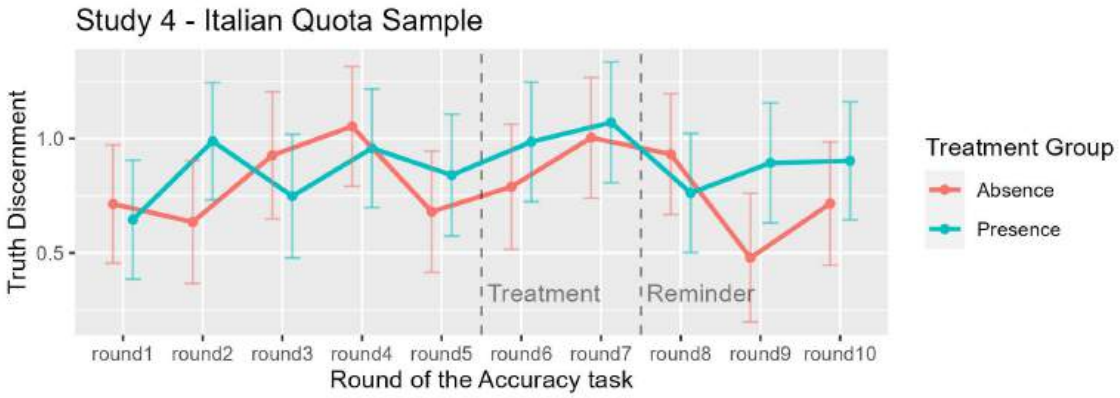


Figure 7.6: Truth Discernment in the Treatment Groups (with 95% CI).

the two treatments. The same is also true for the manipulation check, which also acted as a re-treatment. Specifically, participants in the treatment group exposed to the *Presence* message show an oscillating trend around the average, without much changes after the two interventions. A similar trend is visible for the *Absence* group, although with a drop in rounds nine and ten.

The lack of significance of the treatment effect was consistent across all implemented regression models. In all cases, the direction of the effect was as hypothesized for both groups (for example, β = 0.115, p = 0.122 for the *Presence* group and β = -0.019, p = 0.807 for the *Absence* group in the bivariate model with items FE). However, the significance did not reach the 95% threshold in any of the bivariate or multivariate regression models (p ranging from 0.094 to 0.817). Consequently, *Main-ITA* fails to support Hypothesis 6a and 6b.

Table 7.1: OLS Regression Models - Main-ITA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.555	0.457	0.547	0.460
	0.124 (<0.001) ***	0.123 (<0.001) ***	0.123 (<0.001) ***	0.122 (<0.001) ***
Pol. Consistency	0.303	0.016	0.318	0.026
	0.098 (0.002) **	0.134 (0.908)	0.098 (0.001) **	0.133 (0.843)
Pol. Inconsistency	0.271		0.296	
	0.102 (0.008) **		0.102 (0.004) **	
Pol. Orientation	−0.341	−0.198	−0.276	−0.175
	0.220 (0.120)	0.222 (0.373)	0.225 (0.220)	0.226 (0.439)
Anti-elitism	0.849	0.329		
	0.285 (0.003) **	0.293 (0.262)		
Inst. Trust (original)	0.351	0.398		
	0.261 (0.179)	0.249 (0.110)		
Fact. Bel. - Accurate	0.128	0.066	0.103	0.040
	0.111 (0.250)	0.110 (0.551)	0.111 (0.352)	0.109 (0.716)
Emp. Exp. - Pochi/e partecipanti (...	0.350	0.181	0.342	0.167
	0.223 (0.117)	0.204 (0.373)	0.225 (0.128)	0.202 (0.408)
Emp. Exp. - Più o meno la metà	0.083	0.129	0.094	0.144
	0.227 (0.715)	0.211 (0.540)	0.229 (0.680)	0.209 (0.492)
Emp. Exp. - La maggior parte (into...	0.143	0.014	0.097	−0.033
	0.244 (0.558)	0.222 (0.951)	0.247 (0.694)	0.220 (0.882)
Emp. Exp. - Quasi tutti/e	0.552	0.449	0.595	0.488
	0.324 (0.089) +	0.322 (0.162)	0.330 (0.072) +	0.322 (0.129)
Pers. Norm. Bel. - Accurate	0.128	0.142	0.148	0.157
	0.106 (0.227)	0.105 (0.177)	0.106 (0.166)	0.105 (0.135)
Norm. Exp. - Pochi/e partecipanti (...	−0.014	−0.254	0.019	−0.247
	0.240 (0.953)	0.212 (0.230)	0.238 (0.937)	0.209 (0.238)
Norm. Exp. - Più o meno la metà	−0.023	−0.333	−0.010	−0.332
	0.244 (0.925)	0.218 (0.127)	0.244 (0.966)	0.216 (0.123)
Norm. Exp. - La maggior parte (into...	−0.156	−0.303	−0.119	−0.291
	0.253 (0.537)	0.230 (0.188)	0.253 (0.637)	0.228 (0.201)
Norm. Exp. - Quasi tutti/e	−0.231	−0.377	−0.221	−0.424
	0.306 (0.451)	0.291 (0.196)	0.312 (0.478)	0.292 (0.146)
Treatment - Presence	0.141	0.105	0.148	0.107
	0.088 (0.110)	0.082 (0.202)	0.088 (0.094) +	0.082 (0.190)
Treatment - Absence	−0.035	−0.104	−0.022	−0.102
	0.096 (0.717)	0.093 (0.266)	0.096 (0.817)	0.093 (0.271)
Trust in Universities			0.476	0.366
			0.254 (0.061) +	0.238 (0.124)
Inst.Trust (factor analysis)			−0.159	0.144
			0.275 (0.563)	0.265 (0.586)
Num.Obs.	7470	7369	7511	7410
AIC	42 737.3	40 773.2	42 997.4	40 978.4
BIC	94 032.1	90 344.5	94 616.7	90 859.0

Chapter 8

Results of the US Quota Sample

Main-USA is the final empirical component of this dissertation, serving as the third main study employed to test its primary hypotheses. As mentioned, it has the same structure of *Convenience-ITA* and *Main-ITA*, but it was conducted with a quota sample of approximately 1200 US internet users[1]. Like in the case of *Main-ITA*, answers were gathered through Qualtrics until quotas for age, gender and education were met. Despite efforts to fulfil all quotas, this was not possible in the case of education.

Therefore, the sample exhibits an over-representation of individuals with a bachelor's degree or higher qualifications, leading to an under-representation of those who completed their education at the high school level. Despite a slight under-representation of people aged 55 years or older, the sample closely mirrors the distribution of the US general population in terms of age and gender.

Regarding political views representation, the sample slightly over-represents individuals who self-identify as Democratic-leaning and independent voters.

This chapter's structure, analytical approach, and visualizations align with the examples provided in Chapter 6 and 7. For a detailed explanation of the plots and analysis methodology, please refer to Chapter 4.

[1]The median duration, akin to the other three main studies, was around 15 minutes.

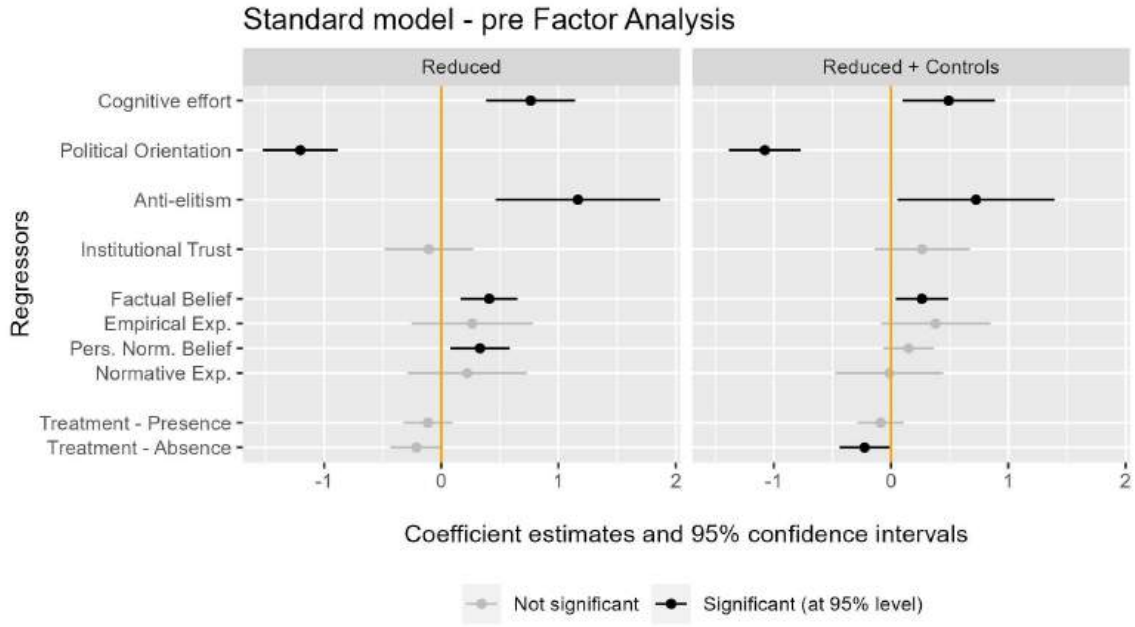


Figure 8.1: Reduced OLS regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

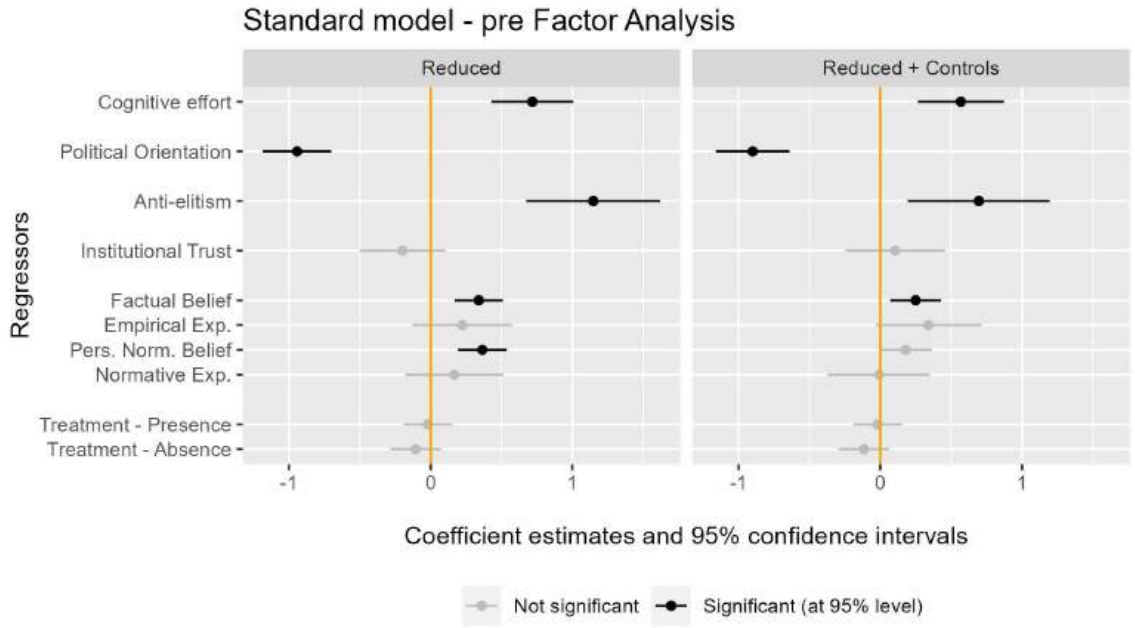


Figure 8.2: Reduced MLM regression model, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

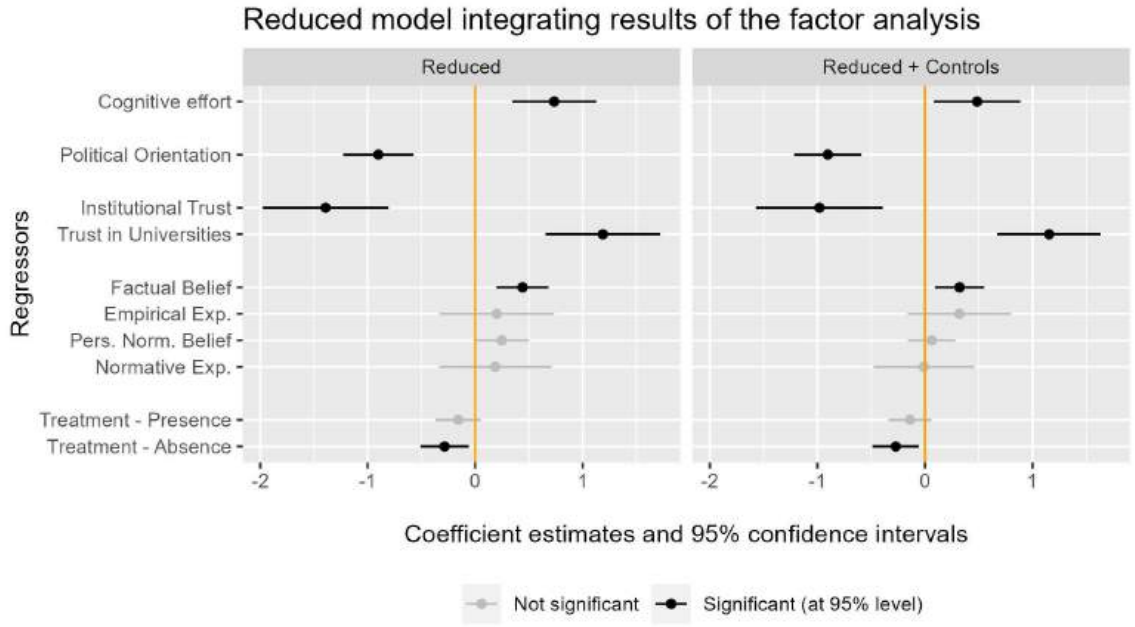


Figure 8.3: OLS regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

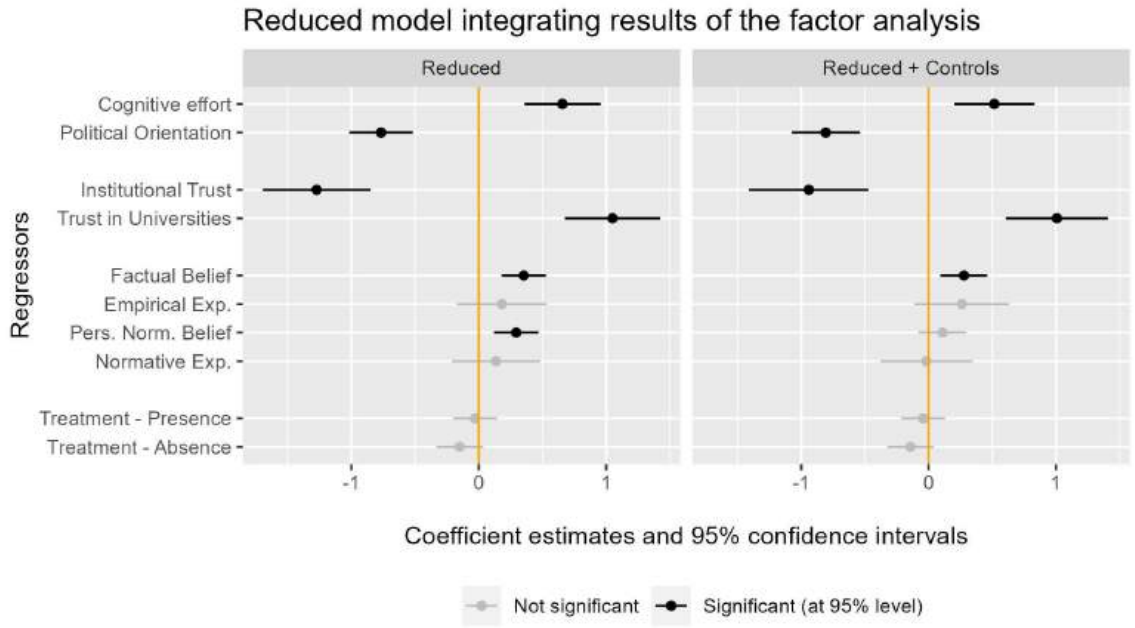


Figure 8.4: MLM regression model integrating the results of the factor analysis, with and without controls. Predicted variable: perceived accuracy. Presented coefficients refer to the interaction between various explanatory variables and truth discernment.

8.1 Cognitive Styles

Participants with a more analytical cognitive style are found to have better truth discernment in *Main-USA* (Figure 8.1 to 8.4, Table 8.1 and A.3), consistent with the findings in the other two main studies. This conclusion is supported by all analyses conducted.

The coefficient for CRT score is significantly associated with a higher truth discernment in bivariate regression models ($\beta = 1.284$, $p < 0.001$). The same is true for OLS multivariate models with ($\beta = 0.492$, $p = 0.014$ in *Reduced* and $\beta = 0.485$, $p = 0.018$ in *Indexes*) and without ($\beta =$ 0.762 and 0.736, $p < 0.001$ in both cases) controls, as well as in MLM multivariate models with (β $= 0.568$, $p < 0.001$ in *Reduced* and $\beta = 0.516$, $p < 0.001$ in *Indexes*) and without ($\beta = 0.718$ and 0.718 respectively, $p < 0.001$ in both cases) controls. Consequently, *Main-ITA* finds full support for Hypothesis 1.

8.2 Motivated Reasoning

In *Main-USA*, the exploration of motivated reasoning is hindered by the fact that all collected reliable posts are perceived as politically neutral. Statistically speaking, the perceived partisanship scores of reliable posts in the US dataset, as judged by participants in *Validation-USA*, always intersect neutrality within their confidence intervals (see Figure 5.8). This means that it is impossible to categorize them as either Democratic- or Republican-leaning. Consequently, reliable posts cannot be labelled as either politically consistent or inconsistent with participants' ideologies.

The absence of "partisan" reliable posts makes it impossible to compute the effect of political consistency on the perceived accuracy of reliable posts, as these are neither consistent nor inconsistent, but just neutral. Consequently, the effect of political consistency on truth discernment is also impossible to calculate, as this latter variable is calculated as the difference in perceived accuracy of reliable posts minus that of fake news posts.

At most, the effect of political consistency can be tested on the perceived accuracy of fake news

for which partisan posts exist. However, this would measure how participants perceive fake news differently, not their ability to distinguish it from reliable posts.

As a potential workaround for this limitation, I computed an alternative variable called *Dichotomous - Pol.Cons.*, as I mentioned in other studies. This variable forces political neutrality into one of the two levels of political consistency, considering only the estimated perceived partisanship, but not its significance. However, this variable never demonstrates significance in bivariate regression models ($p = 0.923$ in model with items FE).

In other words, even when categorizing statistically neutral reliable posts as either Democratic- or Republican-leaning, these models indicate that participants are equally adept at judging politically consistent posts as they are at judging inconsistent ones. Consequently, even when employing an expanded operationalization of motivated reasoning to facilitate its calculation, *Main-USA* fails to support Hypothesis 2.

8.3 Anti-elitism

Generally speaking, the results for anti-elitism in *Main-USA* mirror those found in *Convenience-ITA* and *Main-ITA*, albeit often with a greater number of significant findings (Figure 8.1 and 8.2, Table 8.1 and A.3).

When analysed separately in bivariate regression models *Talk-too-much* ($\beta = 0.957$, $p < 0.001$ in the model with items FE) and *Don't-care-about-people* ($\beta = 1.140$, $p < 0.001$) show significant and positive associations with truth discernment. Instead, the coefficient for *Difference-with-people* is not statistically significant ($p = 0.363$).

Similarly, the index encompassing the three anti-elitism items consistently exhibits significance and a positive association with higher truth discernment in multivariate models. Anti-elitist participants are better at discerning news in OLS multivariate models with ($\beta = 0.724$, $p = 0.034$) and without ($\beta = 1.165$, $p = 0.001$) controls. The same result is replicated in MLM models with ($\beta =$

0.695, $p = 0.007$) and without controls ($\beta = 1.148$, $p < 0.001$).

Contrary to what was hypothesized, participants who perceive the elite as corrupt demonstrate better judgment of news posts. Thus, Hypothesis 3a is not supported, as anti-elitism exhibits a significant, but positive correlation with truth discernment.

Similar to the approach adopted for the other two main studies, *Don't-care-about-people* was integrated into a new index with variables assessing trust in institutions, resulting from both PCA and PFA factor analysis. The resulting index replaced the original institutional trust index in a revised formulation of the pre-registered models. The outcomes of these models are delineated in Figure 8.3 and 8.4, and their implications are discussed in the subsequent section.

8.4 Institutional Trust

Compared to *Convenience-ITA* and *Main-ITA*, bivariate analysis in *Main-USA* finds confirming evidence on the role of trust in universities ($\beta = 0.355$, $p = 0.036$) and trust in social media (β = -1.197, $p < 0.001$). In addition, coefficients for trust in Parliament (β = -0.781, $p < 0.001$), the Government (β = -0.588, $p < 0.001$) and the press (β = -0.588, $p = 0.006$) are also significant and present a negative link with truth discernment.

However, institutional trust as measured by an index aggregating all individual institutions, exhibits no significant association with truth discernment. This is true in OLS models with (p = 0.201) and without ($p = 0.580$) controls, as well as in MLM models with ($p = 0.557$) and without ($p = 0.201$) controls.

Similarly to what was noticed in *Main-ITA*, this absence of significance could be explained with this index incorporating variables with contrasting effects. Consequently, the five variables were regrouped in a new index resulting from PCA and PFA factor analyses. The resulting index includes trust in Parliament, the Government, and the press, as well as the anti-elitism item *Don't-care-about-people*. Trust in universities was included as a separate variable, while trust in social

media was excluded for similar reasons to the ones explained in Chapters 6 and 7. Figure 8.3 and 8.4 present the outcomes of multivariate regression models where the original institutional trust index has been replaced with its factor analysis-derived counterpart.

Consistent with *Main-ITA*, trust in universities exhibits a significant and positive coefficient in these new multivariate formulations (β ranging from 1.008 and 1.156, p always below 0.001). On the contrary, the revised index for trust in "political" institutions, previously non-significant in *Main-ITA*, demonstrates a significant and negative association with truth discernment across all formulations (β ranging from -1.392 and -0.982, p always below 0.001).

In summary, *Main-USA* underscores the significance of institutional trust as a determinant of US respondents' news discernment abilities. However, the nature of this relation varies depending on the specific institutions under consideration. While participants who place trust in institutions typically demonstrate poorer discernment of news accuracy, the contrary holds true for those who exhibit trust in universities, showing better performance. Consequently, *Main-USA* finds support for Hypothesis 3b only in the case of trust in universities.

8.5 Social Norm on the Importance of Accuracy

In this section, akin to the approach taken in *Convenience-ITA* and *Main-ITA*, I will begin by presenting evidence regarding the presence of beliefs and expectations concerning the importance of accuracy in news sharing. The second part of the section will be dedicated to correlations between these beliefs and expectations and truth discernment.

8.5.1 Does an "Accurate Sharing" Social Norm Exist?

Accuracy emerges, by a large margin, as the factor which is chosen most frequently as the first priority by respondents. This holds true for the question concerning factual beliefs and personal normative beliefs. Approximately 55.68% of the sample indicates that accuracy is their first priority when deciding whether to share a piece of news on social media, while another 60.13% indicates

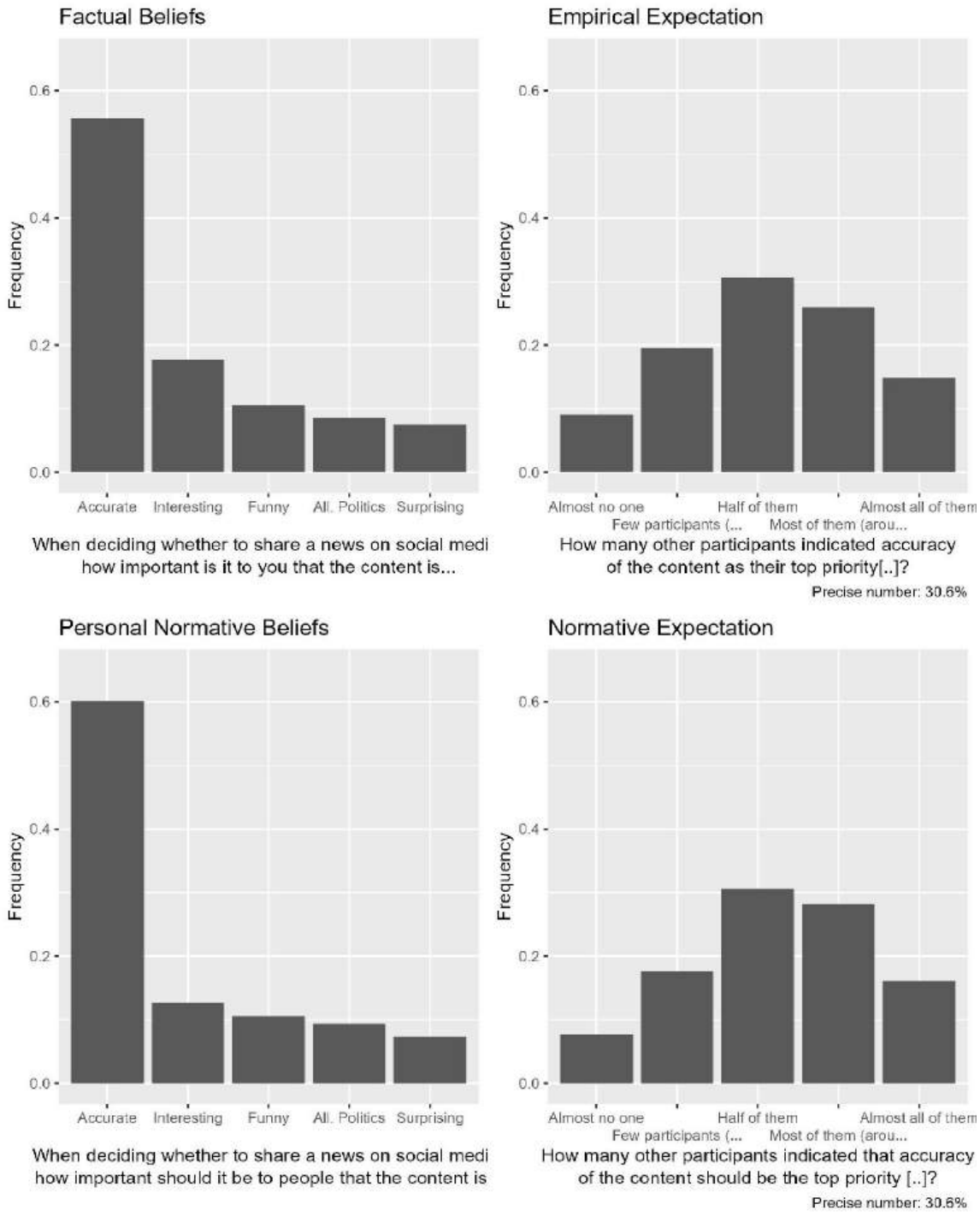


Figure 8.5: Descriptive univariate analysis of the presence of beliefs and expectations regarding the presence of an "accurate sharing" social norm.

that this should be the case for everyone (Figure 8.5).

In contrast to *Convenience-ITA* and *Main-ITA*, a strong preference for accuracy is firmly established among US respondents. Relevance of the news remains the second most selected factor, but with significantly lower percentages of choices compared to the Italian studies. Only around 17.77% of the sample prioritises "interesting" in the factual beliefs question, which further decreases to 12.68% when addressing personal normative beliefs.

Regarding expectations, participants are also far more prone to perceive that their peers prioritise accuracy than in the Italian studies. Similarly to other studies, the mode of both empirical (30.6%) and normative (30.57%) expectations questions indicates that participants believe half of their peers prioritize accuracy. However, the percentage of participants who believe that most (25.91% and 28.19%) or almost all (14.86% and 16.04%) of their peers are doing so is substantially higher in both questions, compared to the Italian studies (where, for example, this latter figure was around 5%). To summarise, the US sample presents a *left* skewed distribution. There are more participants with higher expectations (believing that most or all of their peers prioritize accuracy) than participants with low expectations (believing that few or almost none of their peers prioritize accuracy).

In summary, the potential social norm regarding the importance of accuracy appears relatively established in the US context, especially when compared to its perception by Italian participants. US respondents are much more inclined to declare that: 1. they prioritize accuracy in their sharing behaviour, 2. everybody should do so, 3. a large portion of their peers are, in fact, prioritizing accuracy, and 4. their peers attribute the same normative value to this prioritization.

Additionally, a final observation is that the gap between responses to empirical and normative questions is even smaller in *Main-USA* than what was observed in the Italian studies. Answers to the question about factual beliefs closely resemble those of its normative counterpart, personal normative beliefs. Similarly, the distribution of answers to the empirical expectations question

closely mirrors that of the normative expectation question.

8.5.2 Correlational Evidence on the Relation with Truth Discernment

The procedure followed to explore the correlation between the perception of an "accurate sharing" social norm, and truth discernment is consistent with that used for other variables and studies. Coefficient plots for the most important multivariate regression models are presented in Figure 8.1 to 8.4 and in Table 8.1 and A.3.

Factual beliefs regarding the social norm were found to be significant and positively linked with truth discernment across all model formulations. This result is consistent in bivariate regression models ($\beta = 0.093$, $p < 0.001$), in OLS multivariate models (β ranging from 0.263 to 0.441, p ranging from below 0.001 to 0.022) and MLM multivariate models (β ranging from 0.249 to 0.353, p ranging from below 0.001 to 0.007). Thus, *Main-USA* supports Hypothesis 4a.

This finding is partially consistent with that of personal normative beliefs. Coefficients for this variable are significant and positively linked with truth discernment in bivariate regression models ($\beta = 0.928$, $p < 0.001$), in the OLS *Reduced* model without controls ($\beta = 0.330$, $p = 0.011$), in the MLM *Reduced* ($\beta = 0.365$, $p < 0.001$) and *Indexes* ($\beta = 0.294$, $p = 0.001$) models without controls. In all other formulations, this coefficient is not significant. Consequently, Hypothesis 4b is only partially supported by *Main-USA*.

Variables measuring social expectations are found to be significantly linked with higher truth discernment in bivariate regression models. Participants who believe that their peers are prioritising accuracy are better at judging news ($\beta = 0.635$, $p = 0.002$). The same is true for those who expect all their peers to believe that accuracy should be prioritised by everyone ($\beta = 0.663$, $p = 0.002$). Nevertheless, these coefficients never reach significance in multivariate regression models (p ranging from 0.076 to 0.952). In all of these formulations, participants' performances in discerning fake news from reliable posts are independent of their perception of their peers' priorities in news sharing. Given this lack of consistency when using multivariate models, Hypothesis 5a and 5b are thus not

supported by *Main-USA*.

8.6 Treatment Effect

Like in the others, this study measured the treatment effect using a combination of t-tests, bivariate, and multivariate regression models.

Starting from understanding and recalling the treatment message, data from the manipulation check show that 57.64% correctly answered when asked to remember the intervention. This is a lower percentage than the one observed in the two Italian studies (around 75%).

Regarding the treatment message's effectiveness, participants in both groups recorded a decreased truth discernment after the treatment, although not significantly. Truth discernment for participants in the *Presence* group passed from 0.959 [0.816, 1.102] in the first five rounds, to 0.901 [0.752, 1.050] in the last five ones. Participants in the *Absence* group saw a decrease from 0.973 [0.827, 1.119] to 0.745 [0.594, 0.897] after the treatment.

The same non-significant result is replicated when dividing the analysis on the basis of singular rounds (Figure 8.6). The truth discernment of the two treatment groups did not change significantly before and after exposure to the information message, nor after the re-treatment, as the confidence intervals intersect. Although the group exposed to the *Absence* message saw a consistent decrease in its performances, as hypothesized, the same was registered for the other treatment group, albeit to a lesser extent. In both cases, the decrease was not marked enough to create a significant difference with the two control groups.

When testing the same hypothesis through regression models, the conclusions are generally the same, although some significant results can be found. The *Absence* treatment is significant in bivariate models (β = -0.218, p = 0.019). Furthermore, it is also significant in some OLS multivariate models (β ranging from -0.225 to -0.271, p ranging from 0.013 to 0.040). However, this coefficient is never significant in MLM multivariate models (p ranging from 0.108 to 0.245). On

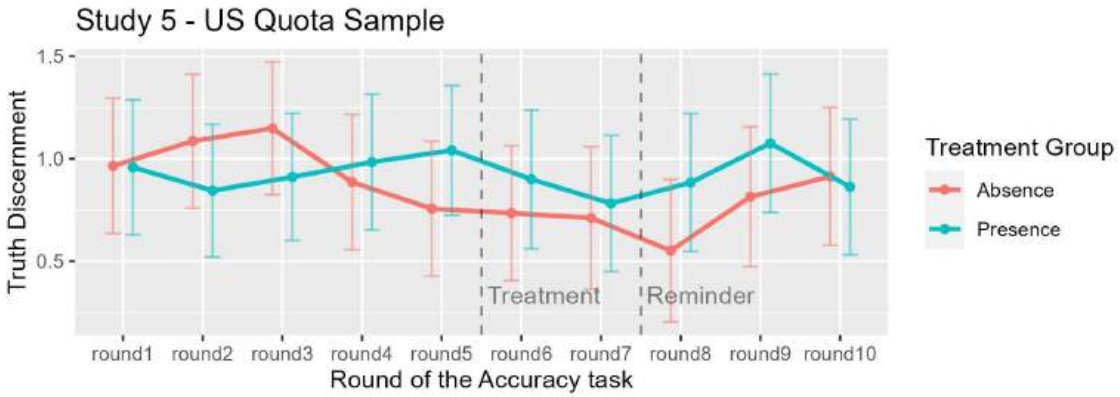


Figure 8.6: Truth Discernment in the Treatment Groups (with 95% CI).

the contrary, the coefficient for the *Presence* treatment effect is never significant (p ranging from 0.143 to 0.952). Consequently, Hypothesis 6a is not supported by *Main-USA*, while some evidence supports Hypothesis 6b, although not consistently.

Table 8.1: OLS Regression Models - Main-USA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.762	0.492	0.736	0.485
	0.193 (<0.001) ***	0.201 (0.014) *	0.199 (<0.001) ***	0.205 (0.018) *
Pol. Orientation	−1.202	−1.076	−0.901	−0.904
	0.164 (<0.001) ***	0.156 (<0.001) ***	0.167 (<0.001) ***	0.159 (<0.001) ***
Anti-elitism	1.165	0.724		
	0.358 (0.001) **	0.341 (0.034) *		
Inst. Trust (original)	−0.107	0.265		
	0.193 (0.580)	0.207 (0.201)		
Fact. Bel. - Accurate	0.408	0.263	0.441	0.323
	0.124 (<0.001) ***	0.115 (0.022) *	0.124 (<0.001) ***	0.116 (0.006) **
Emp. Exp. - Few participants (arou...	0.299	0.339	0.173	0.142
	0.241 (0.215)	0.219 (0.122)	0.247 (0.483)	0.235 (0.547)
Emp. Exp. - Half of them	0.035	0.113	−0.042	−0.051
	0.230 (0.880)	0.207 (0.586)	0.240 (0.860)	0.226 (0.821)
Emp. Exp. - Most of them (around t...	0.300	0.270	0.244	0.169
	0.240 (0.211)	0.220 (0.220)	0.248 (0.326)	0.233 (0.469)
Emp. Exp. - Almost all of them	0.264	0.380	0.201	0.320
	0.265 (0.319)	0.239 (0.112)	0.272 (0.460)	0.246 (0.193)
Pers. Norm. Bel. - Accurate	0.330	0.150	0.248	0.063
	0.129 (0.011) *	0.110 (0.173)	0.129 (0.056) +	0.113 (0.578)
Norm. Exp. - Few participants (arou...	0.039	−0.137	−0.001	−0.123
	0.231 (0.867)	0.210 (0.516)	0.238 (0.996)	0.218 (0.573)
Norm. Exp. - Half of them	0.092	−0.038	0.010	−0.040
	0.230 (0.688)	0.211 (0.856)	0.231 (0.966)	0.215 (0.852)
Norm. Exp. - Most of them (around t...	−0.039	−0.192	0.013	−0.085
	0.233 (0.867)	0.216 (0.373)	0.240 (0.958)	0.224 (0.705)
Norm. Exp. - Almost all of them	0.221	−0.014	0.187	−0.009
	0.260 (0.397)	0.234 (0.952)	0.267 (0.484)	0.239 (0.970)
Treatment - Presence	−0.114	−0.088	−0.158	−0.139
	0.106 (0.284)	0.101 (0.383)	0.108 (0.143)	0.102 (0.172)
Treatment - Absence	−0.213	−0.225	−0.286	−0.271
	0.115 (0.063) +	0.110 (0.040) *	0.115 (0.013) *	0.110 (0.014) *
Trust in Universities			1.187	1.156
			0.272 (<0.001) ***	0.245 (<0.001) ***
Inst.Trust (factor analysis)			−1.392	−0.982
			0.298 (<0.001) ***	0.302 (0.001) **
Num.Obs.	6206	6163	5894	5851
AIC	36 059.9	34 545.6	34 138.3	32 686.2
BIC	77 604.2	74 728.7	73 279.7	70 476.5

Chapter 9

Discussion

9.1 Retesting Cognitive Mechanisms

9.1.1 Confirmatory Evidences on the Role of Cognitive Styles

Across all studies and most model formulations, participants with a more analytical cognitive style consistently demonstrate a better ability to distinguish fake news from reliable posts. This confirms the central argument of a substantial body of literature, which, as outlined in Chapter 2, has already demonstrated the importance of analytical thinking (and its absence) in how individuals process information (e.g. Pennycook and Rand, 2019).

This project extends the validity of such findings to a relatively underexplored type of stimuli: real social media posts. The pivotal role of analytical thinking is demonstrated here outside the polished context of simple factual statements, and with a more realistic information format. In other words, analytical thinking remains a critical factor even, or perhaps even more so, when the interaction with the reader is tarnished by the many intervening features of real-life social media posts. This conclusion and the existing body of evidence further underscores the significance of cognitive styles in explaining misinformation beliefs.

An alternative explanation for these results is the potential role of attentiveness. The real driver for the strong result of CRT might be attentiveness, rather than cognitive styles. In other words, cognitive style might capture the tendency to be attentive or the ease to do so. Consequently, truth

discernment might be explained by simply the attention the participants put into filling out the survey. The CRT might have partially captured this attentiveness, other than just the cognitive style. Following this alternative explanation, if we were able to disentangle cognitive styles and attentiveness we might discover that the former has a decreased significance and that, instead, attentiveness is what explains participants' performances. This operation was not possible with the existing design but might be an interesting path for future research. Furthermore, it might help to explain variations of truth discernment for participants with the same scores in CRT tests.

9.1.2 The Marginal Impact of Left-Right Motivated Reasoning in Everyday News

A common assertion in previous literature on misinformation beliefs is that individuals are inclined to believe fake news because they confirm their preexisting beliefs, regardless of their veracity (see Chapter 2). However, this theory found little, if any, support in the present study.

Across all models and formulations, participants showed equal ability to discern politically consistent and inconsistent social media posts. The only instances where political consistency led to poorer performances were in specific operationalisations of news' perceived partisanship that completely overlooked their potential political neutrality.

If anything, participants were better at judging both consistent *and* inconsistent news than neutral ones. More intensely partisan fake news posts were more easily identified than neutral ones, regardless of the specific political position they favoured.

A possible explanation for the marginal role of motivated reasoning in the present study is that only a minority of the collected social media posts were perceived as partisan. Participants of the two Validation studies judged most of the stimuli as neutral (see Section 5.2). In particular, none of the US-based reliable posts were perceived as significantly favouring one faction or the other, preventing the computation of truth discernment for politically congruent posts.

The theoretical consequences of this potential interpretation depend on the reasons for this neu-

trality. One possibility is that the collected sample is representative of the information encountered by users in their everyday lives and that information in the real world is not as partisan as conventional wisdom would suggest. Even though most of the selected posts discussed news events and public issues, they were not selected to focus solely on political actors with a clear party affiliation. If this holds, it is possible that motivated reasoning plays a less central role in this study because it considers generic news, while the same mechanism becomes more crucial in other studies that focus on strictly political news[1].

Alternatively, the prevalent neutrality of the collected stimuli may be a result of a biased selection. It is possible that, in reality, misinformation is predominantly partisan but that the selected stimuli failed to mirror this situation. If this were the case, then the interpretation would be that this project failed to capture what could in reality still be a much relevant determinant of misinformation beliefs, i.e. motivated reasoning, because of an unrealistically neutral selection of news.

Knowing which option is true is a challenging empirical question, but two considerations slightly point toward the first one. First, stimuli were selected with a procedure designed to be as probabilistic and least arbitrary as possible, exactly to build a representative sample of circulating reliable and fake news (see Section 4.3). For example, I selected *all* the most recent fake news articles in debunking websites, by following minimal exclusion criteria. Secondly, there are some indications in the literature that the information usually consumed by online users is mostly non-political, and even less partisan (for example, see Guess, 2021).

Independently of these considerations, there are also reasons to think that the "neutrality explanation" does not fully account for motivated reasoning's lack of significance. Notably, motivated reasoning is not substantiated by the results even in the two Italian studies (*Convenience-ITA* and

[1]It is important to note that motivated reasoning, as traditionally understood within the left-right political continuum, may not fully capture other relevant frameworks such as globalism-localism or populism-elitism. This work did not study these frameworks, so conclusions about motivated reasoning outside the traditional left-right political spectrum should be cautiously approached.

Main-ITA), for which there are enough partisan posts to test its hypothesis. Even in these cases, participants were equally good at judging congruent and incongruent posts. The only instances where motivated reasoning is supported in the Italian studies is when using a restricted operationalization of political congruence, which omits posts' neutrality. However, this is not the case for *Main-USA*, where even the dichotomous version of political congruence fails to be significant.

Furthermore, motivated reasoning and the neutrality explanation do not explain why partisan posts are the easiest to judge, independently of them being congruent or not with participants' beliefs. Even if few, participants should be better at judging politically inconsistent and neutral posts, when compared to judging consistent posts. The worst truth discernment, however, is observed when judging neutral posts.

This latter result, that partisan posts are the easiest to judge, could be explained by considering another explanation. It is possible that participants utilized the perceived partisanship of posts, i.e. if they are more favourable to one side than the other, and the *intensity* of this partisanship, as cues to investigate their veracity. Impartiality is one of the fundamental qualities of good journalism, as can be seen also in the content of collected posts. More intensely partisan posts are more often fake news than reliable news. Participants may have judged these posts as less accurate because they perceived more overt political bias in them, and indeed, these posts were more frequently fake news.

Another possibility is that politically neutral posts were more ambiguous and less informative overall. Consequently, participants may have had fewer cues to investigate their veracity, leading to more frequent misjudgments.

9.2 Trust in "Trusted" Institutions is Positively Linked with Truth Discernment

A consistent result across all the studies, with some cross-cultural differences, is that truth discernment is positively linked to trust in universities and negatively linked to trust in social media,

even when controlling for cognitive styles and other mechanisms. Trust in Parliament, the Government and the press had null results in the Italian case and negative coefficients in the US sample. Interestingly, this pattern aligns closely with the average levels of trust in these institutions.

Universities are the most trusted institutions, while social media are the least trusted. Parliament, the Government, and the press received scores just below the mid-point of the scale. In summary, within this context, truth discernment is positively linked with trust in institutions with high levels of trust and negatively linked with trust in low-trusted institutions.

This result is unexpected, as all institutions were hypothesized to have the same link with truth discernment. In contrast, this relation seems to be following a meta-level mechanism. The link between institutional trust and truth discernment at the individual level has different directions according to institutional trust at the group level.

It is important to note that the existence of this meta-level mechanism is anything but definitive, even in the present context, given that its discussion is based on observational data and post-hoc analysis, outlined to make sense of unexpected results. However, once the due caution with which it should be taken is specified, I will now proceed with a theoretical discussion of how this meta-level observation can be explained, which could be useful for future research and the formulation of new hypothesis.

The first result is aligned with the pre-registered hypothesis. The positive link between trust in universities and truth discernment can thus be explained with the mechanisms outlined in Chapter 3, or at least some of them.

The first mechanism posited that participants with low trust in institutions and high anti-elitism were expected to believe in fake news because this confirmed their prior beliefs of a corrupted elite. This mechanism can not be easily applied to the results on the role of trust in universities, as trusting universities is not incompatible with believing that the government is corrupt.

On the contrary, it is possible that participants with low trust in universities are more prone

to selecting sources that do not rely on the scientific method to make conclusions about facts and causal links, thus ending up being exposed to fake news sources. Although marginally modified, this is what the second mechanism posited, which justified Hypotheses 3a and 3b.

Finally, it is possible that those participants who are more prone to believing fake news, independently of the reason for doing so, gradually developed low confidence in the scientific method, the scientific community that advances it, and universities as a consequence. This interpretation would support the "reverse causality" mechanism explained in Chapter 3.

All these explanations fail to account for the results regarding low-trusted institutions. Why do individuals who trust social media exhibit poorer truth discernment? Motivated reasoning cannot fully explain this result. First, fake news, per se, does not provide any image of social media. Thus, it is difficult to argue that participants believe in fake news because they confirm a positive image of social media, as it would be if motivated reasoning were applied.

Source selection, instead, could be a more apt explanation, but in a different way as outlined in Chapter 3). It could be that participants who trust social media are exactly those with a different media diet far from mainstream media. Consequently, they encounter more fake news and end up believing it.

Finally, reverse causality offers an even more plausible explanation. Individuals predisposed to believing fake news may have encountered, as the rest of the individuals, some of them in social media and, for various reasons, accepted them. Over time, this could reinforce their confidence in social media as a consistent source of, wrongly judged as accurate, fake news.

Specifically on trust in social media, a final explanation, not hypothesized in Chapter 3, could be that participants who trust social media are exposed to more fake news, as these are more prevalent here than in mainstream media.

In addition to explanations tailored specifically for trust in social media, it is worth considering a more over-arching explanation for the role of trust in distrusted institutions, i.e. institutions with

low levels of trust. Why do participants who have confidence in institutions not trusted by most peers exhibit limited truth discernment?

One possible explanation could be gullibility: individuals who trust social media, while most of their peers have little confidence in them, might be more susceptible to believing everything they encounter, including fake news. On the same line, Talwar et al. (2019) found that online trust negatively correlates with authenticating news before sharing. This suggests that individuals who trust social media, as those who trust online sources, may be less inclined to critically evaluate the information they encounter, leading to lower levels of truth discernment.

9.2.1 Considerations on Anti-elitism.

Among the three items used to measure anti-elitism, only one consistently correlates with truth discernment, although positively: *Don't-care-about-people* (*"Politicians care about what ordinary people think."*). The other two items show a similar positive relationship with truth discernment but with less consistent significance. In all cases, this goes against the hypothesized negative link between anti-elitism and truth discernment.

One potential explanation for this unexpected negative association is, once again, gullibility. It is plausible that participants with relatively lower levels of anti-elitism are more inclined to trust others more generally and have lower scepticism. This tendency to trust may result in a decreased ability to discern the accuracy of information, leading to the observed negative correlation with truth discernment.

Alternatively, it is possible that the selected item for measuring anti-elitism may not have effectively captured participants' perceptions of this concept. A first clue in this direction is the low variability in scores across all three items, with distributions exhibiting pronounced left-skewness. In each study, a significant portion of the sample displayed anti-elitist tendencies according to these measures, which can be understood in light of the low levels of trust in institutions recorded across the studies. However, this restricted variability may have reduced the statistical power necessary

to detect an effect, and it questions the validity of these items more generally. In alignment with this perspective, it is noteworthy that *Main-USA* demonstrates the most consistent significance in the anti-elitism index. Intriguingly, this study also exhibits the lowest skewness of scores across the questions related to anti-elitism. This observation suggests a potential association between the reliability of the anti-elitism index and the distribution variability of scores within the study.

The potential low reliability of the three anti-elitism items is understandable since they are taken from another literature (that on populism) and their intended use as sub-dimensions rather than as standalone measures of anti-elitism. However, despite this limitation, an intriguing pattern emerges regarding the conceptualization of anti-elitism and institutional trust across the three studies. Specifically, *Don't-care-about-people* consistently demonstrate a negative correlation with items measuring trust in institutions. In other words, the absence of institutional trust overlaps with the presence of negative expectations towards those same institutions (government in particular). Coming back to the discussion made in Chapter 3, this finding lends support to the notion of institutional distrust as encompassing the "absence of trust", and the fact that only one dimension (low-high institutional trust) is sufficient to capture this concept comprehensively.

9.3 Contextual Variability of the "Accurate Sharing" Social Norm and Its Disconnection from Truth Discernment

The findings concerning the potential existence "accurate sharing" social norm can be dissected into three primary findings:

- The presence of such a norm appears to be, like other norms, context-dependent, exhibiting variation between the two analysed countries.
- Social expectations regarding the importance of accuracy in news sharing are not found to be associated with truth discernment.
- Personal beliefs regarding the importance of accuracy in news sharing demonstrate a positive

association with truth discernment.

Beginning with results regarding the presence of an "accurate sharing" norm, the studies revealed a cross-cultural difference between Italy and the US. In the two Italian studies, participants evenly chose two factors as the most important, more or less with the same percentages: accuracy and relevance. In the US study, accuracy emerged as the predominant factor. This divergence is consistent across both questions assessing factual beliefs and personal normative beliefs.

Similarly, differences in social expectations were observed, albeit to a lesser extent across cultures. Italian respondents were less convinced of the prevalence of accuracy as a priority among their peers, as evidenced in both empirical and normative expectations questions. In contrast, US respondents were more inclined to be convinced of the prioritization of accuracy among their peers. Notably, participants in both contexts demonstrated an awareness of the levels of importance assigned to accuracy among their peers.

In terms of the association between this social norm and the dependent variable, social expectations are rarely found to be significantly associated with truth discernment. Moreover, causal evidence from experimental information messages designed to inform participants about their peers' social expectations similarly fails to yield significant effects.

Conversely, participants who assert the importance of accuracy in their decision-making regarding news sharing on social media (factual belief), and advocate for this prioritization universally (personal normative belief), consistently exhibit superior abilities in judging news posts. This finding holds true even after controlling for other mechanisms and, notably, cognitive styles. Hence, personal motivations emerge as a pivotal determinant in individuals' truth discernment. Other than being inclined towards analytical thinking and possessing the capacity to focus on content, individuals' *willingness* to ascertain the accuracy of information emerges as a crucial factor.

9.3.1 Explaining the Ineffectiveness of Social Expectations and the Treatment

Several plausible explanations emerge in dissecting the ineffectiveness of social expectations and the treatment. The most straightforward explanation for the absence of association between social expectations and truth discernment is that simply put, these variables are not related. In other words, it is possible that truth discernment is determined more by the cognitive effort we put in reading news and by the importance we give to accuracy rather than perceptions of others' beliefs and expectations.

An alternative explanation is that the "accurate sharing" social norm may indeed be pertinent for misinformation beliefs, but it has not yet taken firm root in the analysed contexts. Considering the relatively recent emergence of fake news as a prominent societal issue, it is plausible that a social norm emphasizing the importance of accuracy has not had sufficient time to permeate the population and significantly alter individuals' behaviour. Consequently, we may find ourselves in a transitional phase characterized by mixed perceptions and evolving social norms.

It is also possible that the items used to measure social expectations (the motivation ranking and the relative questions about others' choices) did not effectively capture this dimension, leading to potential misinterpretation by participants. These items were developed specifically for this study and were not subjected to validation through dedicated studies or qualitative investigations to ensure participants correctly understood the questions. While some evidence suggests that at least a portion of participants correctly interpreted the questions correctly, the lack of validation introduces a degree of uncertainty.

Regarding the effectiveness of the treatment, the most straightforward explanation of its null result is that it exclusively targeted social expectations. If these expectations are not inherently linked to truth discernment, then no discernible effect will be observed.

A second possibility is that the treatment may not have exerted a strong enough influence to alter

participants' behaviour, particularly because it targeted the wrong reference network (Bicchieri, 2006). The treatment messages informed participants about the expectations of their peers, defined as other American or Italian participants of similar survey sessions. However, it is conceivable that for the treatment to be effective, it should have addressed a reference network more closely tied to participants' everyday lives and deemed more significant by them. One potential approach could have been to reference friends, family, or the respondents' circle of social media followers. This would have informed individuals about the expectations of those with whom they regularly interact, making the reference network more direct and potentially impactful. However, conveying information about the expectations of participants' relatives may not have been believable, and it would have been deceptive unless supported by real data. In addition, obtaining such data would have been much more challenging and costly.

It is also possible that respondents' fatigue might have diluted the effectiveness of the treatment. The treatment was exposed to participants nearly at the end of the survey experiment. This was after questions regarding all the measured variables and after five iterations of the accuracy task. It is possible that, at that point in the questionnaire, participants were too tired to change their ability as a consequence of the message. A shorter questionnaire or putting the message at the beginning of the accuracy task might increase its efficacy. However, this was not possible with the available resources without sacrificing other crucial design features. Additionally, descriptive evidence shows that participants' truth discernment did not decrease significantly in the later rounds of the accuracy tasks. Consequently, while a tiring effect is still possible, it is unlikely to have substantially influenced participants' performances.

An interesting point on this potential "lack of strength" in the treatment message is the observation that, despite the lack of statistical significance, the "absence" message appears to have a negative impact in contexts where the norm seems more established (*Main-USA*), while the "presence" message has a positive impact in contexts where the norm seems less established (*Conv.-ITA*

and *Main-ITA*). This suggests that the effectiveness of the treatment may be contingent upon its ability to challenge established perceptions. It is possible that a message highlighting the presence of a norm would have only a marginal effect in a context where the norm is already perceived as prevalent. Similarly, a message highlighting the absence of a norm would have minimal impact in a context where people do not perceive it. If this holds, neither of the two contexts may be sufficiently established to allow the treatments to alter individuals' mixed perceptions significantly.

9.4 General Considerations

After discussing the primary findings concerning the main mechanisms explored in this project, I will now consider its results more broadly. In particular, I will outline potential explanations for the unexpected lack of significance in some hypotheses, examining how they relate to certain methodological characteristics of the studies.

9.4.1 Difficulty of the Chosen Stimuli

As detailed in Methodology (Chapter 4), one significant contribution of this work is the utilization of real social media posts, as opposed to de-contextualized statements and article screenshots commonly employed in similar studies. This decision enhances the external validity of the findings, as the stimuli closely resemble, often in unaltered form, what individuals encounter while navigating social media. On the other hand, a potential unintended consequence is that of having "more difficult" stimuli, i.e., items which are more difficult to judge.

In this sample of posts, reliable and fake news were collected using procedures that do not ensure, by themselves, that fake news is *distinguishable* from reliable posts. The collected posts were validated through two Validation studies, and, on average, reliable posts received higher scores across all measured dimensions of perceived accuracy and quality. Nevertheless, there are instances where the perceived accuracy of the fake news is statistically indistinguishable from the perceived accuracy of the reliable post focused on the same news event (i.e. which contains the

same keywords). Additionally, in few cases, the perceived accuracy of the fake news posts is, on average, slightly higher than their reliable counterpart.

It is conceivable that the collected posts were so challenging to differentiate that between-individual differences hardly emerged. While this might have limited the significance of the results, it is also true that selecting items to facilitate their discernment would have meant selecting survey materials to modify and ultimately facilitate the confirmation of the project's hypothesis.

Another aspect of the utilized materials is that participants were unable to interact with the posts' screenshots to gather more relevant information, as they could in real interactions by clicking on links, browsing the web, or sending posts to peers. This allowed for the isolated study of explored mechanisms without introducing other interactions. However, it may have also heightened the difficulty of judgment even more, compared to real interactions. Overall, the potential difficulty of the collected stimuli is closely linked to another survey feature, which will be discussed next: the inclusion of the "I don't know" option in the perceived accuracy questions.

9.4.2 Inclusion of "I Don't Know" Option

As anticipated in Chapter 4, many papers in the misinformation beliefs literature measure perceived accuracy by asking participants to judge stimuli without providing them with the option to express their uncertainty (e.g. Arechar et al., 2023). In these studies, there is no "I don't know" (DK) option for participants to indicate that they are unsure about the accuracy of what they are reading. Recent evidence (Luskin et al., 2018) shows that this approach may inflate measured misinformation beliefs, as participants who are unsure may be more likely to provide incorrect answers when forced to respond. Consequently, these responses may be recorded as if participants believed fake news when, in reality, they were aware that they did not know the answer.

To compensate for this shortcoming, the surveys in this work always included a DK option in questions measuring perceived accuracy. However, this decision reduced the number of responses available for inclusion in regression models, as the dependent variable was coded as a numerical

variable and "I don't know" answers were thus coded as missing cases.

Beyond purely numerical considerations, it is also important to note that the incidence of these "I don't know" responses likely occurred more frequently in items where participants were more dubious and where differences could emerge more starkly if respondents were pushed to guess an answer by the exclusion of the DK option. Although including the DK option enhances methodological rigour, it could explain why the present work found fewer significant results compared to related works.

9.4.3 Erroneous Exclusion of Forza Italia Voters in *Main-ITA*

While the preceding considerations of the present subsection all started from conscious methodological choices made in an attempt to increase the robustness of this work, another factor, this time implemented unintentionally, could explain some non-significant results in *Main-ITA*.

As outlined in Methodology (Chapter 4), *Main-ITA* contained an error in the coding of the party affiliation question, omitting "Forza Italia", one of the five main Italian political parties. Further examination through robustness checks indicates that Forza Italia voters likely did not spill over to other parties, but instead chose not to respond to this question. Consequently, they were excluded from the sample by Qualtrics as incomplete observations.

Although none of the studies aimed to represent the general population on political variables, the selective exclusion of Forza Italia voters could have influenced the results of *Main-ITA*. For instance, if the truth discernment of these voters were determined by structurally different mechanisms, compared to other ideologically far or close participants, this exclusion might have skewed the findings. While this scenario seems unlikely, it remains a possibility worth acknowledging.

Chapter 10

Conclusion

10.1 Contributions

In Chapter 2, the limited use of real and validated stimuli was identified as a significant gap in the existing literature on misinformation beliefs. This work addressed this gap by examining previously explored mechanisms on a wide collection of real social media posts, containing both reliable and fake news and validated through two Validation Studies.

This project successfully demonstrated the significance of cognitive style in influencing participants' truth discernment, even within the realm of real social media posts. Furthermore, the significance of this role persisted when compared to other explanatory variables included in the analysis, increasing the robustness of cognitive styles, particularly the importance of analytical thinking, in understanding individuals' misinformation beliefs.

The use of an extensive range of news topics, coupled with validated perceived partisanship, on the other hand, allowed us to discover how motivated reasoning can emerge as a marginal explanation of misinformation beliefs. While the relevance of motivated reasoning remains out of question when focusing on news mentioning political actors, this study revealed its diminished relevance in explaining misinformation beliefs concerning general news topics found on social media posts.

For what regards the role of anti-elitism and institutional trust, the present project revealed

nuanced and unexplored relationships between trust in individual institutions and truth discernment. Across *Convenience-ITA*, *Main-ITA* and *Main-USA*, evidence shows that this relation varies depending on the specific institution taken into account and the overall level of trust in it. Participants demonstrating trust in institutions with high levels of public trust, such as universities, exhibited better abilities in discerning news. Conversely, those who exhibited trust in "distrusted" institutions, such as social media, demonstrated poorer performances. Remarkably, this pattern persisted across both analysed countries and diverse sample types.

In addition to examining institutional trust, the present work also analysed the understudied area of social norms surrounding online news sharing. The three main studies allowed, first of all, to *measure* beliefs and expectations regarding the importance of accuracy in online news sharing, something that did not receive the same attention and depth of analysis in past research. The findings underscore that the predominance of accuracy in news sharing is not yet firmly established, neither in individuals' personal beliefs nor in their expectations of their peers' priorities.

Moreover, a notable contribution to the literature is the discovery of cross-country variations in the importance individuals attribute to accuracy and their expectations of others in this regard. Findings reveal that respondents in two studies conducted in Italy exhibit more heterogeneous preferences in their motivations and are less inclined to believe their peers prioritise accuracy when compared to respondents from the US, where the predominance of accuracy seems to be more established.

A final contribution of the present study is that of consistently proving the role of news-sharing personal motivations in determining participants' truth discernment. Past research already gathered evidence on the fact that individuals assigning higher importance to accuracy are better at judging news (Arechar et al., 2023). This project expands that evidence by measuring the importance of accuracy by using rankings instead of Likert scales. Furthermore, other than measuring what participants declare to prioritize, the present work also measured participants' beliefs about what

should be prioritized, finding significant results for this variable as well. The role of both factual beliefs and personal normative beliefs regarding the importance of accuracy is found to be significant even when controlling for the propensity to adopt a more analytical cognitive style.

10.2 Limitations

In addition to manipulating the potential "accurate sharing" social norm, this project aimed to collect correlational evidence regarding the role of various explanatory variables. The objective was to establish initial evidence of the association between these variables and truth discernment, supported by theoretically grounded explanations. In other words, while hypotheses were formulated regarding the mechanisms underlying these links, i.e. why and how a variable influences or is influenced by another, these mechanisms were not empirically explored.

The correlation between trust in certain institutions and truth discernment, consistently observed across the present studies, was not subject to causal testing. The project lacked components that would have allowed exploration into why the link exists between these variables. In other words, a limitation of the present project is that it was not designed to test the mechanisms hypothesized to justify the expected (and unexpected) correlations between those variables.

The same limitation applies to other variables consistently found to be significantly correlated with truth discernment, such as factual and personal normative beliefs regarding the importance of accuracy. The present project does not provide an explanation for *why* participants who prioritize accuracy in news sharing are better at judging news. Suggestions for addressing this gap in future research are presented in the following section (Section 10.3).

Regarding the measurement and manipulation of social expectations about the accurate sharing social norm, the present project always referred to "other participants" as the chosen reference network (Bicchieri, 2006). This decision was primarily driven by practical considerations. Given that the surveys included an experimental message informing participants about these social expec-

tations, the chosen reference network needed to be empirically measurable to ensure the credibility and non-deceptiveness of the message. "Other participants" emerged as the most feasible reference for measurement. However, it is conceivable that other reference networks, such as family, friends, or online followers, may be more relevant to individuals when determining how to behave online.

The present work attempted to use a selection of stimuli which replicate, as closely as possible, the contents encountered by participants in their everyday social media use. In many features, those posts were presented to participants as they are seen on social media. However, the interaction between readers and posts was constrained by a series of lacking features, that, on the contrary, are commonly found on social media platforms. For instance, the survey did not allow participants to click on posts, read comments, browse the internet for additional information, or simply engage in discussions with online and offline peers[1]. Therefore, the generalization of conclusions drawn from this study is limited by the absence of these features. For example, individuals with an intuitive cognitive style but a higher propensity to seek help or engage with peers may have the possibility, in real life, to compensate for their lack of analytical thinking by leveraging their social abilities. However, such compensatory mechanisms were not possible within the confines of the present studies.

10.3 Future Research

This thesis will conclude by suggesting several directions for future research based on the pieces of evidence gathered during this project. One key observation is that the relationship between trust in institutions and truth discernment appears to be contingent upon the overall level of trust. Future research could systematically investigate this observation, which was not pre-registered, to determine its consistency and elucidate the workings of this potential meta-level mechanism. Addi-

[1]It is possible that some participants resorted to some of these resources anyway, as their activity was not monitored as it could have been, for example, in a lab experiment or during in presence survey data collection. However, it is plausible that most of them completed the survey without resorting to external resources to judge posts, as the answers to the accuracy task were not incentivized.

tionally, further research could explore other variables that may interact with trust in institutions to influence truth discernment, thus providing a more comprehensive understanding of the underlying mechanisms at play.

Again on the role of anti-elitism and institutional trust, one hypothesized mechanism of their link with truth discernment was motivated reasoning. This could be studied by measuring, similarly to what was done with perceived partisanship in the two Validation studies, the extent to which social media posts are perceived to talk about institutions and elite, what specific actors they mention and with what sentiment or alignment. Such an approach would allow us to explore motivated reasoning in the context of institutional trust and anti-elitism. Another possible path in explaining this relationship would be to explore how participants' source selection changes according to their levels of trust in different institutions, perhaps by combining survey with trace data (as done, for example, in Guess, 2015).

Regarding the measurement and manipulation of social expectations, a potential limitation of the present work, as mentioned above, is the choice of "other participants" as the main reference network. Future research might explore what reference networks, if any, individuals value when deciding how to behave online. Qualitative investigation might be, in this case, a particularly suitable approach, especially for an initial investigation of the matter, given that it would allow participants to express themselves outside pre-designed blocks of answers actively.

Furthermore, findings from the present work indicate that the "accurate sharing" norm is not established, and it might be in a more initial and flexible stage. Future research might focus on attesting whether this is, in fact, a loose norm (Dimant et al., 2023). This might help explain the observed variability in beliefs and expectations across the three main studies. In addition, it might also indicate that a more local and situated reference network is directing participants' decisions, given that the norm is not uniformly adopted across the entire population.

The exploration of *why* and *how* two variables are correlated could be expanded also to the role

of personal motivations for news sharing. While the present work gathered some initial correlational evidence on this relationship, future works could test the mechanism behind it. For example, it is possible that personal motivations mediate the *amount* of cognitive effort actively put in the judgment of posts. While cognitive styles measure the propensity of individuals to engage in this type of effort, and their ease in doing so, personal motivations might regulate the *willingness* of engaging in this type of effort. Future studies might measure cognitive effort *during* the evaluation of social media posts (or other types of stimuli) by integrating more sophisticated methodologies, like tracking and timing of clicks and mouse movement, eye-tracking, or even facial emotion recognition, in the case of lab experiments.

Regarding the role of the "accurate sharing" norm more broadly, another possible path of exploration regards the role of modalities of news sharing (see Section 3.5). The present work limited its exploration to news sharing broadly speaking. Future research could explore if individuals' beliefs and expectations regarding news-sharing motivations change depending on the context of each specific interaction and on the meaning with which information is shared (e.g. when it is done satirically, to inform others, or to foster a discussion).

As mentioned in the previous paragraph, a limitation of the present work is that participants are asked to judge social media posts without being allowed to resort to many features which, on the contrary, are present on social media. A possible direction for further research is that of the retesting mechanism here analyzed in a context which allows this kind of enhanced interaction in order to increase the external validity of these conclusions even more. In addition, allowing participants to interact more freely with what they are reading and with other online and offline peers could bring out new and unexpected mechanisms and behavioural patterns.

Besides methodological directions and suggestions on what future works might study, some theoretical approaches might also be worth exploring, especially those regarding the role of social norms.

Motivations for news sharing were studied here as they are all alternatives in the same continuum, as in previous research. Participants were asked to judge those factors using the same criteria. However, it is possible for some motivations to be complementary rather than alternatives. Instead of choosing between relevance and accuracy, it is possible that individuals evaluate what to share by considering both of these factors, but as responding to different needs.

For example, it is possible that, for some individuals, posts need, first of all, to be accurate to be shared. At the same time, it might be that, among accurate posts, individuals are more prone to share relevant news. In this approach, accuracy and relevance answer to the different needs of individuals in a framework which is similar to Herzberg's motivation-hygiene theory. Accuracy, or better said inaccuracy, might be a factor of hygiene, which can cause dissatisfaction in the potential sharer, but which does not represent, per se, a reason for the content to be shared. Relevance, on the other hand, might be a motivation factor, which can cause satisfaction in the sharer and justify an active action.

Future research might design studies which allow the exploration of sharing decisions within this framework, which might help to explain the results presented here, where accuracy and relevance are consistently chosen as the two most important factors, especially in Italian studies, where they share similar percentages of choices.

Another alternative theoretical approach to sharing motivations and the presence of a potential social norm might be the distinction between prescribing and proscribing social norms, which was little explored in the present work. It might be that accuracy and relevance are subject to two different and competing norms. A prescribing norm of relevance might be requiring individuals to share content which is as interesting as possible, in order to receive positive feedback from others. On the other hand, a proscribing norm of accuracy might deter them from sharing inaccurate content, as this might result in unwanted negative sanctions. The present project did not allow respondents to express themselves on such a distinction. Future research might adopt this approach

and develop studies that allow more in-depth exploration of this facet.

Bibliography

Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, *21*, 1–12.

Akkerman, A., Mudde, C., & Zaslove, A. (2014). How Populist Are the People? Measuring Populist Attitudes in Voters. *Comparative Political Studies*, *47*(9), 1324–1353. https://doi.org/10.1177/0010414013512600

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media and Society*, *9*(1). https://doi.org/10.1177/20563051221150412

Altay, S., Hacquin, A. S., & Mercier, H. (2022). Why do so few people share fake news? It hurts their reputation. *New Media and Society*, *24*(6), 1303–1324. https://doi.org/10.1177/1461444820969893

Álvarez-Benjumea, A., & Winter, F. (2020). The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(37), 22800–22804. https://doi.org/10.1073/pnas.2007977117

Amaya, A. (2020). Adapting how we ask about the gender of our survey respondents. *Pew Research Center*.

Andı, S., & Akesson, J. (2020). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, *9*(1), 106–125.

Andrighetto, G., Szekely, A., Guido, A., Gelfand, M., Abernathy, J., Arikan, G., Aycan, Z., Bankar, S., Barrera, D., Basnight-Brown, D., Belaus, A., Berezina, E., Blumen, S., Boski, P., Bui, H. T. T., Cárdenas, J. C., Čekrlija, de Barra, M., de Zoysa, P., ... Eriksson, K. (2024). Changes in social norms during the early stages of the COVID-19 pandemic across 43 countries. *Nature Communications*, *15*(1), 1436. https://doi.org/10.1038/s41467-024-44999-5

Angelucci, C., & Prat, A. (2023). Is journalistic truth dead? measuring how informed voters are about political news. *Measuring How Informed Voters Are about Political News (April 8th, 2023)*.

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and Reducing Online Misinformation Across 16 Countries on Six Continents. *Nature Human Behaviour*, *7*(9), 1502–1513.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729.supp

Bakir, V., & McStay, A. (2018). Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Balcetis, E., & Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of personality and social psychology*, *91*(4), 612.

Baptista, J. P., & Gradim, A. (2022). Online disinformation on Facebook: the spread of fake news during the Portuguese 2019 election. *Journal of Contemporary European Studies*, *30*(2), 297–312. https://doi.org/10.1080/14782804.2020.1843415

Bauer, P. C., & Clemm von Hohenberg, B. (2021). Believing and Sharing Information by Fake Sources: An Experiment. *Political Communication*, *38*(6), 647–671. https://doi.org/10.1080/10584609.2020.1840462

Bicchieri, C. (2006). *The Grammar of Society - The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C., Lindemans, J. W., & Jiang, T. (2014). A structured approach to a diagnostic of collective practices. *Frontiers in Psychology*, *5*(DEC), 1–13. https://doi.org/10.3389/fpsyg.2014.01418

Bigley, G. A., & Pearce, J. L. (1998). Straining for Shared Meaning in Organization Science : Problems of Trust and Distrust. *The Academy of Management Review*, *23*(3), 405–421.

Bloch, M. (1921). *Réflexions d'un historien sur les fausses nouvelles de la guerre / Marc Bloch*.

Bosmajian, H. A. (2006). *Burning books*. McFarland.

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(5), 2–4. https://doi.org/10.1073/pnas.2020043118

Brashier, N. M., & Rand, D. G. (2021). *llusory Truth Occurs Even with Incentives for Accuracy*.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Brown, É. (2018). Propaganda, Misinformation, and the Epistemic Value of Democracy. *Critical Review*, *30*(3-4), 194–218. https://doi.org/10.1080/08913811.2018.1575007

Castanho Silva, B., Jungkunz, S., Helbling, M., & Littvay, L. (2020). An Empirical Comparison of Seven Populist Attitudes Scales. *Political Research Quarterly*, *73*(2), 409–424. https://doi.org/10.1177/1065912919833176

Chadwick, A., Hall, N. A., & Vaccari, C. (2023). Misinformation rules!? Could "group rules" reduce misinformation in online personal messaging? *New Media and Society*. https://doi.org/10.1177/14614448231172964

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media and Society*, *20*(11), 4255–4274. https://doi.org/10.1177/1461444818769689

Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. Guilford Press.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

de Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

de Rooij, E. A., Steculła, D. A., & Pickup, M. A. (2022). Populist media diets. *Social Science Quarterly*, *103*(4), 975–991. https://doi.org/10.1111/ssqu.13178

DeVerna, M. R., Guess, A. M., Berinsky, A. J., Tucker, J. A., & Jost, J. T. (2024). Rumors in retweet: Ideological asymmetry in the failure to correct misinformation. *Personality and Social Psychology Bulletin*, *50*(1), 3–17.

Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, *1*(1). https://doi.org/10.37016/mr-2020-001

Dimant, E., Gelfand, M., Hochleitner, A., & Sonderegger, S. (2023). Strategic Behavior with Tight, Loose and Polarized Norms. *SSRN Electronic Journal*, (10233). https://doi.org/10.2139/ssrn.4340092

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Egelhofer, J. L., Aaldering, L., Eberl, J. M., Galyga, S., & Lecheler, S. (2020). From Novelty to Normalization? How Journalists Use the Term "Fake News" in their Reporting. *Journalism Studies*, *21*(10), 1323–1343. https://doi.org/10.1080/1461670X.2020.1745667

Elster, J. (2007). *Explaining Social Behavior - More Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Enders, A. M., & Uscinski, J. E. (2021). Are misinformation, antiscientific claims, and conspiracy theories for political extremists? *Group Processes and Intergroup Relations*, *24*(4), 583–605. https://doi.org/10.1177/1368430220960805

Epstein, J., Santo, R. M., & Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology*, *68*(4), 435–441. https://doi.org/10.1016/j.jclinepi.2014.11.021

ESS. (2021). ESS-9 2018 Documentation Report. Edition 3.1.

EVS. (2020). European Values Study 2017: Integrated Dataset (EVS 2017). https://doi.org/10.4232/1.13560

Faragó, L., Kende, A., & Krekó, P. (2020). We only Believe in News That We Doctored Ourselves: The Connection between Partisanship and Political Fake News. *Social Psychology*, *51*(2), 77–90. https://doi.org/10.1027/1864-9335/a000391

Fawzi, N. (2019). Untrustworthy News and the Media as "Enemy of the People?" How a Populist Worldview Shapes Recipients' Attitudes toward the Media. *International Journal of Press/Politics*, *24*(2), 146–164. https://doi.org/10.1177/1940161218811981

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems*, *38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Goyanes, M., & Lavin, A. (2018). The Sociology of Fake News. Factors affecting the probability of sharing political fake news online. *Media and Communications*, 16. https://researchportal.uc3m.es/display/act503117

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Greene, C. M., Nash, R. A., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory*, *29*(5), 587–604.

Guess, A. M. (2015). Measure for measure: An experimental test of online political media exposure. *Political Analysis*, *23*(1), 59–75. https://doi.org/10.1093/pan/mpu010

Guess, A. M. (2021). (Almost) everything in moderation: new evidence on Americans' online media diets. *American Journal of Political Science*, *65*(4), 1007–1022.

Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Asian-Australasian Journal of Animal Sciences*, *32*(2), 1–9. https://doi.org/10.1126/sciadv.aau4586

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry (United Kingdom)*, *62*(9-10), 1033–1065. https://doi.org/10.1080/0020174X.2018.1508363

Hadiz, V. R., & Chryssogelos, A. (2017). Populism in world politics: A comparative cross-regional perspective. *International Political Science Review*, *38*(4), 399–411. https://doi.org/10.1177/0192512117693908

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., & Puranen, B. (2022). World Values Survey: Round Seven. https://doi.org/10.14281/18241.16

Hakhverdian, A., & Mayne, Q. (2012). Institutional trust, education, and corruption: A micro-macro interactive approach. *Journal of Politics*, *74*(3), 739–750. https://doi.org/10.1017/S0022381612000412

Halpern, D., Valenzuela, S., Katz, J., & Miranda, J. P. (2019). From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News. *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21.* https://doi.org/10.1007/978-3-030-21902-4_16

Hameleers, M. (2022). Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information Communication and Society*, *25*(1), 110–126. https://doi.org/10.1080/1369118X.2020.1764603

Hameleers, M., Bos, L., & de Vreese, C. H. (2017). The Appeal of Media Populism: The Media Preferences of Citizens with Populist Attitudes. *Mass Communication and Society*, *20*(4), 481–504. https://doi.org/10.1080/15205436.2017.1291817

Hameleers, M., Brosius, A., & de Vreese, C. H. (2022). Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European journal of communication*, *37*(3), 237–268.

Hofstede, G., Hofstede, G. J., & Minkov, M. (1991). *Cultures and Organizations*. http://books.google.de/books?id=wFW0AYqIM0AC

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hudson, J. (2006). Institutional trust and subjective well-being across the EU. *Kyklos*, *59*(1), 43–62. https://doi.org/10.1111/j.1467-6435.2006.00319.x

Ignazi, P. (2017). Sartori's party system typology and the Italian case: The unanticipated outcome of a polarised pluralism without anti-system parties. *Contemporary Italian Politics*, *9*(3), 262–276. https://doi.org/10.1080/23248823.2017.1389122

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *176*(1), 5–51. https://doi.org/10.1111/j.1467-985X.2012.01032.x

Iosifidis, P., & Nicoli, N. (2020). The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation. *International Communication Gazette*, *82*(1), 60–81. https://doi.org/10.1177/1748048519880729

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, *65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Kahan, D. M. (2010). Fixing the communications failure. *Nature*, *463*(7279), 296–297. https://doi.org/10.1038/463296a

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424. http://journal.sjdm.org/13/13313/jdm13313.html

Kahan, D. M. (2017). *Misconceptions, Misinformation, and the Logic of Identity-protective Cognition*. https://doi.org/10.1016/j.nut.2015.02.008

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174. https://doi.org/10.1080/13669877.2010.511246

Kahneman, D. (2003). Kahneman2003. *the American Economic Review*, *93*(05), 1449–1474.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Kohlberg, L. (1969). *Stage and sequence; The cognitive-developmental approach to socialization*.

Kopp, C., Korb, K. B., & Mills, B. I. (2018). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to "fake news". *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0207383

Kozyreva, A., Lorenz-Spreen, P., Herzog, S., Ecker, U., Lewandowsky, S., & Hertwig, R. (2022). *Toolbox of Interventions Against Online Misinformation and Manipulation*. https://psyarxiv.com/x8ejt/

Krastev, I. (2007). Is East-Central Europe Backsliding? The strange death of the liberal consensus. *Journal of Democracy*, *18*(4), 56–63.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

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

Legros, S., & Cislaghi, B. (2020). Mapping the Social-Norms Literature: An Overview of Reviews. *Perspectives on Psychological Science*, *15*(1), 62–80. https://doi.org/10.1177/1745691619866455

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. https://doi.org/10.1146/annurev.polisci.3.1.475

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and Distrust : New Relationships and Realities. *The Academy of Management Review*, *23*(3), 438–458.

Lewis, M. A., & Neighbors, C. (2006). Social norms approaches using descriptive drinking norms education: A review of the research on personalized normative feedback. *Journal of American College Health*, *54*(4), 213–218. https://doi.org/10.3200/JACH.54.4.213-218

Loos, E., & Nijenhuis, J. (2020). *Consuming Fake News: A Matter of Age? The Perception of Political Fake News Stories in Facebook Ads* (Vol. 12209 LNCS). Springer International Publishing. https://doi.org/10.1007/978-3-030-50232-4_6

Luskin, R. C., Gaurav, S., Park, Y. M., & Blank, J. (2018). *Misinformation about Misinformation? Of Headlines and Survey Design*.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, *5*(1). https://doi.org/10.1186/s41235-020-00252-3

McIntyre, L. (2018). *Post-truth*. MIt Press.

McKnight, D. H., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. *Trust in Cyber-societies: Integrating the human and artificial perspectives.*, 27–54. https://doi.org/10.1007/3-540-45547-7_3

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-020-20043-0

Mourão, R. R., & Robertson, C. T. (2019). Fake News as Discursive Integration: An Analysis of Sites That Publish False, Misleading, Hyperpartisan and Sensational Information. *Journalism Studies*, *20*(14), 2077–2095. https://doi.org/10.1080/1461670X.2019.1566871

Nyborg, K., & Rege, M. (2003). On social norms: The evolution of considerate smoking behavior. *Journal of Economic Behavior and Organization*, *52*(3), 323–340. https://doi.org/10.1016/S0167-2681(03)00031-3

OECD. (2017). *OECD Guidelines on Measuring Trust*. https://doi.org/10.1787/9789264278219-en

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.

Pantazi, M., Hale, S., & Klein, O. (2021). Social and Cognitive Aspects of the Vulnerability to Political Misinformation. https://doi.org/10.1111/pops.12797

Papapicco, C., Lamanna, I., & D'errico, F. (2022). Adolescents' Vulnerability to Fake News and to Racial Hoaxes: A Qualitative Analysis on Italian Sample. *Multimodal Technologies and Interaction*, *6*(3). https://doi.org/10.3390/mti6030020

Passarelli, G. (2017). Electoral systems in context: Italy. *The Oxford Handbook of Electoral Systems*, (March), 851–870. https://doi.org/10.1093/oxfordhb/9780190258658.013.35

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Binnendyk, J., & Rand, D. (2022). Overconfidently conspiratorial: Conspiracy believers are dispositionally overconfident and massively overestimate how much others agree with them.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., & Rand, D. G. (2021a). Research note: Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *Harvard Kennedy School Misinformation Review*.

Pennycook, G., & Rand, D. G. (2021b). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications*, *13*(1), 2333.

Pereira, A., Monteiro, M. B., & Camino, L. (2009). Social norms and prejudice against homosexuals. *Spanish Journal of Psychology*, *12*(2), 576–584. https://doi.org/10.1017/S1138741600001943

Piaget, J. (1932). *The moral judgment of the child*. Routledge.

Pickles, K., Cvejic, E., Nickel, B., Copp, T., Bonner, C., Leask, J., Ayre, J., Batcup, C., Cornell, S., Dakin, T., Dodd, R. H., Isautier, J. M., & McCaffery, K. J. (2021). COVID-19 misinformation in Australia: key groups and trends over time in a national longitudinal survey. *Journal of medical Internet research*. https://doi.org/10.2196/23805

Pimienta, D., Blanco, Á., & de Oliveira, G. M. (2023). The method behind the unprecedented production of indicators of the presence of languages in the Internet. *Frontiers in Research Metrics and Analytics*, *8*. https://doi.org/10.3389/frma.2023.1149347

Pornpitakpan, C. (2004). The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Powell, M., Yörük, E., & Bargu, A. (2020). Thirty years of the Three Worlds of Welfare Capitalism: A review of reviews. *Social Policy and Administration*, *54*(1), 60–87. https://doi.org/10.1111/spol.12510

Pytlikzillig, L. M., & Kimbrough, C. D. (2016). Consensus on Conceptualizations and Definitions of Trust: Are We There Yet? In *Interdisciplinary perspectives on trust: Towards theoretical and methodological integration* (pp. 17–47). https://doi.org/10.1007/978-3-319-22261-5

Rahhal, T. A., May, C. P., & Hasher, L. (2002). Truth and character: Sources that older adults can remember. *Psychological Science*, *13*(2), 101–105. https://doi.org/10.1111/1467-9280.00419

Rampersad, G., & Althiyabi, T. (2020). Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology and Politics*, *17*(1), 1–11. https://doi.org/10.1080/19331681.2019.1686676

Robertson, C. T., & Mourão, R. R. (2020). Faking Alternative Journalism? An Analysis of Self-Presentations of "Fake News" Sites. *Digital Journalism*, *8*(8), 1011–1029. https://doi.org/10.1080/21670811.2020.1743193

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10). https://doi.org/10.1098/RSOS.201199

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision Making*, *16*(2), 484–504. https://doi.org/10.31234/osf.io/cgsx6

Sartori, G. (2005). *Parties and Party Systems: A Framework for Analysis*. ECPR press. https://doi.org/10.2307/2148863

Scheibenzuber, C., Hofer, S., & Nistor, N. (2021). Designing for fake news literacy training: A problem-based undergraduate online-course. *Computers in Human Behavior*, *121*(March), 106796. https://doi.org/10.1016/j.chb.2021.106796

Schulz, A. (2019). Where populist citizens get the news: An investigation of news audience polarization along populist attitudes in 11 countries. *Communication Monographs*, *86*(1), 88–111. https://doi.org/10.1080/03637751.2018.1508876

Schulz, A., Müller, P., Schemer, C., Wirz, D. S., Wettstein, M., & Wirth, W. (2018). Measuring Populist Attitudes on Three Dimensions. *International Journal of Public Opinion Research*, *30*(2), 316–326. https://doi.org/10.1093/ijpor/edw037

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 15. http://arxiv.org/abs/1708.01967

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-Cultural Translation: Methodology and Validation. *Journal of Cross-Cultural Psychology*, *25*(4), 501–524.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–726. https://doi.org/10.1017/S0140525X00003435

Stavrova, O., Fetchenhauer, D., & Schlösser, T. (2013). Why are religious people happy? The effect of the social norm of religiosity across countries. *Social Science Research*, *42*(1), 90–105. https://doi.org/10.1016/j.ssresearch.2012.07.002

Steffens, M. S., Dunn, A. G., Wiley, K. E., & Leask, J. (2019). How organisations promoting vaccination respond to misinformation on social media: a qualitative investigation. *BMC Public Health*, *19*(1), 1–12. https://doi.org/10.1186/s12889-019-7659-3

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J., & Loker, K. (2019). Who Shared It?: Deciding What News to Trust on Social Media. *Digital Journalism*, *7*(6), 783–801. https://doi.org/10.1080/21670811.2019.1623702

Stier, S., Kirkizh, N., Froio, C., & Schroeder, R. (2020). Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys. *International Journal of Press/Politics*, *25*(3), 426–446. https://doi.org/10.1177/1940161220907018

Sundar, S. S. (1999). Exploring receivers' criteria for perception of print and online news. *Journalism and Mass Communication Quaterly*, *76*(2), 373–386. https://doi.org/10.1177/107769909907600213

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, *51*, 72–82. https://doi.org/10.1016/J.JRETCONSER.2019.05.026

Tandoc, E. C. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), e12724. https://doi.org/10.1111/soc4.12724

Tandoc, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, *21*(3), 381–398. https://doi.org/10.1177/1464884919868325

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Valentim, V. (2021). *Parliamentary Representation and the Normalization of Radical Right Support* (Vol. 54). https://doi.org/10.1177/0010414021997159

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

Van De Walle, S., & Six, F. (2014). Trust and Distrust as Distinct Concepts: Why Studying Distrust in Institutions is Important. *Journal of Comparative Policy Analysis: Research and Practice*, *16*(2), 158–174. https://doi.org/10.1080/13876988.2013.785146

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, Culture and Society*, *42*(3), 460–470. https://doi.org/10.1177/0163443720906992

Vegetti, F., & Mancosu, M. (2020). The Impact of Political Sophistication and Motivated Reasoning on Misinformation. *Political Communication*, *37*(5), 678–695. https://doi.org/10.1080/10584609.2020.1744778

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 1–108. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Weeks, B. E. (2015). Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation. *Journal of Communication*, *65*(4), 699–719. https://doi.org/10.1111/jcom.12164

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). *Misinformation in Social Media: Definition, manipulation, and Detection* (tech. rep. No. 2). https://doi.org/10.1145/3373464.3373475

Wuttke, A., Schimpf, C., & Schoen, H. (2020). When the Whole Is Greater than the Sum of Its Parts: On the Conceptualization and Measurement of Populist Attitudes and Other Multidimensional Constructs. *American Political Science Review*, *114*(2), 356–374. https://doi.org/10.1017/S0003055419000807

Zimmermann, F., & Kohring, M. (2020). Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election. *Political Communication*, *37*(2), 215–237. https://doi.org/10.1080/10584609.2019.1686095

Appendix A

Regression Tables

Table A.1: MLM Regression Models - Convenience ITA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.315	0.390	0.285	0.373
	0.129 (0.015) *	0.155 (0.012) *	0.129 (0.028) *	0.155 (0.016) *
Pol. Consistency	0.655	0.600	0.650	0.603
	0.316 (0.038) *	0.323 (0.063) +	0.316 (0.040) *	0.323 (0.062) +
Pol. Inconsistency	0.485	0.517	0.493	0.524
	0.320 (0.130)	0.327 (0.114)	0.320 (0.124)	0.327 (0.109)
Pol. Orientation	−0.107	−0.101	−0.103	−0.096
	0.020 (<0.001) ***	0.025 (<0.001) ***	0.021 (<0.001) ***	0.025 (<0.001) ***
Anti-elitism	0.540	0.388		
	0.272 (0.047) *	0.312 (0.214)		
Inst. Trust (original)	0.685	0.985		
	0.288 (0.017) *	0.336 (0.003) **		
Fact. Bel. - Accurata	0.380	0.203	0.366	0.189
	0.097 (<0.001) ***	0.111 (0.067) +	0.098 (<0.001) ***	0.111 (0.088) +
Emp. Exp. - Intorno a un quarto	0.034	0.233	0.019	0.198
	0.247 (0.889)	0.297 (0.434)	0.248 (0.938)	0.299 (0.508)
Emp. Exp. - Più o meno la metà	0.058	0.216	0.036	0.191
	0.258 (0.822)	0.304 (0.478)	0.259 (0.890)	0.306 (0.533)
Emp. Exp. - Intorno a tre quarti	−0.080	0.060	−0.111	0.038
	0.272 (0.767)	0.317 (0.851)	0.273 (0.685)	0.320 (0.905)
Emp. Exp. - Quasi tutti/e	−0.634	−1.101	−0.737	−1.152
	0.376 (0.091) +	0.488 (0.024) *	0.394 (0.062) +	0.491 (0.019) *
Pers. Norm. Bel. - Accurata	0.246	0.201	0.257	0.209
	0.092 (0.008) **	0.106 (0.058) +	0.093 (0.006) **	0.107 (0.050) +
Norm. Exp. - Intorno a un quarto	0.360	0.294	0.377	0.315
	0.248 (0.146)	0.303 (0.331)	0.249 (0.130)	0.305 (0.302)
Norm. Exp. - Più o meno la metà	0.116	−0.003	0.144	0.017
	0.255 (0.649)	0.307 (0.993)	0.256 (0.573)	0.309 (0.955)
Norm. Exp. - Intorno a tre quarti	0.484	0.351	0.521	0.376
	0.271 (0.074) +	0.325 (0.281)	0.272 (0.056) +	0.327 (0.250)
Norm. Exp. - Quasi tutti/e	0.421	0.680	0.525	0.753
	0.364 (0.247)	0.449 (0.131)	0.371 (0.157)	0.453 (0.097) +
Treatment - Presence	−0.004	0.027	−0.016	0.012
	0.098 (0.965)	0.107 (0.798)	0.098 (0.872)	0.107 (0.913)
Treatment - Absence	0.039	0.019	0.030	0.022
	0.098 (0.691)	0.105 (0.854)	0.098 (0.757)	0.106 (0.833)
Trust in Universities			0.497	0.433
			0.235 (0.034) *	0.273 (0.113)
Inst.Trust (factor analysis)			0.004	0.279
			0.243 (0.988)	0.280 (0.320)
Num.Obs.	4229	3721	4205	3707
R2 Marg.	0.179	0.228	0.177	0.225
R2 Cond.	0.454	0.480	0.454	0.479
AIC	14 760.6	13 226.3	14 687.0	13 185.7
BIC	15 135.2	14 153.4	15 061.3	14 112.2
ICC	0.3	0.3	0.3	0.3
RMSE	1.18	1.16	1.18	1.16

Table A.2: MLM Regression Models - Main-ITA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.514	0.417	0.507	0.424
	0.093 (<0.001) ***	0.101 (<0.001) ***	0.093 (<0.001) ***	0.100 (<0.001) ***
Pol. Consistency	0.280	0.261	0.299	0.281
	0.319 (0.379)	0.321 (0.417)	0.324 (0.355)	0.326 (0.388)
Pol. Inconsistency	0.296	0.301	0.315	0.318
	0.321 (0.356)	0.324 (0.352)	0.326 (0.333)	0.328 (0.332)
Pol. Orientation	−0.193	−0.142	−0.148	−0.104
	0.163 (0.236)	0.170 (0.405)	0.162 (0.363)	0.169 (0.539)
Anti-elitism	0.740	0.362		
	0.204 (<0.001) ***	0.219 (0.098) +		
Inst. Trust (original)	0.309	0.391		
	0.191 (0.106)	0.200 (0.051) +		
Fact. Bel. - Accurate	0.047	−0.005	0.033	−0.031
	0.078 (0.548)	0.081 (0.951)	0.078 (0.671)	0.080 (0.700)
Emp. Exp. - Pochi/e partecipanti (...	0.317	0.224	0.304	0.225
	0.153 (0.038) *	0.160 (0.161)	0.152 (0.046) *	0.158 (0.156)
Emp. Exp. - Più o meno la metà	0.122	0.172	0.125	0.193
	0.159 (0.441)	0.164 (0.294)	0.158 (0.430)	0.163 (0.238)
Emp. Exp. - La maggior parte (into...	0.146	0.057	0.097	0.027
	0.166 (0.377)	0.171 (0.737)	0.165 (0.559)	0.170 (0.872)
Emp. Exp. - Quasi tutti/e	0.544	0.418	0.563	0.450
	0.245 (0.026) *	0.250 (0.094) +	0.245 (0.021) *	0.249 (0.071) +
Pers. Norm. Bel. - Accurate	0.229	0.215	0.245	0.231
	0.075 (0.002) **	0.078 (0.006) **	0.075 (0.001) **	0.078 (0.003) **
Norm. Exp. - Pochi/e partecipanti (...	−0.068	−0.239	−0.033	−0.226
	0.162 (0.675)	0.169 (0.158)	0.159 (0.836)	0.167 (0.175)
Norm. Exp. - Più o meno la metà	−0.089	−0.303	−0.072	−0.299
	0.167 (0.595)	0.173 (0.080) +	0.164 (0.660)	0.171 (0.081) +
Norm. Exp. - La maggior parte (into...	−0.185	−0.299	−0.139	−0.279
	0.171 (0.278)	0.177 (0.092) +	0.168 (0.409)	0.175 (0.112)
Norm. Exp. - Quasi tutti/e	−0.234	−0.287	−0.204	−0.292
	0.229 (0.307)	0.235 (0.223)	0.228 (0.370)	0.234 (0.212)
Treatment - Presence	0.051	0.059	0.056	0.066
	0.077 (0.507)	0.077 (0.442)	0.076 (0.460)	0.076 (0.387)
Treatment - Absence	−0.080	−0.115	−0.076	−0.116
	0.081 (0.328)	0.082 (0.158)	0.081 (0.351)	0.081 (0.154)
Trust in Universities			0.426	0.347
			0.178 (0.017) *	0.186 (0.062) +
Inst.Trust (factor analysis)			−0.118	0.143
			0.192 (0.537)	0.205 (0.486)
Num.Obs.	7470	7369	7511	7410
R2 Marg.	0.137	0.181	0.135	0.181
R2 Cond.	0.400	0.422	0.401	0.424
AIC	26 773.9	26 553.4	26 911.8	26 677.4
BIC	27 182.1	27 637.5	27 320.3	27 776.2
ICC	0.3	0.3	0.3	0.3
RMSE	1.26	1.25	1.26	1.24

Table A.3: MLM Regression Models - Main-USA. Dependent variable: perceived accuracy of social media posts. Estimates refer to interaction coefficients between various explanatory variables and a dummy variable describing the veracity of posts. Coefficients for categorical variables refer to all their levels except for the reference category. Control variables: age, gender, education, profession, media consumption, item fixed effects (only for OLS models).

	Reduced	Reduced + Controls	Indexes	Indexes + Controls
CRT	0.718	0.568	0.656	0.516
	0.148 (<0.001) ***	0.155 (<0.001) ***	0.153 (<0.001) ***	0.160 (0.001) **
Pol. Orientation	−0.944	−0.901	−0.768	−0.807
	0.124 (<0.001) ***	0.133 (<0.001) ***	0.127 (<0.001) ***	0.137 (<0.001) ***
Anti-elitism	1.148	0.695		
	0.241 (<0.001) ***	0.256 (0.007) **		
Inst. Trust (original)	−0.199	0.106		
	0.155 (0.201)	0.180 (0.557)		
Fact. Bel. - Accurate	0.339	0.249	0.353	0.277
	0.087 (<0.001) ***	0.092 (0.007) **	0.088 (<0.001) ***	0.094 (0.003) **
Emp. Exp. - Few participants (arou...	0.142	0.250	0.006	−0.029
	0.166 (0.393)	0.178 (0.160)	0.168 (0.973)	0.182 (0.874)
Emp. Exp. - Half of them	−0.068	0.105	−0.110	−0.070
	0.163 (0.675)	0.175 (0.549)	0.163 (0.502)	0.176 (0.691)
Emp. Exp. - Most of them (around t...	0.201	0.265	0.174	0.132
	0.169 (0.236)	0.181 (0.142)	0.170 (0.308)	0.182 (0.469)
Emp. Exp. - Almost all of them	0.224	0.339	0.181	0.261
	0.180 (0.212)	0.191 (0.076) +	0.180 (0.316)	0.190 (0.170)
Pers. Norm. Bel. - Accurate	0.365	0.180	0.294	0.110
	0.088 (<0.001) ***	0.094 (0.056) +	0.090 (0.001) **	0.096 (0.253)
Norm. Exp. - Few participants (arou...	0.103	−0.082	0.081	−0.031
	0.172 (0.551)	0.182 (0.655)	0.174 (0.641)	0.185 (0.866)
Norm. Exp. - Half of them	0.129	−0.082	0.030	−0.112
	0.165 (0.435)	0.176 (0.642)	0.163 (0.854)	0.175 (0.524)
Norm. Exp. - Most of them (around t...	0.042	−0.167	0.087	−0.086
	0.168 (0.805)	0.180 (0.352)	0.167 (0.602)	0.181 (0.635)
Norm. Exp. - Almost all of them	0.167	−0.011	0.134	−0.017
	0.177 (0.345)	0.185 (0.952)	0.177 (0.448)	0.185 (0.925)
Treatment - Presence	−0.020	−0.020	−0.029	−0.042
	0.087 (0.819)	0.087 (0.819)	0.089 (0.742)	0.089 (0.633)
Treatment - Absence	−0.106	−0.116	−0.150	−0.142
	0.091 (0.245)	0.092 (0.204)	0.093 (0.108)	0.094 (0.130)
Trust in Universities			1.050	1.008
			0.191 (<0.001) ***	0.205 (<0.001) ***
Inst.Trust (factor analysis)			−1.273	−0.940
			0.216 (<0.001) ***	0.240 (<0.001) ***
Num.Obs.	6206	6163	5894	5851
R2 Marg.	0.229	0.271	0.224	0.267
R2 Cond.	0.474	0.499	0.468	0.496
AIC	22 660.6	22 696.8	21 464.0	21 496.3
BIC	22 990.5	23 806.7	21 791.4	22 597.6
ICC	0.3	0.3	0.3	0.3
RMSE	1.28	1.26	1.28	1.25

Appendix B

Complete List of Utilized Social Media Posts

(a) 1- False

(c) 2- False

(e) 3- False

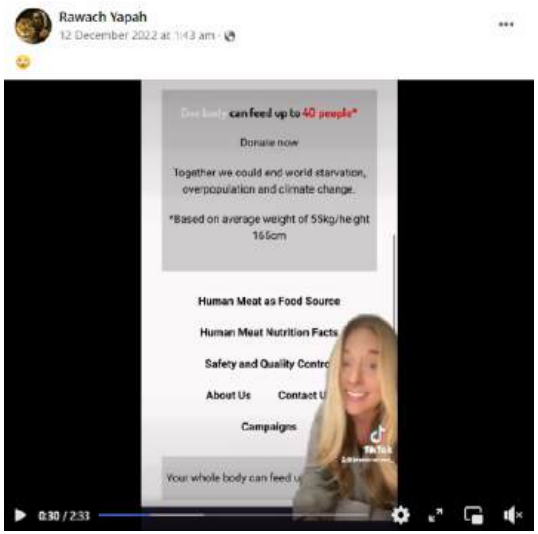
(a) 1- Reliable

(c) 2- Reliable

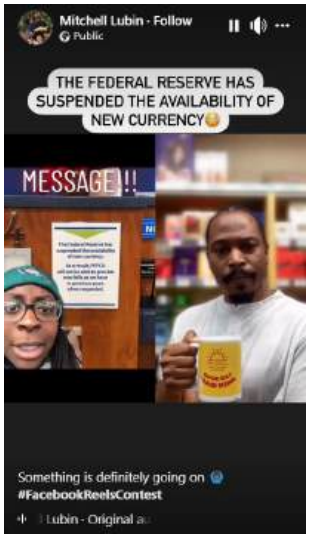
(e) 3- Reliable

Figure B.1: Social Media Posts Collected for The US Study.

(a) 4- False



(c) 5- False



(e) 6- False

(a) 4- Reliable

(c) 5- Reliable

(e) 6- Reliable

Figure B.2: Social Media Posts Collected for The US Study.

(a) 7- False

(c) 8- False

(e) 9- False

(a) 7- Reliable

(c) 8- Reliable

(e) 9- Reliable

Figure B.3: Social Media Posts Collected for The US Study.

(a) 10- False

(c) 11- False

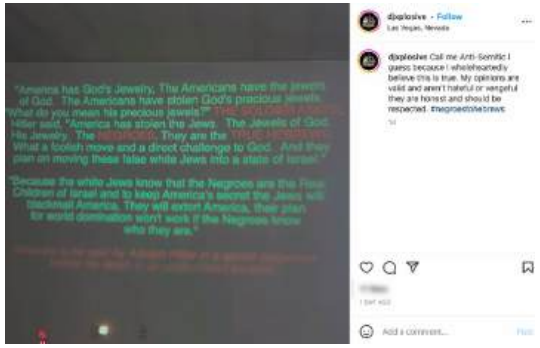
(e) 12- False

(a) 10- Reliable

(c) 11- Reliable

(e) 12- Reliable

Figure B.4: Social Media Posts Collected for The US Study.



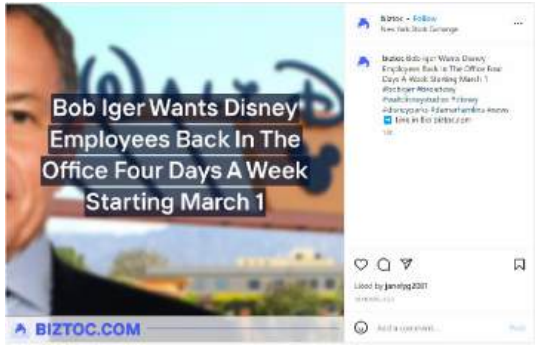
(a) 13- False

(c) 14- False

(e) 15- False

(a) 13- Reliable

(c) 14- Reliable



(e) 15- Reliable

Figure B.5: Social Media Posts Collected for The US Study.

(a) 16- False

(c) 17- False

(e) 18- False

(a) 16- Reliable

(c) 17- Reliable

(e) 18- Reliable

Figure B.6: Social Media Posts Collected for The US Study.

(a) 19- False

(c) 20- False

(a) 19- Reliable

(c) 20- Reliable

Figure B.7: Social Media Posts Collected for The US Study.

(a) 1- False

(c) 2- False

(e) 3- False

(a) 1- Reliable

(c) 2- Reliable

(e) 3- Reliable

Figure B.8: Social Media Posts Collected for The Italian Studies.

(a) 4- False

(c) 5- False

(e) 6- False

(a) 4- Reliable

(c) 5- Reliable

(e) 6- Reliable

Figure B.9: Social Media Posts Collected for The Italian Studies.

(a) 7- False

(c) 8- False

(e) 9- False

(a) 7- Reliable

(c) 8- Reliable

(e) 9- Reliable

Figure B.10: Social Media Posts Collected for The Italian Studies.

(a) 10- False

(c) 11- False

(e) 12- False

(a) 10- Reliable

(c) 11- Reliable

(e) 12- Reliable

Figure B.11: Social Media Posts Collected for The Italian Studies.

(a) 13- False

(c) 14- False

(e) 15- False

(a) 13- Reliable

(c) 14- Reliable

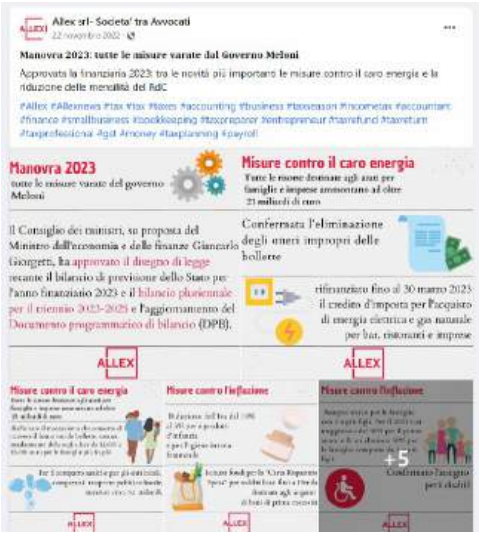
(e) 15- Reliable

Figure B.12: Social Media Posts Collected for The Italian Studies.

(a) 16- False

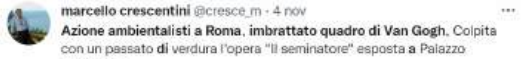
(c) 17- False

(e) 18- False



(a) 16- Reliable

(c) 17- Reliable



(e) 18- Reliable

Figure B.13: Social Media Posts Collected for The Italian Studies.

(a) 19- False

(c) 20- False

(a) 19- Reliable

(c) 20- Reliable

Figure B.14: Social Media Posts Collected for The Italian Studies.

Appendix C

Questionnaire of Main Studies - English Version

Misinformation Survey - US

Survey Flow

Start of Block: Informed Consent

JS

INTRO Welcome and thanks for participating in this survey. This research project is conducted by Fabio Torreggiani (University of Milan) and Dr. Aron Szekely (Collegio Carlo Alberto - Turin). It is a study to learn more about people's knowledge of news.

Your participation is voluntary and it involves the completion of a survey. The survey consists in seven sections. All the sections are mandatory and you will be automatically excluded if you will not complete them. The survey will take around fifteen to twenty minutes to be completed. Risks to participation are minimal, and you will be helping to further scientific knowledge.

* { box-sizing: border-box; } /* Create three equal columns that floats next to each other */ .column { float: left; width: 33.33%; padding: 10px; height: 300px; /* Should be removed. Only for demonstration */ } /* Clear floats after the columns */ .row:after { content: ""; display: table; clear: both; } /* Responsive layout - makes the three columns stack on top of each other instead of next to each other */ @media screen and (max-width: 600px) { .column { width: 100%; } }

Page Break

RULES Important Rules

The survey will not contain any form of deception. All that we will communicate to you is true. Your participation is voluntary and you can leave the survey at any time. However, in that case, you will not receive any payment. You must be age 18 or older to participate. Your answers will be kept confidential. You are asked not to communicate with other people during the survey and to make your answers in an independent manner. You may not share your participation link with anyone.

TWH Optional Supplement - X (formerly known as Twitter) Username
Do you have a personal X (Twitter) account?

- ○ Yes (1)
- ○ No (2)

JS

HELPTXT The authors of this research would be interested in being able to add people's, and specifically, your answers to this survey to publicly available information from your X (Twitter) account, such as your profile information, tweets in the past and in future, the accounts you follow.

Your X (Twitter) information will be treated as confidential and given the same protections as your survey data. Your X (Twitter) username, and any information that would allow you to be identified, will not be published. To thank you for participating in this part of the survey, you will be offered an additional reward of 2$.

The participation in this part of the survey is completely optional and it will not invalidate the rest of the survey: you can still participate in the rest of the survey, without participating in the X (Twitter) part.

You can find further information on how we will manage your data by clicking on the FAQ below.

A- What information will you collect from my X (Twitter) account?
We will only collect information from your X (Twitter) account that is publicly available. This will include information from your account (such as your profile description, who you follow, and who

follows you), the content of your tweets (including text, images, videos and web links), and background information about your tweets (such as when you tweeted, what type of device you tweeted from, and the location the tweet was sent from). We will collect information from your past tweets (up to the last 3,000). B- What will the information be used for?

The information will be used for social research purposes only. At the end of the survey period you will be able to find further information on this research project by visiting this webpage: https://fabtorreggiani.github.io/mis-survey2023-project-page/ C- Who will be able to access the information?

Matched data which includes both your survey answers and X (Twitter) information will be made available for social research purposes only. Researchers who want to use your matched X (Twitter) and survey information must apply to access it and present a strong scientific case to ensure that the information is used responsibly and safely. Matched statistical information from your X (Twitter) account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same access controls as your other survey answers. At no point will any information that would allow you to be identified be made available to the public. D- What will you do to keep my information safe?

All information we collect will be held in accordance with the California Consumer Privacy Act (CCPA). Because X (Twitter) information is public data that anyone can search, it is impossible to anonymise completely. To keep your information safe, researchers will only be able to access the matched survey answers and detailed X (Twitter) information in a secure environment set up to protect this type of data. Only approved researchers who have gone through special training may access this information, and they will have to apply to do so. Statistical information from your X (Twitter) account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same level of protection as your other survey answers. E- What if I change my mind?

This information will be collected and stored for as long as they are useful for research purposes, or until you contact us to withdraw your permission. You can do this at any time by emailing us at fabio.torreggiani@unimi.it, and do not have to give a reason.

Display This Question:

If TWH = 1

TWCONS Are you willing to tell us your personal X (Twitter) username and for your X (Twitter) information to be added to your answers to this survey? (You will be asked your X (Twitter) username just after the approval of the informed consent at the end of this page)

- ○ Yes (1)
- ○ No (2)

PAYMENT Payment

You will be paid at the end of the survey. You will receive a flat sum for completing the survey and this sum will NOT depend on the answers you make during the survey. Additional instructions on how you can access to extra payments can be found below.

Additional Payments

During the survey you can access to additional rewards in two ways: Among other things, you will be asked to estimate other participants' answers to two questions. You will receive an extra 0.3$ for each correct answer. As anticipated, if you have a X (Twitter) account and you agree to participate in the X (Twitter) part of the survey (see above), you will be offered an additional sum of 2$ as a tribute of gratitude.

CONSENT Consent

"I have read the information regarding the procedures and confidentiality of the study, and I agree to participate in it."

(Whenever you feel ready, click on the "->" button at the end of the page. If you have any doubts or questions about this project, please get in touch with Fabio Torreggiani at fabio.torreggiani@unimi.it.)

- ○ Yes, I agree to participate (1)
- ○ No, I do not agree to participate (2)

Skip To: End of Survey If CONSENT = 2

End of Block: Informed Consent

Start of Block: Demographics

Display This Question:

If TWH = 1

And TWCONS = 1

TWNAME What is your X (Twitter) username (e.g. @johnsmith)?

_____

qualtrics XM

AGE How old are you?

▼ Select your age (99) ... 99 or older (98)

GENDER How do you identify yourself?

- ○ Man (1)
- ○ Woman (2)
- ○ Non-binary / third gender (3)
- ○ Prefer to self-identify (4) _____
- ○ Prefer not to say (5)

EDU What is the highest level of education you have completed?

- ○ No education (1)
- ○ Early childhood education (2)
- ○ Primary education (elementary school) (3)
- ○ Lower secondary education (middle school) (4)
- ○ Upper secondary education (high school) (5)
- ○ Post-secondary non-tertiary education (6)
- ○ Short-cycle tertiary education (7)
- ○ Bacherlor's or equivalent (8)
- ○ Master's or equivalent (9)
- ○ Doctorate or equivalent (10)

PROFESSION What is/was the name or title of your main job?

- Manager (1)
- Professional (2)
- Technician or associate professional (3)
- Clerical support worker (4)
- Service and sales workers (5)
- Skilled agricultural, forestry and fishery workers (6)
- Craft related trade workers (7)
- Plant and machine operators, and assemblers (8)
- Elementary occupations (9)
- Armed forces occupations (10)
- I have never worked yet (11)

End of Block: Demographics

Start of Block: Twitter Usage

TWLURK On average, how frequently do you engage in passive usage on X (Twitter), such as reading tweets or browsing content without actively participating?

- Never (1)
- Once a year (2)
- Once every few months (3)
- Once a month (4)
- Once a week (5)
- 2-3 times a week (6)
- 4-6 times a week (7)
- Daily or more (8)
- Prefer not to say (9)

TWACTIVE On average, how frequently do you engage in active interaction on X (Twitter), including tweeting original content, commenting, liking or retweeting others' content?

- Never (1)
- Once a year (2)
- Once every few months (3)
- Once a month (4)
- Once a week (5)
- 2-3 times a week (6)
- 4-6 times a week (7)
- Prefer not to say (8)

End of Block: Twitter Usage

Start of Block: Cognitive Tasks

COGINTRO In the following section you will be asked three questions. Please do your best to answer as accurately as possible.

COGT1 The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?

_____

COGT2 If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take 50 printers to print out 50 pages of paper?

_____

COGT3 On a loaf of bread, there is a patch of mold. Every day, the patch doubles in size. If it takes 40 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?

_____

End of Block: Cognitive Tasks

Start of Block: Political affiliation

POL1 Which party do you align with?

- Strong Democrat (1)
- Democrat (2)
- Independent Lean Democrat (3)
- Independent (4)
- Independent Lean Republican (5)
- Republican (6)
- Strong Republican (7)
- No party appeals to me (8)
- I prefer not to answer (9)

POL2 In politics people sometimes talk of left and right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right?

- ○ Left (1)
- ○ 1 (2)
- ○ 2 (3)
- ○ 3 (4)
- ○ 4 (5)
- ○ 5 (6)
- ○ 6 (7)
- ○ 7 (8)
- ○ 8 (9)
- ○ 9 (10)
- ○ Right (11)
- ○ I prefer not to answer (12)

POL3 How would you place your views on these scales?
Equality and Individual Effort

- Incomes should be made more equal (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- There should be greater incentives for individual effort (11)
- I prefer not to answer (12)

POL4 Responsibility for well-being

- Government should take more responsibility to ensure that everyone is provided for (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- People should take more responsibility to provide for themselves (11)
- I prefer not to answer (12)

POL5 How much do you agree or disagree with the statement that nowadays one often has trouble deciding which moral rules are the right ones to follow?

- Completely agree (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- Completely disagree (11)
- I prefer not to answer (12)

POL6 How important is it for you to live in a country that is governed democratically?

- Not at all important (1)
- 1 (2)
- 2 (3)
- 3 (4)
- 4 (5)
- 5 (6)
- 6 (7)
- 7 (8)
- 8 (9)
- 9 (10)
- Absolutely important (11)
- I prefer not to answer (12)

End of Block: Political affiliation

Start of Block: Anti-Elitism and Inst. Trust

INST Please indicate how much confidence you have in the following institutions

	None at all (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	A great deal (10)	I prefer not to answer (11)
The press (1)	○	○	○	○	○	○	○	○	○	○	○
The Parliament (2)	○	○	○	○	○	○	○	○	○	○	○
The Government (3)	○	○	○	○	○	○	○	○	○	○	○
Social media (4)	○	○	○	○	○	○	○	○	○	○	○
Universities (5)	○	○	○	○	○	○	○	○	○	○	○

SCI1 How much do you agree or disagree with the statement that science and technology are making our lives healthier, easier, and more comfortable?

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)
- ○ I prefer not to answer (8)

SCI2 All things considered, would you say that the world is better off, or worse off, because of science and technology?

- A lot worse off (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 (7)
- 8 (8)
- 9 (9)
- A lot better off (10)
- I prefer not to answer (11)

ANTIEL Below, you will find a list of opinions on the political class and the society.

qualtrics XM

For each of the following statements, please indicate your degree of agreement or disagreement.

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)	I prefer not to answer (8)
Politicians talk too much and take too little action (1)	○	○	○	○	○	○	○	○
The differences between the people and the so-called elite are greater than within the people (2)	○	○	○	○	○	○	○	○
Politicians care about what ordinary people think (3)	○	○	○	○	○	○	○	○

End of Block: Anti-Elitism and Inst. Trust

Start of Block: Social norms

SOCINTRO Below you will be asked some questions; some of them are about other participants' answers.

In responding to those questions, remember that they are about **participants of a session of a**

similar survey.

SOCNORM1 When deciding whether to share a news on social media, **how important is it to you** that the content is... (Please order the following factors from the most to the least important)
_____ Surprising (1)
_____ Accurate (2)
_____ Interesting (3)
_____ Aligned with your politics (4)
_____ Funny (5)

SOCNORM2 In your opinion, how many other participants **indicated accuracy** of the content as **their top priority** when deciding whether to share a news on social media? (If your estimate is correct, you will earn a bonus sum of 0.30$)

- Almost no one (1)
- Few participants (around a quarter) (2)
- Half of them (3)
- Most of them (around three quarters) (4)
- Almost all of them (5)

SOCNORM3 When deciding whether to share a news on social media, **how important should it be** to people that the content is... (Please order the following factors from the most to the least important)

_____ Surprising (1)
_____ Accurate (2)
_____ Interesting (3)
_____ Aligned with your politics (4)
_____ Funny (5)

SOCNORM4 In your opinion, how many other participants indicated that **accuracy** of the content **should be** the top priority when deciding whether to share a news on social media? (If your estimate is correct, you will earn a bonus sum of 0.30$)

- Almost no one (1)
- Few participants (around a quarter) (2)
- Half of them (3)
- Most of them (around three quarters) (4)
- Almost all of them (5)

End of Block: Social norms

Start of Block: Media consumption

MEDIACON People learn what is going on in this country and the world from various sources. For each of the following sources, please indicate whether you use it to obtain information daily, weekly, monthly, less than monthly or never:

	Daily or more (1)	Weekly (2)	Monthly (3)	Less than monthly (4)	Never (5)	I prefer not to answer (6)
Newspaper (1)	○	○	○	○	○	○
TV News (2)	○	○	○	○	○	○
Radio News (3)	○	○	○	○	○	○
Mobile phone (4)	○	○	○	○	○	○
Email (5)	○	○	○	○	○	○
Internet (6)	○	○	○	○	○	○
Social media (Facebook, X (Twitter), etc.) (7)	○	○	○	○	○	○
Talk with friends or colleagues (8)	○	○	○	○	○	○

ATTCHECK Help us keep track of who is paying attention, please select - 'Somewhat disagree' in the options below.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

End of Block: Media consumption

Start of Block: ACCTASKINTRO

Q230 In the next section, you will be asked to analyse a series of posts from various social media. They contain information on various news events. For each post, you will be asked how much, in your opinion, that post accurately describes the events it mentions.

End of Block: ACCTASKINTRO

Start of Block: Fake 1

IMAGE1F

BELIEF1F To the best of your knowledge, how accurate is the above post?

- Extremely inaccurate (1)
- Moderately inaccurate (2)
- Slightly inaccurate (3)
- Slightly accurate (4)
- Moderately accurate (5)
- Extremely accurate (6)
- I don't know (7)

Display This Question:

If round = 3

Or round = 6

Or round = 9

SHARING1F If you were to see the above post online, how likely would you be to share it?

- Extremely unlikely (1)
- Moderately unlikely (4)
- Slightly unlikely (5)
- Slightly likely (6)
- Moderately likely (7)
- Extremely likely (8)

End of Block: Fake 1

Start of Block: True 1

qualtrics XM

IMAGE1T

BELIEF1T To the best of your knowledge, how accurate is the above post?

- ○ Extremely inaccurate (1)
- ○ Moderately inaccurate (2)
- ○ Slightly inaccurate (3)
- ○ Slightly accurate (4)
- ○ Moderately accurate (5)
- ○ Extremely accurate (6)
- ○ I don't know (7)

Display This Question:

If round = 3

Or round = 6

Or round = 9

SHARING1T If you were to see the above post online, how likely would you be to share it?

- Extremely unlikely (1)
- Moderately unlikely (4)
- Slightly unlikely (5)
- Slightly likely (6)
- Moderately likely (7)
- Extremely likely (8)

End of Block: True 1

Start of Block: Fake 2

IMAGE2F

BELIEF2F To the best of your knowledge, how accurate is the above post?

- Extremely inaccurate (1)
- Moderately inaccurate (2)
- Slightly inaccurate (3)
- Slightly accurate (4)
- Moderately accurate (5)
- Extremely accurate (6)
- I don't know (7)

qualtrics XM

Display This Question:

If round = 3

Or round = 6

Or round = 9

SHARING2F If you were to see the above post online, how likely would you be to share it?

- Extremely unlikely (1)
- Moderately unlikely (4)
- Slightly unlikely (5)
- Slightly likely (6)
- Moderately likely (7)
- Extremely likely (8)

End of Block: Fake 2

Start of Block: True 2

IMAGE2T

BELIEF2T To the best of your knowledge, how accurate is the above post?

- Extremely inaccurate (1)
- Moderately inaccurate (2)
- Slightly inaccurate (3)
- Slightly accurate (4)
- Moderately accurate (5)
- Extremely accurate (6)
- I don't know (7)

Display This Question:

If round = 3

Or round = 6

Or round = 9

SHARING2T If you were to see the above post online, how likely would you be to share it?

- Extremely unlikely (1)
- Moderately unlikely (4)
- Slightly unlikely (5)
- Slightly likely (6)
- Moderately likely (7)
- Extremely likely (8)

End of Block: True 2

Start of Block: TREAT1

TREATINTRO **Before answering the following questions, we want to share the following information with you:** We asked American respondents in an earlier survey session the same questions that you just answered about the importance of different factors when deciding whether to share a news article or not.

TREAT1 **A large majority** of these Americans stated that **accuracy is their first priority** and that **it should be the top priority of people** when deciding whether to share news on social media.

End of Block: TREAT1

Start of Block: TREAT2

TREATINTRO
Before answering the following questions, we want to share the following information with you: We asked American respondents in an earlier survey session the same questions that you just answered about the importance of different factors when deciding whether to share a news article or not.

TREAT2 **Only a small minority** of these Americans stated that **accuracy is their first priority** and **that it should be the top priority of people** when deciding whether to share news on social media.

End of Block: TREAT2

Start of Block: MANCHECK

ManCheck Some minutes ago we showed you a message containing data about Americans' priorities in news sharing. Can you recall what it said?

- ○ That a large majority of Americans indicated 'accuracy' as their top priority when deciding whether to share a news or not (1)
- ○ That only a small minority of Americans indicated 'accuracy' as their top priority when deciding whether to share a news or not (3)
- ○ I do not remember (4)

Display This Question:

If TREAT1 Is Displayed

And ManCheck = 1

OKPRES Exactly, your answer is correct. You chose *"${ManCheck/ChoiceGroup/SelectedChoices}"*.

The message we showed you reported the following information:
${TREAT1/QuestionText}

Display This Question:

If TREAT1 Is Displayed

And ManCheck != 1

WRONGPRES Unfortunately, your answer is wrong. You chose *"${ManCheck/ChoiceGroup/SelectedChoices}"*.

Instead, the message we showed you reported the following information:
${TREAT1/QuestionText}

Display This Question:

If TREAT2 Is Displayed

And ManCheck = 3

OKABS Exactly, your answer is correct. You chose *"${ManCheck/ChoiceGroup/SelectedChoices}"*.

The message we showed you reported the following information:
${TREAT2/QuestionText}

Display This Question:

If TREAT2 Is Displayed

And ManCheck != 3

WRONGABS Unfortunately, your answer is wrong. You chose *"${ManCheck/ChoiceGroup/SelectedChoices}"*.

Instead, the message we showed you reported the following information:
${TREAT2/QuestionText}

End of Block: MANCHECK

Start of Block: COMMENTS

COMMENTS Comments (optional)

If you have comments or suggestions feel free to use this space to help us improve this survey. We appreciate comments on every part of the survey.

Otherwise, go ahead with the conclusion of the survey.

_____

_____

_____

_____

_____

End of Block: COMMENTS

Start of Block: DEBRIEFING

qualtrics XM

Display This Question:

If TREAT1 Is Displayed

DEBRIEFINTRO1 Important information before the Conclusion During this survey we showed you the following message:
${TREAT1/QuestionText}

The information it contains is true. In fact, in a session of a similar survey, which took place on 4 August 2023, in which 101 people completed the survey, the answers were the following:
58.42% of participants indicated accuracy as their first priority 57.43% of participants indicated that accuracy should be the first priority of people
However, a total of 633 people participated to the full survey (composed of various survey sessions). The answers referring to the full sample of participants are the following:
69.36% of participants indicated accuracy as their first (52.78%) or second (16.9%) priority 73.24% of participants indicated that accuracy should be the first (57.75%) or second (15.5%) priority of people
In this survey we showed you a list of posts containing information on various news events. All screenshot are taken from real-life posts published on various social media. Some of them contain false information on the facts or events they mention and they were labelled as "fake news" by the debunking site PolitiFact.com.

Here below you can find the complete list of post you saw during the survey. For each post, it is indicated whether the information it contains is true or false. Besides, next to each post you can find a link to its relative sources.

Display This Question:

If TREAT2 Is Displayed

DEBRIEFINTRO2 Important information before the Conclusion During this survey we showed you the following message:
${TREAT2/QuestionText}

The information it contains are true. In fact, in a session of a similar survey, which took place on 28 July 2023, in which 28 people completed the survey, the answers were the following:
32.14% of participants indicated accuracy as their first priority 35.71% of participants indicated that accuracy should be the first priority of people
However, a total of 633 people participated to the full survey (composed of various survey sessions). The answers referring to the full sample of participants are the following:
69.36% of participants indicated accuracy as their first (52.78%) or second (16.9%) priority 73.24% of participants indicated that accuracy should be the first (57.75%) or second (15.5%) priority of people

In this survey we showed you a list of post containing information on various news events. All screenshot are taken from real-life posts published on various social media. Some of them contain false information on the facts or events they mention and they were labelled as "fake news" by the debunking site PolitiFact.com.

Here below you can find the complete list of post you saw during the survey. For each post, it is indicated whether the information it contains is true or false. Besides, next to each post you can find a link to its relative sources.

Display This Question:
If block1FAKE = 1

DEB1F ${IMAGE1F/QuestionText}
THIS POST CONTAINS FALSE INFORMATION

Source: https://www.politifact.com/factchecks/2023/feb/02/tweets/tweets-distort-bidens-comments-on-tanks-ukraine-an/

Display This Question:
If block2FAKE = 1

DEB2F ${IMAGE2F/QuestionText}
THIS POST CONTAINS FALSE INFORMATION

Source: https://www.politifact.com/factchecks/2023/jan/04/instagram-posts/damar-hamlin-remains-hospital-anti-vaxxers-spread/

Display This Question:
If block1TRUE = 1

DEB1T ${IMAGE1T/QuestionText}
THIS POST DOES NOT CONTAIN FALSE INFORMATION

Source: https://www.bbc.com/news/world-us-canada-64404928

Display This Question:

If block2TRUE = 1

DEB2T ${IMAGE2T/QuestionText}

THIS POST DOES NOT CONTAIN FALSE INFORMATION

Source: https://www.nytimes.com/2023/01/02/sports/football/damar-hamlin-bills-hit.html

End of Block: DEBRIEFING

Appendix D

Ethics Committee Certificate

Institutional Review Board

ETHICS COMMITTEE of Collegio Carlo Alberto

On December 7, 2022 the Ethics Committee of Collegio Carlo Alberto met online and examined the research project (Title: "*Social Mechanisms of Disinformation Perception. Survey Experiment on the Role of Social Norms and Anti-Elitism in Italy and the USA.*") with the questionnaire, the related documentation, and the informed consent forms presented by **Fabio Torregiani**.

The Ethics Committee examined the rationale of the project, the suitability and completeness of the information contained in the documentation, as well as the suitability of the protocol for the objectives of the study and the eligibility of the investigators testified by the curriculum vitae attached to the project.

After ample discussion, and having ensured sufficient time for a thorough examination of the issues related to the study, the Ethics Committee made up of Prof. Christopher Flinn, Prof. Valeria Marcenò, Prof. Michele Graziadei, approves, for the part within its powers, the protocol of the study the inform consent forms and the questionnaire.

Turin, 10 December 2022

Michele Graziadei

The Chair
Prof. Michele Graziadei
Collegio Carlo Alberto

Appendix E

Pre-registration

Social Mechanisms of Disinformation Perception: A Survey Experiment on the Role of Social Norms and Anti-Elitism

Created: 23th December 2022
Updated: 16th August 2023

Authors
Fabio Torreggiani (University of Milan & Collegio Carlo Alberto) – fabio.torreggiani@unimi.it
Aron Szekely (Collegio Carlo Alberto) – aron.szekely@carloalberto.org

1. Data Collection - Have any data been collected for this study already?

No data for this study have been collected until now. Data for two pilot studies have been collected and partially analysed. Additionally, a pretest of 600 participants has been collected and partially analysed. Both the pilot studies and the pretest have been collected using different samples from the ones that will be used for this study.

2. Hypothesis – What's the main question being asked or hypothesis being tested in this study?

Research question 1: What are the key mechanisms influencing how people (mis)perceive the credibility of fake news? In particular, are there any *social and political* mechanisms?

Cognitive Effort
Hypotheses 1 (confirmatory): Participants with high scores in the CRT tasks will be more accurate in judging news.

Political Affiliation and Ideology
Hypotheses 2a (confirmatory): Participants will be less accurate in judging news with a political leaning which is concordant with their political affiliation, than when judging politically discordant news.

Hypotheses 2b (confirmatory): Participants will be less accurate in judging news with a political leaning which is concordant with their political ideology, than when judging politically discordant news.

Anti-elitism and Institutional (dis)trust
Hypotheses 3a: Participants with high anti-elitism will be less accurate in judging news.

Hypotheses 3b: Participants with low institutional trust will be less accurate in judging news.

Social norms about the importance of accuracy
Hypotheses 4a: Participants with a high factual belief about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Hypotheses 4b: Participants with a high personal normative belief about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Hypotheses 5a: Participants with a high empirical expectation about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Hypotheses 5b: Participants with a high normative expectation about the existence of a social norm about sharing only accurate news will be more accurate in judging news.

Research question 2: Can a *social priming message* shape fake news perception?

Hypotheses 6a: A priming message highlighting the presence of an accuracy social norm is expected to increase participants' truth discernment.

Hypotheses 6b: A priming message highlighting the absence of an accuracy social norm is expected to decrease participants' truth discernment.

Research question 3: Do the findings from research question 1-2 *vary* between Italy and the US?

Hypotheses 7: The mechanisms driving fake news beliefs will be similar across the two contexts

Research question 4 (exploratory): What are the links between fake news *perceptions* and online social media *behaviour*?

3. Dependent Variable - Describe the key dependent variable(s) specifying how they will be measured.

(1) Perceived accuracy
During the accuracy task, participants will rate the accuracy of social media posts including news content on a 6-point scale (extremely inaccurate, moderately inaccurate, slightly inaccurate, slightly accurate, moderately accurate, extremely accurate).

(2) Intention to share
During the accuracy task, participants will indicate their willingness to share social media posts including news content using a 6-point scale (extremely unlikely, moderately unlikely, slightly unlikely, slightly likely, moderately likely, extremely likely).

4. Conditions - How many and which conditions will participants be assigned to?

Participants will be presented with 10 news posts (randomly picked from a bigger dataset containing both fake and reliable news). The experiment includes three conditions in a mixed within and between subjects study design.

During the first five items, all participants will be assigned to the **control condition**, where news posts will be presented without any priming message.

After the fifth item, participants will be randomly assigned to one of two conditions:

1. A **"presence of social norm"** priming condition where they will see a priming message highlighting the presence of a social norm regarding the importance of sharing only reliable news.

2. A **"absence of social norm"** priming condition where they will see a priming message highlighting the absence of a social norm regarding the importance of sharing only reliable news.

Phase of the accuracy task	*Group 1*	*Group 2*
First 5 news posts	No message (control)	No message (control)
Last 5 news posts	"Presence of accuracy social norm" message	"Absence of accuracy social norm" message

5. Analyses – Specify exactly which analyses you will conduct to examine the main question/hypothesis

Primary Analyses

RQ 1: What are the key mechanisms influencing how people (mis)perceive the credibility of fake news? In particular, are there any *social and political* mechanisms?

We will test all the different mechanisms with the same procedure. The level of analysis will be that of ratings, i.e. individual answers. Regression analysis will be preceded by descriptive univariate and bivariate analysis. We will use a series of OLS linear regressions with robust standard errors clustered on participants and fixed effects for the items. The dependent variable will be perceived accuracy. The independent variables will include a dummy for item veracity (0=false, 1=true), a coefficient for each explanatory variable, and a interaction coefficient between the veracity dummy and each explanatory variable; along with sociodemographic control variables. The main focus of our analysis, for each mechanism, will be the interactions between the dummy variable indicating the veracity of the item (i.e. whether it is a fake or a true news) and the coefficients of the explanatory variables (e.g. participant's CRT score).

Thus, the full regression model will be the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_A \mathbf{X} + \beta_B \mathbf{X} x_1 + \beta_C \mathbf{Z} + \beta_D \mathbf{I} + \epsilon$$

Where:

- 'y' represents the **dependent variable**, which is the perceived accuracy of social media news posts.
- 'β_0' represents the **intercept**.
- 'x_1' represents the **veracity** of items, acting as a dummy variable to indicate whether the answer refer to a reliable post or not. 0 = contains false information; 1 = does not contain false information.
- 'β_1' is the coefficient of veracity, i.e. how the perceived accuracy of a reliable post changes when compared to a post containing false information. Conceptually, this can be thought of as truth discernment.
- 'X' represents the vector of **independent variables**, encompassing variables for the four main mechanisms examined in the survey: cognitive effort, political affiliation, anti-elitism & institutional trust, and social norms.
- 'β_A' corresponds the vector of coefficients associated with the independent variables included in vector X.

- 'β_B' represents the vector of coefficients linked to the **interactions** between the main independent variables and veracity of posts. This is the main focus of the analysis as this term identifies the effects of the various independent variables on truth discernment.
- 'Z' is the vector of **control variables**.
- 'β_C' represents the vector of coefficients associated with the control variables included in vector Z.
- 'Z' represents the vector of **items fixed effects**.
- 'β_D' represents the vector of coefficients associated with the items fixed effects included in vector Z, indicating how the average perceived accuracy changes based on which one of the 40 images we consider.
- 'ε' represents the **error term**.

Initially, we will test the different hypotheses one by one, including just the dependent variable, the main explanatory variable of the mechanism currently tested, the veracity of the items and the interaction effect between the last two (i.e. veracity and mechanism).

For example, HP1 (regarding the role of cognitive ability) will be initially tested with the following model:

$$y = \beta_0 + \beta_1 veracity + \beta_2 CRT + \beta_3 veracity * CRT + \beta_A I + \epsilon$$

To this model, we will then add controls for sociodemographic variables and media consumption.

After running this procedure for each hypotheses individually, we will use the full model to test all the hypotheses simultaneously, including item veracity, all the explanatory variables, all their interactions with item veracity, controls and items fixed effects. This is exploratory in the sense that any pattern of result is revealing.

RQ 2: Can a *social priming message* shape fake news perception?

HP 6a and HP 6b will be tested with the same procedure described for the testing of hypotheses 1 to 5b. The level of analysis will be that of ratings, i.e. individual answers. We will use a series of OLS linear regression models with robust standard errors clustered on participants and items fixed effects. The dependent variable will be perceived accuracy. The independent variables will include a dummy for item veracity (0=false, 1=true), a dummy variable for each treatment condition (0= control, 1 = "presence" treatment & 0 = control, 1 = "absence" treatment), and a interaction coefficient between the veracity dummy and each treatment effect; along with sociodemographic control variables. The main focus of our analysis, for each mechanism, will be the interactions between the dummy variable indicating the veracity of the item (i.e. whether it is a fake or a true news) and the coefficients of the treatment effect.

Initially, we will test the effect of the treatment, without the inclusion of other variables. For example, HP6a will be tested with the following model:

$$y = \beta_0 + \beta_1 veracity + \beta_2 treatment + \beta_3 veracity * treatment + \beta_A I + \epsilon$$

To this model, we will then add controls for sociodemographic variables and media consumption.

After running this procedure for both hypotheses individually, we will use a full model to test the effect of the treatment alongside with the other explanatory variables. This is exploratory in the sense that any pattern of result is revealing.

Thus, the full regression model will be the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_A X + \beta_B X x_1 + \beta_C Z + \beta_D I + \epsilon$$

This model mirrors the full model explained in RQ1, with one exception: the inclusion of x_2 which represents whether the answer received treatment (with either an "absence" or "presence" message) or not.

RQ 3: Do the findings from research question 1-3 *vary* between Italy and the US?

HP 7: The mechanisms driving fake news beliefs will be similar across the two contexts

We will replicate the same models and analysis used in previous hypotheses for both the US and Italian samples. We will report eventual differences, but we do not held hypotheses regarding the nature of these differences.

RQ 4 (exploratory): What are the links between fake news *perceptions* and online social media *behaviour*?

Given the exploratory nature of this research question, we do not pre-register any hypotheses nor we anticipate the exact analysis that we will run. We outline some analytical options that we will initially pursue in the "Other" section of this document. This is exploratory in the sense that any pattern of result is revealing.

6. Outliers and Exclusions - Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

We do not plan to automatically exclude any participant during the survey experiment. We check for the percentage of people who fail an attention check during the survey. During the analysis we control for those participants who miss the attention check. We also check for the percentage of people who fail a manipulation check towards the end of the survey and control for it during the analysis.

7. Sample Size - How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.

We will collect a sample size of 1200 subjects per country (Italy and US), for a total of around 2400 participants.

8. Other - Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

RQ 1: What are the key mechanisms influencing how people (mis)perceive the credibility of fake news? In particular, are there any *social and political* mechanisms?

As a robustness check, we will run the full model by also adding rounds fixed effects, indicating in which round each answer has been given.
We will also run the same analysis separately for fake and true items. Similarly to what said for the primary analysis, the dependent variable will be perceived accuracy. In these separated models, however, the focus will be on explanatory variables' coefficients (and not their interaction with veracity). We will do this by using OLS linear regression models with robust standard errors on participants and fixed effects for the items.
We will also run the same analysis by using OLS linear mixed regression models with cross-classified structure (answers refer to participants and items, which are not nested hierarchically). In these mixed models, we will use random intercepts at both the participants and items levels and random slopes at the participants level.

RQ 2: Can a *social priming message* shape fake news perception?

As a robustness check, we will re-run the primary analysis by following the procedures already specified in the robustness checks of RQ 1.

RQ 4 (exploratory): What are the links between fake news *perceptions* and online social media *behaviour*?

The final decision on which analysis to run will be made in phase of exploration of the final data. However, among the many ways in which we could analyse this RQ, the followings are some of the options we plan to explore.

In order to explore RQ 4, we will create a series of variables starting from participants' behaviour on Twitter. We will then test the correlation of these variables with the variables collected through the survey. Our primary focus will be on misinformation perception, but we will also analyse political affiliation, institutional trust and anti-elitism (see Secondary Analysis).

HP 8: Perceptions predict own social media behaviour

We will test the correlation of the following Twitter-based variables with the perceived accuracy of fake news posts as measured in the survey:

1. Followed pages:
    a. Number of pages recognized as fake news spreaders by debunking sites and specialized sources followed on Twitter
    b. Number of hyper-partisan sources followed on Twitter
2. Number of shared tweets coming from fake news-related sources (as defined in the points a . and b. above).
3. Number of liked tweets coming from fake news-related sources (as defined in the points a. and b. above).

We will also run secondary analysis with two aims:

1. Test whether Twitter behaviour reflects beliefs measured during the survey also on dimensions other than belief in fake news (RQ 2).
2. Test whether the correlations we hypothesize to be present between stated independent variables (e.g. anti-elitism) and belief in fake news are replicated when using their behavioural corresponding variables (e.g. number of followed populist pages) built on Twitter data.

For the first aim, we will run the following analysis:

4. **Political affiliation.** We will test the correlation of the following Twitter-based variables with subjects' political affiliation as measured in the survey:
    a. Followed pages:
        i. Number of pages belonging to political parties, parties' representatives or elected politicians of one faction or the other followed on Twitter
        ii. Number of pages belonging to media recognized as being leaning to one political position or the other by independent sources followed on Twitter
    b. Number of shared tweets coming from politicians or "political" media (as defined in points i. and ii. above)
    c. Number of liked tweets coming from politicians or "political" media (as defined in points i. and ii. above)
5. **Institutional trust.** We will test the correlation of the following Twitter-based variables with subjects' institutional trust as measured in the survey. The analysis will be firstly pursued in an aggregated way (including all types of institutions) and then distinguishing for the various type of institutions mentioned in the survey:
    a. Followed pages:
        i. Number of pages belonging to various public (primarily national) institutions followed on Twitter.
    b. Number of shared tweets coming from various public institutions
    c. Number of liked tweets coming from various public institutions
6. **Anti-elitism and populism.** In the Italian sample, we will test the correlation between anti-elitism as measured in the survey and presence of Italian versus European flag in participants' account name and biography on Twitter.

For the second aim, we will test the correlations between the behavioural Twitter-based independent variables described above and the perceived accuracy of fake news as measured in the survey.

9. Name - Give a title for this AsPredicted pre-registration Suggestion: use the name of the project, followed by study description.

Social Mechanisms of Disinformation Perception: A Survey Experiment on the Role of Social Norms and Anti-Elitism.

Finally. For record keeping purposes, please tell us the type of study you are pre-registering.

Survey Experiment.