

Misinformation on social media: Investigating motivated reasoning through an identity-protection model

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Philosophie (Dr. phil.)

genehmigte kumulative Dissertation

von

Magdalena Wischnewski
aus
Olpe

Gutachterin 1: Prof. Dr. Nicole Krämer

Gutachter 2: Prof. Dr. Matthias Brand

Tag der mündlichen Prüfung: 27.10.2021

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

ub | universitäts
bibliothek

Diese Dissertation wird via DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt und liegt auch als Print-Version vor.

DOI: 10.17185/duepublico/75400

URN: urn:nbn:de:hbz:465-20220308-103029-8

Alle Rechte vorbehalten.

ACKNOWLEDGEMENTS

This thesis would not be what it is today without my wonderful colleagues, fantastic friends, and an utmost supporting family. First and foremost, I thank Prof. Dr. Nicole Krämer for seeing merit in my initial thoughts on motivated reasoning and helping me to develop those ideas. You gave me the freedom to explore and gently redirected me whenever needed. Most importantly, I want to thank you for the trust you bestowed in me, even when that meant letting me wander off around the globe.

I would also like to thank Prof. Dr. Matthias Brand. You were there from the beginning when I interviewed for the position, throughout my first attempts of writing a paper, and until the final stretch of this PhD project. Thank you for all the time you took to guide me through this endeavor.

I am grateful for my colleagues with whom I had the privilege of working with. Thank you to former and present colleagues from SP at the University of Duisburg-Essen and the UCSM research training group. Similarly, I owe the deepest gratitude to my colleagues from the DMRC at QUT. A shout out to my fantastic co-authors: Rebecca Bernemann, Axel Bruns, Tobias Keller, Thao Ngo, and Martin Jansen. Especially Thao, who has become my PhD-sister with whom I could discuss all my joys and frustrations that come along with pursuing a PhD—thank you!

I am indebted to my family and friends who have been there for me and continue to be loyal and steadfast: Dhanyavad and danke! Your patience, love, and care helped me cross this milestone. I am sorry for the times I was a negligent daughter or friend, too stuck up with this work and not seeing beyond it!

I dedicate this thesis to my grandparents, Margot and Günter, Helga and Willi, who showered me in their love. To my father, Andreas, who sowed in me the curiosity it takes to choose this path. To my mother, Karin, whose sacrifices and endurance got me where I am today. And to my partner, Smith, who is my sharpest critique, safe harbor, and most loving companion. We two are, after all, “two explorers in one boat, in a wild sea, who met at an irrelevant point in their voyage, who smile, talk, and share stories of their lives, who look into each other’s eyes—reflecting on how similar, dissimilar and yet powerful they can be as one”.

ANNOTATION OF THE PAPERS INCLUDED IN THE CUMULUS

Study 1

Wischnewski, M., Bruns, A. & Keller, T. (2021). Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter. *Digital Journalism*, 9(5), 549–570. <https://doi.org/10.1080/21670811.2021.1903960>.

Study 2

Wischnewski, M., Bernemann, R., Ngo, T., & Krämer, N. (2021). Disagree? You must be a bot! How beliefs shape Twitter profile perceptions. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan*. New York, NY: ACM. <https://doi.org/10.1145/3411764.3445109>

Study 3

Wischnewski, M., Ngo, T., Bernemann, R., Jansen, M., & Krämer, N. (2022). “I agree with you, bot!” How users (dis)engage with social bots on Twitter. *Manuscript under review for publication in the journal New Media & Society*, online first. <https://doi.org/10.1177/14614448211072307>

Study 4

Wischnewski, M. & Krämer, N. (2020). I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption. In *Proceedings of the 11th International Conference on Social Media and Society*, 148–155. <https://doi.org/10.1145/3400806.3400824>

Study 5

Wischnewski, M. & Krämer, N. (2021). The role of emotions and identity-protection cognition when processing (mis) information, *Technology, Mind, and Behavior*, 2(1). <https://doi.org/10.1037/tmb0000029>

ABSTRACT

In recent years, the unprecedented dissemination of misleading or false information on social media platforms has become a public concern by posing fundamental threats to democratic political systems. The increased dissemination of such misinformation has intensified research efforts to understand how psychological mechanisms facilitate this misinformation dissemination. One answer comes from motivated reasoning, suggesting that information is not always processed evenly but to maintain or protect existing attitudes, beliefs, or identities (Kunda, 1990). For the context of misleading or false information, motivated reasoning proposes that content confirming a person's view (congruent) is quickly passed on without further questioning, whereas content that contradicts a person's view (incongruent) is more likely to be identified as false.

This cumulative doctoral dissertation aims to further explore the relationship of misinformation on social media with motivated reasoning. To this end, two broader strategies are applied: First, the empirical effects of motivated reasoning on misinformation sharing are scrutinized. In Study 1, it is tested whether motivated reasoning can explain sharing of hyper-partisan news content on Twitter. By collecting data directly from Twitter, this observational study confirms a sharing process driven by motivated reasoning. Similarly, Study 2 and 3 tested whether motivated reasoning can explain users' perception and engagement with automated accounts, so-called social bots, on Twitter. Results of both studies indicated that users' perceptions are, as predicted, biased. Users perceive congruent accounts as more human-like and incongruent accounts as more bot-like. In addition, while users mostly ignore incongruent accounts, independent whether bot or human-run, congruent accounts that behave like social bots are less likely to perceive engagement.

Consolidating the effects of motivated reasoning through empirical data in the first three studies, in a second step, the underlying psychological processes of motivated reasoning are investigated in Study 4 and 5. In both studies, an identity-centric model of motivated reasoning is examined, focusing on identity threat/affirmation and cooccurring emotional reactions that contribute to motivated reasoning. The findings of both studies yield mixed results. While both studies show that emotions are involved in identity-protection cognition, clear inferences about the precise role of emotions cannot be made. Summarizing all results, a refined model of motivated reasoning as identity-protection cognition is introduced.

Taken together, the results of all five studies extend previous findings of motivated reasoning and contribute to a better understanding of how motivated reasoning affects misinformation on social media.

ZUSAMMENFASSUNG

Die Verbreitung irreführender oder falscher Informationen auf Social-Media-Plattformen wurde in den letzten Jahren zunehmen politisch und öffentlich diskutiert, da sie demokratische Systeme grundlegend bedroht. Diese zunehmende Verbreitung solcher Fehlinformationen hat Forschungsbemühungen intensiviert, die darauf abzielen zu verstehen, wie psychologische Mechanismen diese Verbreitung von Fehlinformationen erleichtern. Eine Theorie, dies zu erklären, ist motivierte Kognition. Mit motivierter Kognition wird eine Informationsverarbeitung bezeichnet, deren Ziel es ist, bestehende Meinungen zu bestätigen, anstatt akkurate Schlüsse zu ziehen (Kunda, 1990). Für den Kontext von Fehlinformationen bedeutet dies, dass Inhalte, die die Meinung einer Person bestätigen (kongruent), schnell ohne weitere Hinterfragung weitergegeben werden, während Inhalte, die der Meinung einer Person widersprechen (inkongruent), eher als falsch identifiziert werden.

Diese kumulative Dissertation hat zum Ziel den Zusammenhang von Fehlinformationen in sozialen Medien mit motivierter Kognition zu untersuchen. Zu diesem Zweck wurden zwei Strategien verfolgt: Zunächst wurden die Auswirkungen motivierter Kognition auf das Teilen von Fehlinformationen auf Social Media untersucht. In Studie 1 wurde getestet, ob motivierte Kognition das Teilen von hyper-partisan Nachrichteninhalten auf Twitter erklären kann. Dazu wurden Twitter-Daten ausgewertet, die einen Sharing-Prozess angetrieben durch motivierte Kognition bestätigen. In Studie 2 und 3 wurde getestet, ob motivierte Kognition die Wahrnehmung von und die Interaktion mit automatisierten Accounts, sogenannten Social Bots, auf Twitter erklären kann. Die Ergebnisse beider Studien zeigten, dass die Wahrnehmungen der Nutzer, wie vorhergesagt, verzerrt sind. Benutzer nehmen kongruente Konten als menschenähnlicher und inkongruente Konten als Bot-ähnlicher wahr. Während Benutzer inkongruente Konten, unabhängig davon, ob sie von Bots oder Menschen geführt werden, meist ignorieren, nehmen kongruente Konten, die sich wie soziale Bots verhalten, weniger wahrscheinlich Engagement wahr.

Durch die empirische Bestätigung des Einflusses von motivierter Kognition der ersten drei Studien, wurden in einem zweiten Schritt in Studie 4 und 5 die zugrunde liegenden psychologischen Prozesse von motivierter Kognition untersucht. In beiden Studien wird ein identitätszentriertes Modell der motivierten Kognition mit Fokus auf Identitätsbedrohung/-bestätigung und gleichzeitig auftretende emotionale Reaktionen getestet. Die Ergebnisse beider Studien führen zu gemischten Ergebnissen. Obwohl beide Studien zeigen, dass Emotionen an der Identitätsschutzkognition beteiligt sind, können keine klaren Rückschlüsse auf die genaue Rolle von Emotionen gezogen werden.

Zusammengenommen erweitern die Ergebnisse aller fünf Studien frühere Erkenntnisse über motivierte Kognition und tragen zu einem besseren Verständnis dafür bei, wie motivierte Kognition Fehlinformationen in sozialen Medien beeinflusst.

TABLE OF CONTENTS

Acknowledgements	I
Annotation of the papers included in the cumulus	III
Abstract	IV
Zusammenfassung	V
1. Introduction	1
2. Theoretical Background	3
2.1 Misinformation: A definition.....	3
2.2 Misinformation and social media	4
2.3 A psychological perspective to misinformation–motivated reasoning	6
2.3.1 What is motivated reasoning?	7
2.3.2 Previous empirical findings connecting misinformation and motivated reasoning... 9	9
2.4 Motivated reasoning affecting news-sharing, perceptions, and interactions on social media	10
2.4.1 Hyper-partisan news-sharing through the lens of motivated reasoning.....	10
2.4.2 Social bots through the lens of motivated reasoning	12
2.5. The psychology of motivated reasoning.....	13
2.5.1 Motivated reasoning and misinformation: The central role of identity	14
2.5.2 The theory of identity-protection cognition	17
2.5.3 Emotions in motivated reasoning.....	18
2.5.4 Combining emotions and identity-protection cognition	21
2.5.5 Identity-protection or cognitive sophistication?	23
2.5.6 Preventing misconceptions through identity-salience manipulations.....	24
2.6 Conclusion and research objectives.....	25
3. Summary of the research papers included in the cumulus	29
3.1 Study 1: Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter	29
3.2 Study 2: Disagree? You must be a bot! How beliefs shape Twitter profile perceptions	30
3.3 Study 3: “I agree with you, bot!” How users (dis)engage with social bots on Twitter ..	32
3.4 Study 4: I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption.....	33
3.5 Study 5: The role of emotions and identity-protection cognition when processing (mis)information	34
4. Discussion	36
4.1 Overview of the findings	36
4.1.1 Motivated reasoning and misinformation on social media	36

4.1.2 Connecting identity-protection cognition and emotions.....	38
4.1.3 Motivated reasoning as identity-protection?.....	40
4.2 Theoretical implications	42
4.2.1 Content-emotions and identity-emotions.....	44
4.2.2 Identity and identity-constituting attitudes in identity-protection cognition	46
4.2.3 Credibility cues and identity-protection cognition	48
4.2.4 The influence of processing style on identity-protection cognition.....	50
4.3 Practical implications	51
4.4 Limitations and future studies	56
4.4.1 Emotions	56
4.4.2 Identity threat/affirmation	58
4.4.3 Individual differences	59
4.4.4 Alternative explanations	60
4.4.5 Motivations and processing style.....	61
4.5 Conclusion	63
5. References	64
Appendix	82

TABLE OF FIGURES

Figure 1: Elicitation of anger, anxiety, and enthusiasm as the result of identity-threatening or identity-affirming (mis)information.....	22
Figure 2: The mediation model of identity-protection including the moderating role of cognitive sophistication.....	24
Figure 3: Visualization of the the contribution of each study in relation to the two research questions.....	28
Figure 4: An updated model of identity-protection cognition..	43
Figure 5: Content emotion and identity emotion in identity-protection cognition.	45
Figure 6: Identity strength moderating the (mis)information identity threat/affirmation link.	47
Figure 7: Credibility cues in identity-protection cognition.....	49
Figure 8: Effects of cognitive sophistication and task affordances on the processing style. ..	51

1. INTRODUCTION

The digitalization of mediated communication has enabled quicker and broader information sharing than ever before. It has also paved the way for unprecedented dissemination of misleading or false information, typically described as misinformation (Del Vicario et al., 2016; Egelhofer & Lecheler, 2019; Lazer et al., 2018). Previous studies specifically highlight the role of social media platforms to explain the increased circulation of misinformation (Del Vicario et al., 2016; Tucker et al., 2018). On the one hand, social media platforms such as Facebook, Twitter, YouTube, Instagram, or Reddit, lack editorial verification and, thereby, facilitate the creation of (false) content through “a direct path from producers to consumers of content” (Del Vicario et al. 2016, p. 554). On the other hand, social media’s network structures allow for rapid dissemination of (false) content (Vosoughi et al., 2018). This increased circulation of misinformation has detrimental consequences for individuals (Kim et al., 2020; Oyeyemi et al., 2014; Radzikowski et al., 2016).

Most prominent cases of misinformation and its negative consequences relate to political communication (Allcott & Gentzkow, 2017; Bessi & Ferrara, 2016; Schäfer et al., 2017). In particular, many authors suggest that misinformation and subsequent misperceptions pose a fundamental threat to democratic political systems. For example, Zimmermann and Kohring (2020) found that misinformation affected the parliamentary election outcomes in Germany, 2017. Collecting panel data, the authors found that believing misinformation impacted voting choices and corroded trust in news media and politics.

The increased effects of misinformation on individuals have intensified studies on how psychological mechanisms facilitate misinformation dissemination. One promising direction refers to *motivated reasoning*. Motivated reasoning suggests that information is not always processed to arrive at an accurate conclusion but to maintain or protect existing attitudes, beliefs (Kunda, 1990), or identities (Kahan et al., 2007). In the context of misinformation, motivated reasoning hypothesizes that misinformation confirming one’s view (congruent) is quickly passed on without further questioning, whereas misinformation that contradicts one’s view (incongruent) is more likely to be questioned and scrutinized.

While previous findings on motivated reasoning could elucidate its effects on public opinion formation (Strickland et al., 2011), economic perceptions (Bisgaard, 2015), science denial (Washburn & Skitka, 2017), as well as political information seeking and processing (Druckman et al., 2016; Peterson & Allamong, 2021; Taber & Lodge, 2006), fewer studies have empirically examined the hypothesized effect of motivated reasoning on political

misinformation. For instance, Ribeiro et al. (2017) investigated whether partisanship predicted how users on Twitter labelled news with tags such as #FakeNews or #AlternativeFacts. In line with motivated reasoning, Ribeiro et al. (2017) found that labelling something as #FakeNews or #AlternativeFacts was more likely when the news item was incongruent to one's partisanship. Similarly, Anthony and Moulding (2019) found that misinformation that was congruent with participants' political beliefs was more likely to be accepted than misinformation that was incongruent. Considering the limited number of empirical studies examining motivated reasoning and political misinformation, I address this gap by asking:

RQ1: How does motivated reasoning affect misinformation on social media?

To answer RQ1, this thesis examines the contribution of hyper-partisan news outlets and automated social media accounts, so-called social bots. In three studies, it was first examined whether motivated reasoning can explain hyper-partisan news sharing. Second, it was experimentally investigated whether motivated reasoning can affect users' perceptions and engagement with social bots. In applying motivated reasoning to perceptions and behavior on social media, theoretical assumptions were tested, and context-specific variables, such as previous knowledge about deceptive tools, credibility cues, and interindividual differences, were explored.

While knowing how motivated reasoning affects misinformation on social media and how these effects play out, it is also important to examine its underlying psychological processes to develop adequate countermeasures. To this end, this thesis examines, firstly, the effects of emotions in motivated reasoning. Previous investigations concerning motivated reasoning have primarily focused on cognitive and motivational processes driving motivated reasoning, and less on the influence of emotions. With being tailored to appeal to specific emotions (Bakir & McStay, 2018), emotional processes are, however, particularly relevant in the context of misinformation. Secondly, previous research has connected motivated reasoning to various constructs such as attitudes, partisanship, and identities but is lacking one unifying framework. Consequently, in a second research question, I address both gaps by asking:

RQ2: How can emotions and identity explain motivated reasoning?

To this end, I theoretically link the experience of emotion to the experience of identity-threat and identity-affirmation. In doing so, I connect previous insights on identity-protection cognition (Kahan, 2017) with emotional processes. To answer RQ2, I build on and extend previous theoretical models of motivated reasoning, which have included cognitive (Lord et al., 1979; Mercier, 2016; Nickerson, 1998), motivational (Chaiken et al., 1996; Kunda, 1990), and, to a limited extent, emotional processes (Suhay & Erisen, 2018; Lodge & Taber, 2000; Weeks,

2015), as well as identity-based explanations (Kahan et al., 2011; Kahan, 2017; Van Bavel & Pereira, 2018).

Overall, this thesis contributes to a better understanding of the effects of motivated reasoning on misinformation circulation on social media by (1) identifying three specific cases in which motivated reasoning affects misinformation circulation and by (2) elucidating the underlying psychological processes leading to motivated reasoning, particularly, the contribution of emotions related to identity-threat and identity-affirmation.

2. THEORETICAL BACKGROUND

2.1 Misinformation: A definition

The term ‘fake news’ has been applied in many contexts and has been associated with different phenomena. For example, fake news has been used to describe news satire and parody—content low in factuality but produced to entertain and not deceive (Nielsen & Graves, 2017; Wardle & Derakhshan, 2018). In contrast, fake news has also been used as a derogatory buzzword to discredit mainstream media (Egelhofer & Lecheler, 2019). Most commonly, however, the term fake news is used to describe content with varying levels of factuality: (1) factually correct but misleading information, (2) addition or deletion of information, and (3) complete fabrication (Quandt et al., 2019). In addition to low levels of factuality, such fake news are also characterized by an attempt to come under the guise of a journalistic format, copying structural elements from mainstream media such as headlines, news text, and pictures (Allcott & Gentzkow, 2017; Horne & Adali, 2017; Lazer et al., 2018). To differentiate such content of low factuality from satire and a derogatory buzzword, many scholars refer to *misinformation* or *disinformation*. While the former usually refers to unintentionally false content, the latter infers an intention of deceit. Although it is essential to differentiate between fabrication and dissemination intentions, for example, from a legal perspective, intentions do not affect the consequences of mis- or disinformation. Individuals, who are presented with false content, are equally likely to believe unintentionally and intentionally shared misinformation which is why I solely refer to misinformation in this thesis.

With factuality and format constituting misinformation, it is often also related to other concepts such as conspiracy theories, rumors, and hyper-partisan media. Conspiracy theories allegedly uncover and connect events and public figures to a greater "machination of powerful people, who attempt to conceal their role" (Sunstein & Vermeule, 2009, p. 205). In doing so, conspiracy theories rely on questionable or unsupported sources and information and use

oversimplifications to provide a narrative that explains complex connections. The resulting narrative frequently incorporates misinformative content, which allegedly proves the conspiracy's claims. However, some conspiracy theories have also been found to be true.

In contrast, rumors do not offer a whole narrative but promote information with insufficient evidence of its veracity. Rumors are often circulated in events of threat or uncertainty and are characterized by rapid social transmission (DiFonzo & Bordia, 2007). Similar to conspiracy theories, rumors can turn out to be true.

Conspiracy theories and rumors are often promoted by fringe or hyper-partisan media (Faris et al., 2017). These outlets portrait mainstream journalism as hegemonic and biased, to which they provide an *alternative* news source, arriving at so-called partisan alternative journalism (Benkler et al., 2018). Critically, hyper-partisan news often contain at least a kernel of truth, decontextualized to the degree that it creates an entirely misleading view of the world (Faris et al., 2017) (see more in Chapter 2.4.1).

Having established a definition of misinformation, the following section connects the recent rise of misinformation dissemination with technological advances in mediated communication. In doing so, I focus on research connecting misinformation dissemination to social media platforms.

2.2 Misinformation and social media

In 2017, the American Dialect Society¹ and Collins Dictionary² voted 'fake news' word of the year, signaling "an anecdotal indicator of the general popularization of the term during that time" (Quandt et al., 2019, p. 1). With its newly gained popularity, it might be assumed that misinformation³ is a modern invention. However, misinformation can be found throughout history. To this end, Burkhardt (2017) aptly argues that misinformation has been around "as long as humans have lived in groups where power matters" (p. 5). In her essay, she connects misinformation to means of communication that have evolved throughout history, ranging from early word of mouth, the invention of the printing press to mass media, arriving at today's internet era. Burkhardt (2017) concludes that, while misinformation is not a new phenomenon, the recent upsurge of misinformation is attributed to "a vastly increased scale" (p. 6) of dissemination through the internet. Similarly, connecting the development of the internet to the upsurge of misinformation, Tandoc et al. (2017) argue that "[n]ow that online platforms,

¹ <https://www.americandialect.org/fake-news-is-2017-american-dialect-society-word-of-the-year>

² <https://blog.collinsdictionary.com/language-lovers/etymology-corner-collins-word-of-the-year-2017/>

³ While it was the term „fake news” that was included in both dictionaries, I use the term misinformation to differentiate misleading content from satire and political smear campaigns (see previous paragraph).

particularly social media, are becoming the main sources of news for a growing number of individuals, misinformation seems to have found a new channel" (p. 2).

It is argued that social media's distinct affordances, such as limited content moderation, network structures, and popularity metrics, are catalysts for misinformation (e.g., Tucker et al., 2018). On social media, anyone with an account can produce and share information. While this horizontal communication enabled greater citizen engagement, it also made way for *dark participation* (Quandt, 2018), such as the distribution of misinformation. Because of minimal editorial verification and journalistic standards (also called *disintermediation*, see Del Vicario et al., 2016), misinformation can easily be fabricated and made publicly available. While social media platforms engage in user-generated content moderation and disperse accounts that share misinformation on a large scale to limit the reach of misinformation, media and communication scholars argue that these interventions operate "under a complex web of nebulous rules and procedural opacity" (Roberts, 2018, Commercial content moderation and the logic of opacity section, para.1).

Moreover, while one side calls for stronger regulations of what can be shared on social media, the other side argues for fewer restrictions under the umbrella of freedom of speech. However, Roberts (2018) argues that, ultimately, platforms are guided not by political and democratic standards but by revenue generation. Therefore, if misinformation is profitable, she argues, platforms are less likely to interfere through content moderation.

In addition to content production without editorial verification, social media's networked structures allow users to share misinformation not only with close ties (direct connections) but beyond users' physical and social proximity (indirect connections). Once misinformation is shared by one user, it can spread through the platform's unique engagement options such as likes, shares/retweets, or comments. Depending on an account's connectedness (centrality), misinformation can spread far and quickly. Different approaches have been used to explain how misinformation spreads through these ties by applying insights from social network theory (Borgatti et al., 2009) and diffusion models (Schubert et al., 2021). Vosoughi and colleagues (2018), for example, compared the spread of true with false information on Twitter. They found that incorrect information spread "farther, faster, deeper, and more broadly than the truth" (p. 5). One of the reasons explaining these results is that many misinformation stories employ emotional language to generate attention (Bakir & McStay, 2018), which has increased content diffusion on social networks (Brady et al., 2017).

Once information is shared, social media's popularity metrics such as Facebook's *like* or Twitter's *retweet* functions increase visibility within the network and credibility judgments.

For example, Winter and colleagues (2015) found that negative comments decreased the persuasive influence of news shared on social media. The authors also found that the effect of comments depended on the personal relevance of the topic for an individual. This effect of popularity metrics is exacerbated by accounts that are no longer run by a human user but have been programmed for specific purposes like content amplification—so-called social bots (Howard & Kollanyi, 2016) (see also Chapter 2.4.2). Moreover, it has been argued that the personalization of social media platforms, as well as algorithmic curation, amplify the effects of misinformation. With social media feeds but also search engines, such as Google, and e-commerce platforms, such as Amazon, remembering users' search histories, users can find themselves quickly in an amplified misinformation environment (Juneja & Mitra, 2021).

To sum up, social media's affordances such as limited content moderation, network structures, popularity metrics, and algorithmic curation contribute to the rise of misinformation. While these factors promote a misinformation ecosystem, they do not explain *why* individuals accept and share misinformation. To understand why individuals fall prey to misinformation, a psychological perspective needs to be taken which is offered in the next section.

2.3 A psychological perspective to misinformation—motivated reasoning

The previously described increased circulation of misinformation, facilitated by today's new media landscape, has resulted in a proliferation of academic studies trying to understand how individuals contribute to the upsurge of misinformation. In particular, many scholars are interested in *why* individuals succumb to misinformation to find ways to implement countermeasures.

Answers to why individuals succumb to misinformation fall into three broader categories. First, when reading the news or participating in a conversation, individuals assume information to be truthful and relevant unless contextual or content cues like source credibility or plausibility evaluations indicate otherwise (Gilbert et al., 1993). In other words, any misinformation is generally more likely to be accepted instead of rejected. Previous research has connected this initial truth assumption to Bayesian inferences, arguing that most everyday life information/situation is more likely to be mundane and true, indicating a higher base rate for true events (Brashier & Marsh, 2020). Hence, it is rational to assume truthfulness when encountering new information.

A second approach suggests that misinformation can go unnoticed because of cognitive biases such as *illusory truth*, which refers to empirical findings indicating that repetition of statements increased their plausibility and accuracy ratings (Brashier & Marsh, 2020; Dechêne

et al., 2010). Explaining empirical findings of illusory truth, previous works point to the effects of processing fluency, where “people learn that fluency typically leads to the correct judgment in less time than other strategies” (Brashier & Marsh, 2020, p. 504). Because repeated exposure to misinformation increases its processing fluency, accuracy ratings increase.

The third approach follows what the first and second already suggested. While credibility and plausibility judgments are decisive when judging the veracity of information, the application of these mechanisms can be distorted by biases such as illusory truth. Different than illusory truth, some biases include a motivational component which distorts information processing. One such bias is described as motivated reasoning which suggests that information is not always processed evenly but to maintain or protect existing attitudes, beliefs (Kunda, 1990), or identities (Kahan et al., 2007). According to motivated reasoning, any incoming information that threatens a core attitude or identity is more likely to be rejected. In contrast, incoming information that affirms a core attitude or identity is more likely to be accepted. Transferring this to why people fall for misinformation, motivated reasoning suggests that misinformation confirming one’s view (congruent) is passed on without further questioning, whereas misinformation that contradicts one’s view (incongruent) is more likely to be questioned and scrutinized.

In the following two sections, I introduce motivated reasoning as the central theory of this thesis, focusing on theoretical milestones leading to today’s conception of motivated reasoning. This is followed by a section on empirical results connecting motivated reasoning to misinformation.

2.3.1 What is motivated reasoning?

The idea that individuals tend to interpret incoming information to maintain prior attitudes and beliefs developed from early philosophical understandings to contemporary theoretical models and empirical findings. In his famous quote from 1620, Francis Bacon claims that “the human understanding, when it has once adopted an opinion (...) draws all things else to support and agree with it”. Later empirical findings support his view. For example, Lord et al. (1979) found in their seminal study on biased assimilation and attitude polarization that arguments in favor or against the death penalty were evaluated depending on the individuals’ prior attitude. The authors observed that proponents perceived arguments supporting their view as stronger and more credible, whereas opponents evaluated the same arguments as less strong and less credible. While relying on past knowledge and experiences is indispensable to human reasoning, Lord et al. (1979) point out that the detrimental effects of biased assimilation arouse

from individuals' "readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially 'justified' the processing bias" (p. 2107).

In an effort to arrive at a theoretical explanation, which summarizes previous observations and explains why motivated reasoning can be found, Kunda (1990) suggests an interplay of cognitive and motivational processes. She argues that, while any reasoning is motivated, the specific reasoning-motivation is decisive for the reasoning outcome. Such reasoning-motivations direct reasoning to either pursue an accurate conclusion or a preconceived conclusion, which Kunda (1990) coined as *accuracy goals* and *directional goals*.

Given sufficient cognitive capacity and motivation, Chaiken (1987) argues that accuracy goals are individuals' default motivation. Moreover, accuracy goals are commonly activated when individuals expect strong negative consequences for being wrong (e.g., personal harm) or when being accurate is more attractive (e.g., receiving financial incentives). As a consequence of accuracy goals, individuals engage in deeper and more careful task deliberation, spend more time to arrive at an accurate conclusion, consider more alternatives (Tetlock & Kim, 1987), and overcome cognitive biases such as primacy or anchoring effects (Kruglanski & Freund, 1983) and the fundamental attribution error (Tetlock, 1985). However, in some cases, accuracy goals can worsen the reasoning outcome as they exacerbate bias. For example, when given non-diagnostic information in an inference task, individuals driven by accuracy goals were more likely to wrongly include this information (Tetlock & Boettger, 1989). In this case, the more complex processing due to accuracy motivations led to a poorer reasoning outcome. However, a more prevalent cause for faulty reasoning outcomes are directional goals.

In contrast to accuracy goals, directional goals are activated when individuals "want to arrive at a particular conclusion" (Kunda, 1990, p. 484). In turn, the desire for such a preconceived conclusion may arise from perceptions of a threatened identity or attitude (defense motivation) or the need to make a desirable impression on others (impression motivation) (Chaiken et al., 1996). As a consequence of directional goals, Lord et al. (1979) suggest that individuals engage in more peripheral information processing and produce more counterarguments or discount disconfirming evidence (Klayman & Ha, 1987). These consequences can be worsened by contextual factors that facilitate heuristic processing like time pressure, a complex reasoning task (Chaiken et al., 1989), and weak consequences of being wrong (Tetlock, 1985). Kunda (1990) concludes that directional motives "influence which beliefs and rules are accessed [from memory] and applied on a given occasion" (p. 485). In other words, she argues that motivated reasoning is essentially what it claims to be: a motivational selection of cognitive reasoning strategies to suit one's attitude and identity.

2.3.2 Previous empirical findings connecting misinformation and motivated reasoning

Empirical work testing the effects of motivated reasoning on misinformation supports the hypothesis that motivated reasoning makes individuals more vulnerable to misinformation (Peterson & Iyengar, 2021). Researchers consistently find that misinformation are more likely to be believed if they are congruent to participants' beliefs and preferences (Anthony & Moulding, 2019; Ecker et al., 2014; Kahne & Bowyer, 2017; Kuklinski et al., 2000; Nyhan & Reifler, 2010). Testing whether these effects can partially be explained by source credibility, Clayton et al. (2019) found that source cues contribute less to the observed effects. The authors showed participants false and true news originating either from CNN (Democrat congruent; Republican incongruent source), FoxNews (Democrat incongruent, Republican congruent), or without a source. Unlike previous results on the impact of source congruency (Bolsen et al., 2014), only the content of the news could explain which news were believed to be false or real, whereas the source did not matter. Similarly, the prominence of a source, for example, legacy media versus unknown news forums, showed no effect on truth judgments (Clemm von Hohenberg, 2019; Tsang 2020). In light of previous findings which indicate that more credible sources are generally more likely to be persuasive (e.g., Flanagin & Metzger, 2000), these findings might implicate that, in recent years, legacy media has seen a decrease in trust and credibility (see also Turcotte et al., 2015).

Another strand of research investigates how individual differences shape these processes. While previous research on motivated reasoning indicates that individuals with greater cognitive sophistication skills show greater bias (Taber & Lodge, 2006), newer evidence suggests this is not the case. For example, Pennycook and Rand (2019) could show that individuals with greater cognitive sophistication skills could better differentiate between false and accurate news, leading the authors to suggest that individuals are not biased in their perception but are cognitive misers (see in more detail Chapter 2.5.5). Moreover, other research suggests that political knowledge and media literacy skills affect how strongly motivated reasoning affects truth judgments. For example, Kahne and Bowyer (2017) could show that students with greater political knowledge were more biased, and, in contrast, students with greater media literacy skills were less biased.

Psycho-physiological data confirm these results: Moravec and colleagues (2018) tested behavioral and neurological reactions towards congruent and incongruent misinformation. Their results support the motivated reasoning hypothesis. Through EEG data, they found that, when misinformation was congruent with an individual's opinion, that information received increased cognitive attention, making users more likely to fall for misinformation. However,

Moravec and colleagues (2018) also point out that incongruent misinformation receives little to no attention, suggesting that motivated reasoning protects against misinformation.

While all results cited above support the hypothesis that motivated reasoning influences misinformation, most studies do not go beyond the question of credibility and veracity judgments. However, motivated reasoning can affect misinformation beyond the question of false or true. For example, Pennycook et al. (2021) asked participants to rate news headlines concerning their veracity and asked how likely participants would share these. Results indicate that some participants would share news headlines which they perceive as inaccurate. Pennycook's et al. (2021) results stress how important it is to differentiate between veracity judgements and the actual news sharing. Hence, I suggest moving one step further, scrutinizing the role of individuals in *sharing* misinformation on social media and how motivated reasoning affects this sharing process.

2.4 Motivated reasoning affecting news-sharing, perceptions, and interactions on social media

Once misinformation is created and entered into social media, it can be shared with and by other accounts of the network, opening the possibility of misinformation going viral (e.g., Tambuscio et al., 2015). These sharing entities can be both human users but also automated social media accounts, so-called social bots. In the following two subsections, I discuss how both entities contribute to the proliferation of misinformation and suggest how motivated reasoning alters these sharing processes.

2.4.1 Hyper-partisan news-sharing through the lens of motivated reasoning

Hyper-partisan news media describes news outlets that position themselves outside or at the fringes of traditional media systems, offering so-called *alternative facts* (Figenschou & Ihlebæk, 2019). Essential for this thesis, hyper-partisan news media have been identified as a central source of misinformation as well as its circulation (Faris et al., 2017). By going further than promoting completely fabricated misinformation, hyper-partisan news media have been accused of decontextualizing truth, repeating falsehoods, and leaps of logic to create a fundamentally misleading view of the world (Faris et al., 2017). Because news stories often contain a kernel of truth, debunking hyper-partisan news is more complex than completely fabricated claims. Moreover, different hyper-partisan news media act in cross-citation and cross-linking networks, creating propaganda networks that amplify and echo misleading information, making it even more difficult to retract and debunk information (Faris et al., 2017). While conventional news outlets on social media still outnumber hyper-partisan news media,

hyper-partisan outlets profit from social media's platform affordances and increase, especially among right-wing partisans in the USA, their readership (Faris et al., 2017).

With hyper-partisan news as a central medium for the generation and circulation of misinformation and misleading information online, in this thesis, the concept of *shareworthiness* is joined with motivated reasoning theory to understand hyper-partisan news media dynamics. In doing so, extrinsic content characteristics (shareworthiness factors) that influence the share-likelihood are combined with intrinsic user motivations that also influence the share-likelihood.

Shareworthiness extends from the earlier concept of newsworthiness which introduced specific news factors that determine which events are most likely to be reported (Galtung & Ruge, 1965; Östgaard, 1965). Such factors, like unexpectedness, reference to individuals (human interest), proximity, or the presence of a conflict, have been shown to influence journalists' selection of news coverage as well as audience selection and preferences alike (Eilders, 2006). Trilling and colleagues (2017) argued that shareworthiness factors drive news-sharing in online contexts similar to newsworthiness. The authors found that the shareworthiness factors *geographical* and *cultural distance*, *negativity*, *positivity*, and *conflict* predicted sharing on Facebook and Twitter. Similarly, Valenzuela and colleagues (2017) successfully applied the concept of shareworthiness to news frames and sharing-likelihoods. While shareworthiness research generated valuable insights, it has rarely been applied to hyper-partisan news sharing. As hyper-partisan news display different features than mainstream news and are tailored to a different audience (Holt et al., 2019), it is assumed that a different subset of shareworthiness factors should drive hyper-partisan news sharing (see Study 1 in Chapter 3.1).

In addition to these extrinsic content features that drive news-sharing, *intrinsic* user motives should also drive online (hyper-partisan) news-sharing. As hyper-partisan news is inherently tailored to accustom partisan views and, consequently, individuals who identify with a particular partisanship, predictions according to motivated reasoning are suitable. In line with motivated reasoning, it is suggested that users are more likely to share news that align with their political attitude/identity than to share news that contradict their attitude/identity, intending to affirm their views and bolster their partisanship. Partisanship-contradicting news are more likely to be perceived as a threat to one's partisan identity, leading to, for example, *defensive inattention* (Chaiken et al., 1996).

2.4.2 Social bots through the lens of motivated reasoning

Human users generate not all communication on social media platforms. (Semi-) automated accounts, so-called social bots, run by computer algorithms, populate social networking sites, trying to disguise their automated nature by blending with human activities (such as, e.g., “liking” or “retweeting”). Hence, their automated nature can go unnoticed by users. Functions of these social bots have been described as copy-paste bots (Schäfer et al., 2017), amplifier bots that boost particular sources (Howard & Kollanyi, 2016), or fake followers which boost popularity (Cresci et al., 2015).

While these social bot functions are not necessarily harmful, they have also been employed for fraudulent purposes and have been connected to malicious activities. The resulting detrimental influence of social bots has become public and academic concern, especially in the context of political persuasion. Social bots have been found, for example, to stir political discussion online to promote misinformation (Wang et al., 2018), to engage in political astroturfing (Keller et al. 2020), and to influence election outcomes (Bessi & Ferrara, 2016; Ferrara, 2017; Schäfer et al., 2017).

To gauge the harmful influence of social bots, researchers have employed two different methodological strategies: social bot detection tools and modeling approaches. Both methods make mutually exclusive trade-offs in the two dimensions (1) ecological validity and (2) accuracy. Social bot detection tools, which employ machine-learning models to automatically detect social bot accounts by scrutinizing an accounts’ behavior, such as the timing of postings and social networking structures (see Karataş & Şahin, 2017, for a review of automated detection solution), are high in ecological validity as they use data generated on social media platforms. However, it has recently been found that such tools are characterized by lower accuracy due to false-positive errors (Rauchfleisch & Kaiser, 2020). In contrast, modeling approaches *simulate* social bots’ impact in social networks, frequently employing agent-based modeling to understand the impact of social bots on a network (Keijzer & Mäs, 2021; Ross et al., 2019). Such an agent-based model clearly defines which accounts are social bots and which are human users (scoring high in accuracy). Likewise, the model determines connectedness and similarity between users/bots and the number of users/bots in a network. Nevertheless, neither are all network characteristics known and translated into a set of rules nor is there sufficient data to describe how users interact with social bots, reducing a model’s ecological validity. For example, Ross et al. (2019) implemented two different agent-based models to understand the impact of social bots on the spiral of silence. The author state that bots “had equal influence on the confidence of their neighbours. In reality, bots might be perceived differently from humans”

(p. 14). While Ross et al. (2019) drew meaningful conclusions from both models, uncertainty remains.

To better inform modeling approaches and, ultimately, better comprehend the impact of social bots on humans, researchers need to understand how human users perceive and engage with social bots. However, such research is scarce (Lazer, 2020). Leaning on insights from motivated reasoning, it is proposed that (a) users are biased in their perceptions of social bots and that (b) users engage in a biased manner with social bots. Motivated reasoning offers a helpful perspective because previous research about political communication has shown that individuals' perceptions are frequently biased towards preconceived attitudes and opinions (Druckman et al., 2016; Taber & Lodge, 2006). Hence, if an account shares a user's attitude, the user might become "blind" or overconfident in the authenticity of an account. In contrast, attitude-incongruent accounts might be perceived as more suspicious.

Moreover, previous research investigating users' attitudes towards social bots has found that two-thirds of US Americans have heard about social bots, of which 80% think that social bots serve a malicious purpose (Stocking & Sumida, 2018). If most users perceive social bots as negative, it can be expected that users do not want to be associated with such accounts. However, if an account is part of one's affinity group by sharing one's political attitude, users would be more likely to accept an association with a social bot account. According to motivated reasoning, to avoid such an association, users should react with a defense mechanism. Hence, motivated reasoning proposes that users perceive accounts sharing the users' attitude or prior opinion as less bot-like. Conversely, users perceive accounts they disagree with as more bot-like (see Study 2 in Chapter 3.2).

While a better understanding of users' perception is a first step to understanding the impact of social bots, a similar rationale is applied to understand users' engagement with social bots. It is proposed (a) that users prefer to engage with like-minded accounts and (b) that users prefer to engage with human users over bot accounts. Combining both assumptions, it is concluded that both effects should interact (see Study 3 in Chapter 3.3).

2.5. The psychology of motivated reasoning

In the previous chapters, I have focused on how motivated reasoning affects misinformation in social media. By providing previous empirical findings, I could show that motivated reasoning has mostly been connected to biased veracity judgments, increasing the acceptance of attitude-congruent misinformation. In the next step, I argued to move beyond veracity judgments to scrutinize how motivated reasoning could also shape misinformation

sharing. To do so, two phenomena, which are closely related to misinformation distribution, are examined: hyper-partisan news media and social bots. In addition to previous research on misinformation acceptance, this approach allows to gauge the impact of motivated reasoning on misinformation diffusion, which can deviate from misinformation acceptance, as Pennycook et al. (2021) have found (see also Chapter 2.3.2).

But why investigate the effects of motivated reasoning on misinformation in the first place? The underlying premise here is that to reduce the detrimental implications of misinformation, one needs to understand which mechanisms are likely to affect misinformation diffusion. Confirming that motivated reasoning affects misinformation diffusion is, hence, only one step towards mitigating misinformation diffusion. In a second step, it is essential to understand the underlying psychological processes of motivated reasoning to successfully implement countermeasures. Consequently, having introduced motivated reasoning earlier (see Chapter 2.3.1), in the next section, I specify motivated reasoning as an identity-protective mechanism and introduce an identity-centric model of motivated reasoning.

2.5.1 Motivated reasoning and misinformation: The central role of identity

Different empirical and theoretical reasons suggest that the construct of identity is important to understand both motivated reasoning and misinformation. First, as introduced earlier (Chapter 2.3.1), motivated reasoning describes how individuals react to information incongruent to participants' ideology, partisanship, or previous attitudes. However, the concepts of ideology and partisanship can also be understood as (social) identities (Huddy, 2001; Bankert et al., 2017) and group-memberships (Mason, 2018). Bankert et al. (2017) argue, for example, that partisanship is less based on informed deliberation about a party's performance but is better characterized as an expressive identity of one's support for a party, independent of its performance. Similarly, previous attitudes can constitute core identity elements, depending on how relevant the attitude is for an individual (Howe & Krosnick, 2017). Especially religious, moral, and political attitudes are inherently related to identities (Jones, 1999) and "matter the most for an individual's thoughts, intentions, and behavior" (Howe & Krosnick, 2017, p. 328). In addition, some group-identities arise due to specific attitudes such as in opinion-based groups (Bliuc et al., 2007). Thus, the construct of identity joins the constructs of ideology, partisanship, and previous attitudes.

Furthermore, the construct of identity is central in many empirical studies on misinformation. For example, Mourão and Robertson (2019) found that "outright and sensational fabrications are only a modicum of content produced" (p. 2091) and that misinformation is better defined as (hyper-) partisan viewpoints. On similar lines, Tripodi

(2018) found that susceptibility to misinformation is often not the result of lacking media literacy but identity. She observed that conservatives in the USA “carefully and meticulously constructed a political reality to support Trump’s presidency by relying on media literacy practices taught to them” (p. 19). Moreover, in an integrative review, Trevors (2019) found that misinformation corrections were more likely to be unsuccessful when these misinformation were congruent to individuals’ prior attitudes (supporting a motivated reasoning account). Identifying several mechanisms that promote this intentional correction resistance, Trevors (2019) found that misinformation which is closely associated with an individual’s or a group’s identity are less likely to be successfully corrected.

In addition to empirical reasons, there are several theoretical reasons for an identity approach to motivated reasoning. In her seminal work (see Chapter 2.3.1), Kunda (1990) connects directional reasoning to cognitive dissonance theory. While the original understanding of cognitive dissonance suggests that two (or more) inconsistent cognitive elements induce a psychologically uncomfortable state (Festinger, 1957) which results in the motivation to reduce this dissonance, Kunda (1990) questions the notion that mere inconsistent beliefs induce motivated reasoning. Instead, relying on empirical evidence from Steele (1988), she argues that “dissonance motivation appears to arise from directional goals rather than from inconsistency among beliefs” (p. 492), with “the goal of concluding that one is not a fool” (p.492) and protecting one’s integrity. According to this, a smoker reading a report that shows how smoking kills feels bad because the report implies that the smoker is not a sensible and competent person. Experiencing this uncomfortable feeling of being less sensible and less competent arouses, according to Kunda, the directional goal to restore one’s integrity.

In doing so, Kunda (1990) follows later evidence, suggesting that dissonance only arises when the individual self is engaged and threatened (Aronson, 1968). Hence, according to Kunda’s (1990) account, directional reasoning results from a threatened self. With identity as a “label attributed to the attempt to differentiate and integrate a sense of self along different social and personal dimensions” (Bamberg, 2011, p. 6), I argue that a threat of the self is inherently also a threat of an identity. Consequently, although not directly mentioning it, Kunda (1990) makes way for an identity-centric model of motivated reasoning by posing cognitive dissonance as the source of motivated reasoning.

In contrast, Lodge and Taber (2000) understand motivated reasoning as an unconscious, biased memory-retrieval and develop a tripartite theory of motivated reasoning. According to Lodge and Taber (2000), past evaluations, also denoted as preexisting knowledge or prior attitudes, about any social object (e.g., issues, political leader, groups) are stored in long-term

memory. In other words, any social object in the long-term memory is affectively charged which has also been labeled as *hot cognition*. Through on-line processing, objects stored in long-term memory can be activated and are brought to the working memory, including their affective label. Hence, according to Lodge and Taber (2000), the affect attached to any cognitive object induces the “how-do-I-feel?” heuristic, which, in turn, colors any reasoning about the object. The result is that “decision makers would know which way they would like the new evidence to point at the very moment they are evaluating it” (p. 196), leading to directional reasoning.

However, in their model, Lodge and Taber (2000) (see also, Taber & Lodge, 2006, 2016) do not explain how individuals arrive at evaluations or attitudes which, in turn, affectively charge social objects. With hot cognition and the “how-do-I-feel?” heuristic as central components of their model, this step is, however, crucial to understand why we find motivated reasoning. To answer this, we need to consider previous research on attitude formation, which suggests that attitudes serve a functional value for the individual but can also be based on social identities (Jones & Regan, 1974). As social identities, they are “grounded in the groups we belong to, and they serve to define and proclaim who we are in terms of our relationship to others who are members of the same or different groups” (Smith & Hogg, 2008, p.338). Hence, although not explicitly stating, Lodge and Taber’s (2000) work is also grounded in matters of self and identity, in a way that identities inform attitude formation, which, in turn, affects downstream processing.

Studies about identity-protection cognition (Kahan et al., 2007; Kahan et al., 2011; Kahan, 2017; Van Bavel & Pereira, 2018) position motivated reasoning firmly in identity processes. As the terminology of identity-protection cognition suggests, the notion of *identity protection* is central to this approach. In works by Kahan and colleagues, social and cultural identities become associated with attitudes and vice versa. Consequently, to have a specific attitude implies to be part of a specific social identity (for an in-depth theoretical account, see the following Chapter 2.5.2). Similarly, Van Bavel and Pereira (2018) focus in their identity-based model of motivated reasoning on identity goals such as goals of belonging, existential goals, status goals, or morals goals which can compete with accuracy goals.

Having established the central role of identity for motivated reasoning in understanding misinformation dissemination, I introduce the theory of identity-protection cognition in the next section. In doing so, I provide the theoretical and empirical milestones that led to the theory, emphasizing on identity defense mechanisms.

2.5.2 The theory of identity-protection cognition

Identity-protection cognition is, just as motivated reasoning, rooted in dissonance theory (see Chapter 2.3.1). Following cognitive dissonance, the underlying psychological function of identity-protection cognition is to preserve a positive sense of self. This manifests in the motivation of every individual to maintain (1) a consistent and stable sense of the self, (2) a competent sense of self, and (3) a morally good sense of self (Aronson et al., 1974).

Unlike cognitive dissonance theory, identity-protection cognition suggests that cognitive inconsistencies threaten not the individual self-concept but also individuals' *social* or *group identity*. In his work on identity-protection cognition, Kahan (2017) argues that political views and beliefs become “a badge of membership with identity-defining affinity groups” (p. 2). When confronted with opposing political views, the individual protects her group identity by rejecting or counterarguing the information—arriving at motivated reasoning. Especially political parties and political attitudes offer a strong group identification for individuals (Leeper & Slothuus, 2014). For the context of misinformation, identity-protection cognition hypothesizes that, due to a close identification with the misinformation, misinformation is more readily accepted and shared (see Chapter 2.3.2).

Although this protection mechanism may seem flawed on a normative level, Kahan (2017) argues that it is *individually rational* to accept confirming information and reject opposing information because it is “rationally suited to the ends of the agents who display it” (p.1). He argues that it is less detrimental for an individual to hold an inaccurate belief than to become an outcast of one's affinity group by agreeing to opposing information (Kahan, 2013; Van Bavel & Pereira, 2018). Research on the *black sheep effect* (Marques et al., 1988) supports this thinking. The black sheep effect describes how judgments of in-group members are more extreme than judgments of out-group members. This means that denying strongly held in-group opinions likely leads to harsher judgments when being part of the in-group as compared to being a member of an out-group.

Similarly, in their identity-based belief model, Van Bavel and Pereira (2018) delineate specific (partisan) identity goals, like belonging goals or status goals, which can contradict accuracy goals. Moreover, identity-protection cognition draws also on evolutionary psychology and utility maximization theory. Evolutionary psychology suggests that belonging to a social group fulfills inherent human needs of belonging (Baumeister & Leary, 1995), whereas utility maximization theory implies that individuals are likely to outweigh benefits and costs (Gilad et al., 1987).

Empirical studies that root information acceptance and rejection in identity affirmation and defense could confirm this approach. For example, for climate change communication and science communication on violent video games, Nauroth and colleagues (2017) examined the importance of identity affirmation and identity defense in evaluating scientists. Just as identity-protection cognition suggests, their results indicate that identity-affirming information leads to positive evaluations of scientists while identity-threatening information leads to negative evaluations. Nauroth et al.'s (2017) results align with previous findings that could show that climate change skepticism in the USA can be explained by strong ingroup thinking and high identification with a partisan group (Postmes, 2015; Unsworth & Fielding, 2014). In a similar vein, de Hoog (2013) found that negative group information increased threat perceptions and defensive thoughts, which mediated the relationship between identification and information evaluation. In the context of media and communication studies, Doosje and colleagues (2002) could show that identity-threats lead to defensive reactions such as attacking message credibility. In contrast, Hartmann and Tanis (2013) found that identity threats increase perceptions of hostile media. On a neurological level, Kaplan et al. (2016) could find that being confronted with opinion-incongruent information activated similar neural paths than threat or physical violence.

To sum up, this section provides theoretical and empirical evidence for identity-protection cognition and connects identity-protection cognition with an identity defense reaction. To go forward, in the next sections, I propose that, as part of the response to identity threats, emotions should be an inherent part of identity-protection cognition, especially for the context of misinformation. Before I connect identity-protection cognition and emotions, I review previous work which relates emotions to motivated reasoning.

2.5.3 Emotions in motivated reasoning

The idea that emotions contribute to motivated reasoning is not entirely new. But what are emotions and how do they contribute?

For this work, I will follow Scherer's (2005) definition of emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns for the organism" (p. 697). The five organismic subsystems Scherer (2005) refers to in his definition are *information processing*, *support*, *execution*, *action*, and *monitoring*. According to this definition, an episode qualifies as an emotion only when most of these subsystems respond with synchronized change to an event. In contrast, the terms affect, mood, and feeling refer to relatively stable and mild psychological states that primarily engage the

monitoring subsystem. Hence, affect and feeling are part of an emotional reaction but cannot be treated as synonyms (Frijda, 1993; Scherer, 2005).

Relying on recent insights from neuroscience, emotions' function is associated with *allostasis*—a brain process to anticipate and prepare the body's needs before they occur (Sterling, 2012). In turn, allostasis encompasses at least four different functions: dynamically regulating resources (e.g., increasing blood pressure before getting up), signaling the need for resources (a sensation of hunger), preparing the intake of resources (salivating before eating) and spending energy to collect new resources (e.g., going grocery shopping) (Feldman-Barrett, 2017). In turn, emotions as part of allostasis become a carrier of information that helps the brain to anticipate and prepare the body's needs. Central to allostasis is that the brain not merely regulates, as in homeostasis, but predicts what is needed (Sterling, 2012).

The conception of emotion as allostasis goes in hand with earlier findings suggesting that emotions function as information. For example, Schwarz and Clore (1983) proposed that individuals use emotions to infer conclusions in a “how-do-I-feel-about-it” manner. Positive evaluations are signaled by positive emotions signal. In contrast, negative evaluations are signaled by negative emotions. Slovic and colleagues (2007) understand this affective evaluation process as an *affect heuristic* that replaces effortful and time-consuming reasoning.

Findings by Schwarz and Clore (1983) and Slovic et al. (2007) are no exception. After being regarded as corroding rational thought (for an overview, see Blanchette et al., 2018), since the early 2000s, such emotional reactions have been more and more included to supplement cognitive models (Dukes et al., 2021). This development was also reflected in research on motivated reasoning. In their seminal work, Lodge and Taber (2000) combined the theory of hot cognition, on-line processing, and the “how-do-I-feel?” heuristic into an overarching model of politically motivated reasoning—the John Q. Public model of motivated reasoning. The central argument of their theory is that early affective responses color all downstream processing through associative pathways. In other words, early affective responses prime subsequent information processing, leading the authors to assume that an unconscious affective reaction causes motivated political reasoning. The authors strengthen the John Q. Public model's theoretical claims through empirical findings (Lodge & Taber, 2013; Taber & Lodge, 2016).

However, the idea of early affective responses that determine downstream information processing is challenged by appraisal theories of emotions. According to appraisal theory, early affective responses might occur in the first instance, yet, subsequent cognitive appraisals of the affective response should result in more differentiated emotional reactions (Lazarus, 1991). For

example, after being cognitively appraised, an early negative affective response can result in feelings of anger or anxiety (Lerner & Keltner, 2000). This suggests that emotion-specific influences override early affective responses. Appraisal theories, nevertheless, assume that emotions affect downstream information processing by prompting so-called action tendencies like fear-induced flight or anger-induced retaliation (Lazarus, 1991).

While the discussed theories by Lazarus (1991), as well as Lerner and Keltner (2000), examine emotions in general, other theories, similar to Lodge and Taber (2000), have tried to understand emotions in the context of politics. One commonly applied theory at the intersection of political psychology and political science comes from Marcus and colleagues (2000), who developed *affective intelligence theory* (AIT). Stemming on insights from neuroscience, the authors suggest a dual emotional system with the two subsystems—the dispositional system and the surveillance system—which guide information processing. While the dispositional system is primarily responsible for reliance on previously learned strategies and habits, the surveillance system is predominantly responsible for detecting novel or threatening stimuli and shifting attention to reasoned considerations.

Similar to appraisal theories of emotion, both systems relate to specific emotional states and subsequent action tendencies. The surveillance system triggers feelings of anxiety as a reaction to the detection of personal threats. In turn, the experience of anxiety motivates individuals to allocate greater attention to the threatening stimulus and results in more careful information seeking and processing and less reliance on habit. As a result, Marcus et al. (2000) assume that individuals rely less on prior attitudes or partisanship and become more open to attitude incongruent information. This was supported by empirical results (Brader et al., 2008).

In contrast, the dispositional system is related to two contrasting emotional states: enthusiasm and anger. The former is triggered when existing habits are adequate to respond to given stimuli, and individuals will seek to maintain this level of enthusiasm. The latter, anger, is an aversive reaction towards a reoccurring threat that is, unlike anxiety, met with information avoidance and the reliance on prior attitudes or partisanship.

In previous research, different authors have associated AIT with motivated reasoning. Weeks (2015), for example, found that individuals who received an anger manipulation were more likely to believe attitude congruent misinformation than individuals who received an anxiety manipulation. However, Weeks (2015) investigated incidental emotion's influence by exogenously priming either anger or anxiety. Hence, emotions were conceptualized as moderators of information processing and were not elicited as part of the experimental stimulus.

In contrast, Suhay and Erisen (2018) tested the integral influence of anger, anxiety, and enthusiasm on motivated reasoning. They suggested a mediating role of emotion that contributes to the effect of motivated reasoning. In other words, if beliefs are met with confirming/contradicting evidence, individuals react with enthusiasm/anger, which significantly contributes to the biasing effect of motivated reasoning. In line with AIT, the authors did not propose anxiety to affect motivated reasoning. In two studies, Suhay and Erisen (2018) found empirical evidence for their hypotheses: Enthusiasm and, much more so, anger contributed significantly to motivated reasoning, whereas anxiety did not contribute. While the authors provide a detailed account of how emotions affect information processing and report interesting findings, their account is, unfortunately, void of a theoretical explanation of why anger, anxiety, and enthusiasm are elicited.

To conclude, I established in this chapter a definition and the function of emotional reactions and how emotional reactions have been an integral part of research on motivated reasoning. The only model that formally includes affective reactions in motivated reasoning processes is the John Q. Public model by Lodge and Taber (2000; 2013) which is challenged by appraisal theories of emotion. First empirical findings by Weeks (2015) rely on exogenously priming emotions and Suhay and Erisen (2018) find that emotional reactions significantly contribute to motivated reasoning but do not provide a theoretical explanation for their findings. To overcome these limitations, in the following, I suggest an identity-protection model of motivated reasoning, which incorporates emotional reactions.

2.5.4 Combining emotions and identity-protection cognition

Identity-protection cognition assumes an identity-threat defense motivation to cause biased information processing (see Chapter 2.5.2). An identity-threat defense motivation elicits emotional reactions such as anger or anxiety (e.g., De Hoog, 2013; Huddy et al., 2005). Hence, emotions must inherently be part of identity-protection cognition. But, how exactly do emotions contribute to identity-protection cognition?

In the context of identity threat, anger is associated with a perceived violation of one's standards and coincides with an approach motivation (Carver & Harmon-Jones, 2009). In contrast, threat-induced anxiety results from lacking personal control and increased uncertainty, coinciding with an avoidance motivation (Eysenck et al., 2007). Moreover, although anger and anxiety are fundamentally different theoretically, they often co-occur empirically (Carver & Harmon-Jones, 2009). Transferring this to the context of misinformation, incoming misinformation either violates one's identity-relevant standards (e.g., attitudes, partisanship)

and elicits threat-induced anger or questions identity-relevant norms to the degree that one is uncertain about the veracity of one's standards eliciting threat-induced anxiety.

Although not originally anticipated by Kahan (2016, 2017), identity affirmation just be included in identity-protection cognition just like identity threat because it was previously found that individuals react to identity affirming stimuli (e.g., Marcus et al., 2000). For example, Marcus et al. (2000) found that, like identity threat eliciting anger and anxiety, identity affirmation elicited positive emotions such as enthusiasm. Transferring this to the context of misinformation, incoming misinformation affirms an individual's identity to an extent that the person feels elevated and enthusiastic.

Subsequently, following the affect heuristic (Slovic et al., 2007), anger, anxiety, and enthusiasm serve as additional information, reinforcing effects of identity threat/affirmation. Because anger, anxiety, and enthusiasm result from an identity threat/affirmation, I suggest a mediation model of identity-protection cognition (see Figure 1).

For the context of misinformation, the same information can be perceived as an identity threat as well as an identity affirmation depending on the individuals' identity. The occurrence of emotions is tested in Paper 4 (see Chapter 3.4) and the mediating role of the emotions anger, anxiety, and enthusiasm in Paper 5 (see Chapter 3.5).

While I suggest that emotional reactions should inherently be elicited as part of identity-protection cognition, emotions are also closely tied to the context of misinformation. Hence, situating identity-protection cognition in the context of misinformation, the connection of emotional reactions and identity-protection cognition is strengthened in two ways: First, emotionally laden (mis)information are more likely to be shared in social networks. For

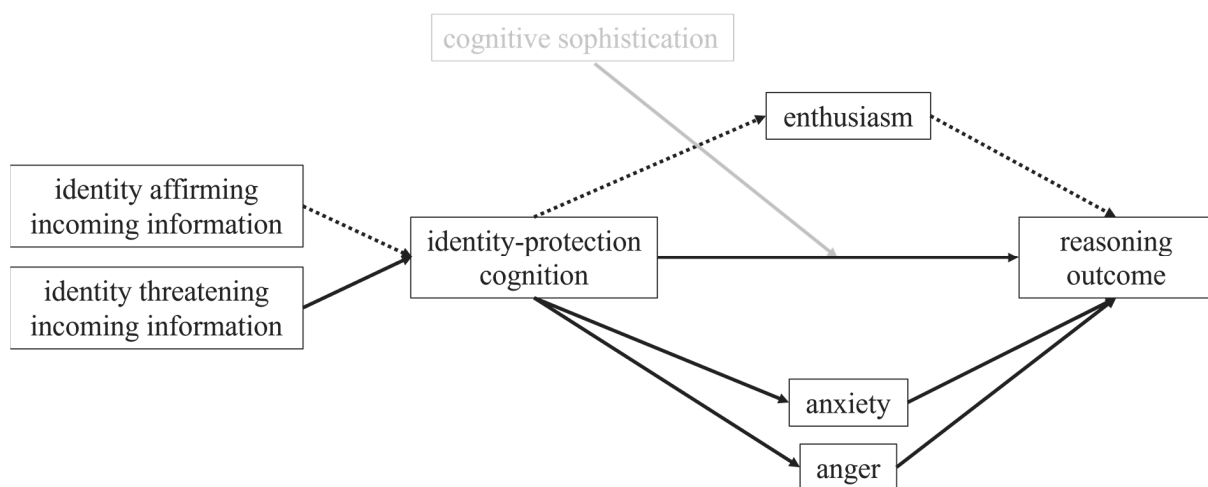


Figure 1: Elicitation of anger, anxiety, and enthusiasm as the result of identity-threatening or identity-affirming (mis)information (highlighted in black print).

example, Brady et al. (2017) found that moralizing emotions in political messages on Twitter shaped information diffusion within social networks. This can be explained by results from Wang and colleagues (2020). The authors found that misinformation are associated with the experience of negative emotions, which were, in turn, related to the diffusion of information online. In other words, if misinformation induced negative emotions, individuals were more likely to share them.

Second, Bakir and McStay (2018) argue that misinformation are often tailored to appeal to specific user groups. This development is intensified through algo-journalistic technological affordances, such as sentiment analysis, gauging users' emotions, and disseminating emotionally congruent news. Once platforms know how users feel about specific issues, the authors also propose that they can manipulate these feelings through emotional contagion. However, Sivek (2018) states that "the degree to which emotions are a focus of this personalization, and how feelings may be manipulated with new emotion analytic tools that assess and respond to users' emotional states" (p. 124) is less well known. Furthermore, knowing that misinformation often contains emotional appeals and is likely to be promoted through emotional messages on social media, researchers have developed tools that exploit these characteristics to detect misinformation (Guo et al., 2019).

2.5.5 Identity-protection or cognitive sophistication?

Previous research on identity-protection cognition in the context of susceptibility to falling for misinformation has also investigated the role of specific individual dispositions. Especially the role of cognitive sophistication, meaning the individual's disposition to engage in deliberative information processing, has resulted in contradicting results. For example, Kahan et al. (2017) found that individuals high in cognitive sophistication showed stronger identity-protection cognition. The authors derive from these results that individuals with greater cognitive sophistication skills are better equipped to counter-argue attitude-incongruent information than individuals with lesser skills. In other words, individuals with greater cognitive sophistication skills are better at using "all their cognitive resources at their disposal to form and persist in identity-consistent beliefs" (Kahan, 2017, p. 7).

In contrast to these results, more recent investigations found opposite results. For example, Pennycook and Rand (2019) examined how individuals rated false and true news. The authors also measured participants' partisanship as well as cognitive sophistication skills. Results indicated that, while partisanship could account only to some degree for how individuals rated the veracity of false and true news, individuals with higher cognitive sophistication skills were more likely to discern between false and true news. These results

were supported by findings from Bago and colleagues (2020). They experimentally manipulated cognitive sophistication by instructing participants to rate the veracity of news under time pressure. In a second step, participants were given additional time to rethink their choice, thus allowing for more deliberation which improved discernment of false and true news. Similarly, Lind et al. (2018) found that higher numeracy skills led to less motivated reasoning about numerical data, leading the authors to conceptualize the underlying process as *motivated-reasoning-as-analysis*.

While results for cognitive sophistication are mixed, previous findings make it evident that cognitive sophistication plays a vital role in identity-protection cognition. As most studies indicate a negative relationship between cognitive sophistication skills and identity-protection cognition as well as a moderating role of cognitive sophistication, this view is adapted in this thesis (see Figure 2).

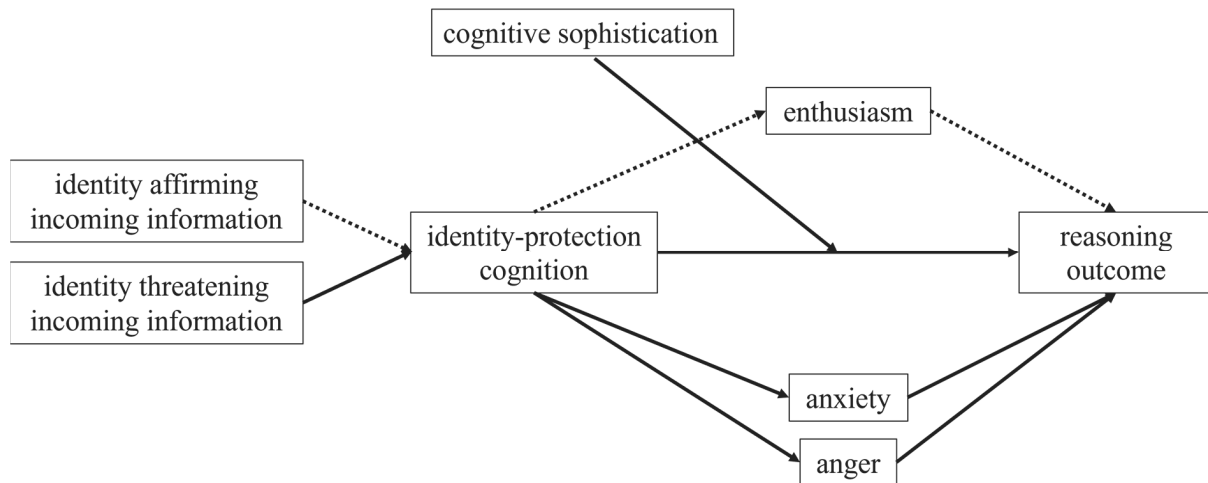


Figure 2: The mediation model of identity-protection cognition introduced in the previous chapter extended by the moderating role of cognitive sophistication (highlighted in black print).

2.5.6 Preventing misconceptions through identity-salience manipulations

Adopting identity-protection cognition in the context of misinformation also allows for interventions to increase misinformation rejection and factual information acceptance by drawing on insights from social identity theory (Tajfel & Turner, 1979; 1986). Social identity theory proposes that a person can identify either as an individual or a group member. The later developed self-categorization theory extends this view by asserting that the individual experiences the self through varying identities (Turner et al., 1994). In response to contextual or social cues, the prevalent identity can shift. In turn, the salient identity of the moment guides downstream cognition and emotions (Hornsey, 2008).

Interventions, shifting identity saliences towards, for example, shared identities or higher-order identities could successfully change individuals' perception of and behavior towards others. It was found, for example, that identity salience manipulations resulted in decreased stereotype susceptibility and stereotype threat (McGlone & Aronson, 2006; Shih et al., 1999) as well as changes in policy support (Unsworth & Fielding, 2014). Maitner et al. (2010) explain similar effects of identity salience manipulations in their study by changing information relevance. Depending on which identity is (made) salient, information become less or more relevant.

Similarly, I suggest that changing identity salience from a threatened to an unthreatened identity should result in increased misinformation rejection and factual information acceptance, which should also be reflected in the individual's emotional experience (see Study 4).

2.6 Conclusion and research objectives

The two central aims of this thesis are (1) to establish empirical effects of motivated reasoning on misinformation on social media with a specific emphasis on how misinformation is shared, and (2) to gain a deeper understanding of the driving mechanisms of motivated reasoning by investigating the role of emotions and identity. To that end, motivated reasoning has already been identified as a central driver of misinformation online theoretically (Kahan 2017) and empirically (e.g., Anthony & Moulding, 2019; Moravec et al., 2018; Weeks, 2015). With most empirical research focusing on veracity judgments, asking participants to differentiate between true and false news, Study 1-3 go beyond this veracity question. In particular, it is investigated how two entities facilitate the propagation of misinformation on social media: users and social bots. To this end, Study 1 provides observational evidence for the role of motivated reasoning by examining users' hyper-partisan news sharing behavior on Twitter. Scrutinizing hyper-partisan news, which has previously been identified as the central source of misinformation, Study 1 suggests that motivated reasoning drives the sharing process in a way that users are more likely to share attitude-congruent hyper-partisan news than attitude-incongruent hyper-partisan news.

Because hyper-partisan news are not the only facilitator of misinformation, in Study 2 and Study 3, automated accounts, so-called social bots, are investigated. Study 2 aims to understand how users perceive these accounts, suggesting that partisan-congruent accounts are perceived as less bot-like and partisan-incongruent accounts as more bot-like. Going one step further, Study 3 examines how these, possibly biased, perceptions affect engagement intentions with social bots. Relying on insights from studies scrutinizing social influence (e.g., Tussyadiah

et al., 2018), examining users' engagement intentions with social bots is especially important as social bots can influence users through interactions. In line with motivated reasoning, Study 3 hypothesizes that users are more likely to engage with attitude-congruent social bots than with attitude-incongruent social bots.

While the first three papers illuminate how motivated reasoning plays out on social media, Study 4 and Study 5 take a deeper look into the underlying psychological processes of motivated reasoning. With misinformation being closely related to identity politics and partisanship (Mourão & Robertson, 2019; Trevors, 2019; Tripodi, 2018), I argue in these papers for an identity-centric approach to motivated reasoning in the context of misinformation. In doing so, I build on and extend insights about identity-protection cognition as the source of motivated thinking (Chapter 2.5.2). Entrenched in cognitive dissonance (Festinger, 1957), I ground identity-protection cognition in the experience of identity threat and identity affirmation, characterized by subsequent emotional experiences. I hypothesize that attitude-congruent (mis)information affirms one's identity, eliciting enthusiasm and increasing (mis)information acceptance, whereas attitude-incongruent (mis)information threatens one's identity, eliciting anger and anxiety and decreasing (mis)information acceptance. The aim of Study 4 is to test these assumptions. In addition, Study 4 also offers an approach to mitigate the effects of identity-protection cognition through identity-salience manipulations.

Extending the efforts of Study 4, Study 5 explicates the underlying psychological processes of identity-protection cognition by including experienced emotions as a third explanatory variable (see also Suhay & Erisen, 2018). In doing so, I follow recent advances in the study of cognition and emotion, which suggest that emotions affect subsequent information processing (e.g., Lerner et al., 2015; Marcus et al., 2000; McKasy, 2020; Nabi, 1999; Redlawsk, 2006). In addition to the mediating role of emotions, individual differences in cognitive sophistication are also expected to affect identity-protection cognition, as previous studies have found contradicting results for this disposition. While Kahan et al. (2017) found that individuals higher in cognitive sophistication showed a stronger bias, others found opposite results (Lind et al., 2018; Pennycook & Rand, 2019; Tappin et al., 2020), implying that greater cognitive sophistication skills decrease the effects of identity-protection cognition, which I follow in Study 5.

The contribution of each study in relation to the two research questions is visualized in Figure 3. In addition, Figure 3 also depicts the different approaches (observational–Study 1 versus experimental–Study 2-5) and the dependent variables of each study (the response column). While all studies contribute to how motivated reasoning might affect misinformation

on social media, contributions by Study 1-3 allow for more direct implications concerning the social media platform Twitter but are less conclusive about theoretical implications. In contrast, Study 4 and 5 are more theoretically oriented, allowing only for indirect implications concerning social media platforms.

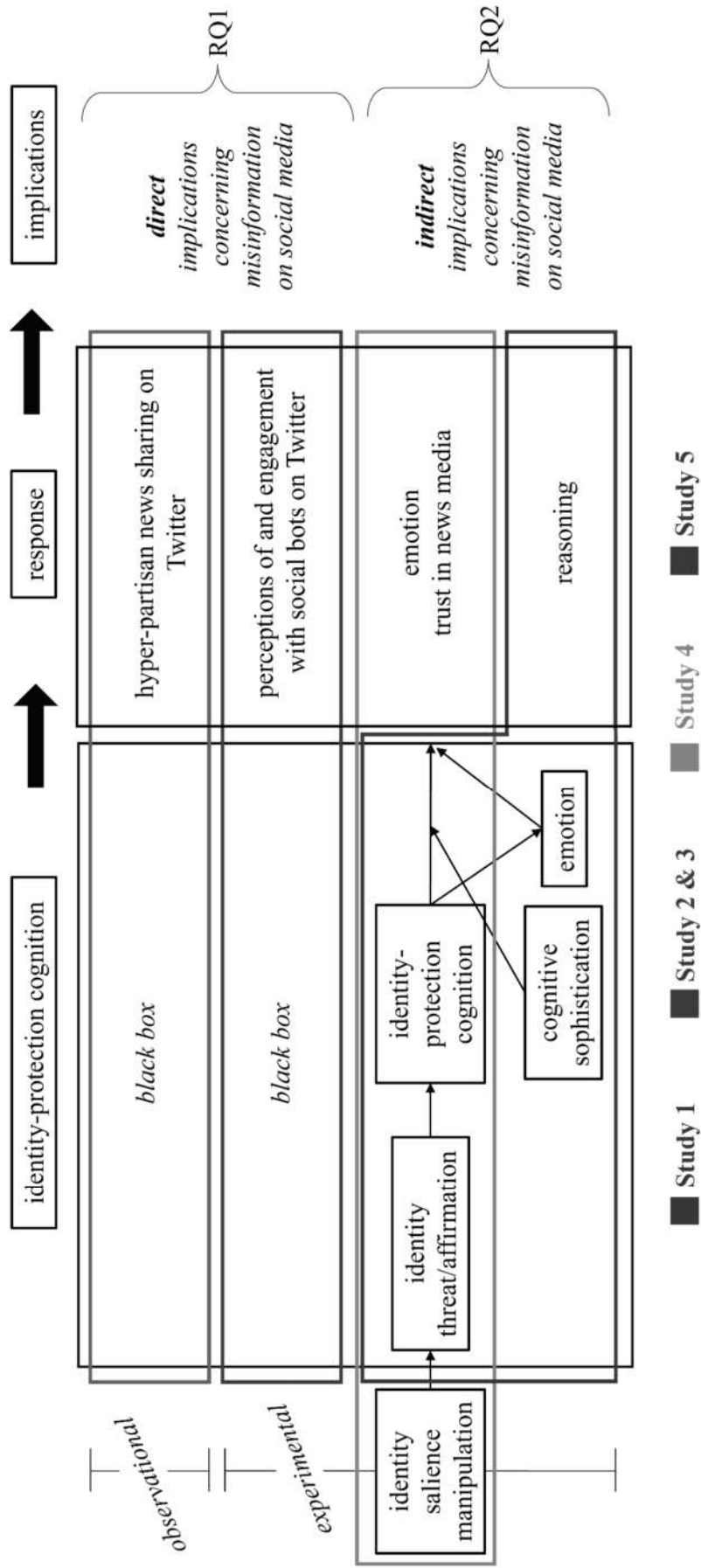


Figure 3: Visualization of the contribution of each study in relation to the two research questions.

3. SUMMARY OF THE RESEARCH PAPERS INCLUDED IN THE CUMULUS

3.1 Study 1: Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter

Previous studies have already investigated why human users share news (e.g., Trilling et al., 2017). In doing so, most investigations of news-sharing practices have scrutinized news-sharing from two perspectives: (1) news content that is more attractive to be shared, and (2) user motivations that drive the sharing process. The first perspective proposes that specific news content is more likely to be shared by users than other news content, for example, news containing conflict or physically and culturally proximate news coverage (Kümpel et al., 2015; Trilling et al., 2017), disregarding that users differ in their sharing motivations. The second perspective suggests that users' sharing activity is driven by different motivations, like reputation motives or information-seeking motives (Kümpel et al., 2015), disregarding that news stories cover a broad range of topics. In this study, we suggest that both perspectives are two sides of the same coin and should, hence, be jointly considered. Study 1 brings together both perspectives and applies these to hyper-partisan news sharing.

Similar to the previous studies, the central hypothesis guiding Study 1 suggests that hyper-partisan news sharing is driven by motivated cognition in a way that users are more likely to share news that are in line with their political identity than news that oppose their political identity. It is also suggested that hyper-partisan news which contain physically and culturally proximate content, conflict, a human angle (human interest), moralizing content, or visual content is more likely to be shared.

One week's content from Infowars.com, as it was shared on Twitter (tweets and Infowars-URLs), was collected to investigate these hypotheses. Twitter as a social media platform for hyper-partisan news sharing was selected because it is a popular medium for news dissemination and consumption (Tandoc & Johnson, 2016). Infowars.com was selected as it is one of the most prominent hyper-partisan news outlets in the USA (Newman et al., 2019). Moreover, Infowars presents an interesting case as its dissemination is entirely driven by third-party accounts since Infowars and its founder Alex Jones were banned from Twitter in September 2018.

All account descriptions that shared an Infowars URL were assessed through Twitter's API as a proxy for how individuals identified themselves. Through a semi-automated clustering approach with keyword lists, accounts were then classified into seven categories. Similarly,

Infowars content was classified into the same categories. To determine which Infowars content was shared on Twitter and which was not shared, URLs shared on Twitter were compared with a dataset from GDELT⁴, an open-data project, monitoring global news coverage in real-time. In the next step, all news stories shared by Infowars were manually coded into the hypothesized shareworthiness factors: physically and culturally proximate content, conflict, a human angle (human interest), moralizing content, or visual content. Two different measures were applied to assess an Infowars stories' success: the number of tweets generated and a retweet factor that measured the amplification of initial sharing by subsequent retweeting. To arrive at the retweet factor, the total number of retweets was divided by the count of original shares.

Results indicated that almost all content that was published on Infowars was also shared on Twitter. Moreover, the motivated reasoning hypothesis could partially be supported. Especially accounts that endorsed former president Donald Trump were more likely to share Infowars stories about Donald Trump. Similarly, accounts with strong affinities for Infowars as well as accounts endorsing Christianity were more likely to share attitude-congruent content than content that did not discuss Infowars content or content on Christianity. Furthermore, it was found that Infowars articles that contained conflict and a human angle were more likely to be shared on Twitter when measuring an Infowars sharing success through the tweet count measure. Interestingly, it was also found that for the second measure, the retweet factor, only stories with a human angle were more likely to be shared.

3.2 Study 2: Disagree? You must be a bot! How beliefs shape Twitter profile perceptions

Human user does not always generate communication on social media. Social bot accounts, run by machine learning algorithms, act autonomously and can create content and disseminate it online as well as comment, like, or follow. Recent empirical investigations could connect social bots to the circulation of misinformation on social media platforms (Wang et al., 2018). Moreover, social bots have also been accused of being engaged in political astroturfing (Keller et al., 2020) and influencing election outcomes (Bastos & Mercea, 2019; Ferrara et al., 2016; Schäfer et al., 2017) as well as public health through the promotion of e-cigarettes (Allem et al., 2017) and anti-vaccination campaigns (Broniatowski et al., 2018). While much academic efforts went into detecting social bots, applying machine learning models for automated detection, study 3 focusses on when and why humans perceive an account as a social bot. In line with identity-protection cognition, it is hypothesized that accounts which are congruent to

⁴ <https://www.gdelproject.org/>

a person's attitude would be less likely to be perceived as a social bot, whereas accounts which are incongruent to a person's attitude would be more likely to be perceived as a social bot. Yet, it was assumed that participants could generally differentiate between a bot profile and a human profile. In addition, it was hypothesized that the relationship between a person's attitude and the profile perception is mediated by the perceived credibility of the profile. All hypotheses as well as the data analyses were pre-registered.

To test these hypotheses, a 2 (attitude: congruent, incongruent) x 3 (profile: human, ambiguous, bot) within-subject online experiment was conducted. Participants viewed overall 24 mock-up Twitter profiles and were asked to indicate whether they thought the profile displayed a human account or a social bot and how credible they found the profile. In addition, participant's political partisanship was measured. Because the perception of Twitter profiles is likely to be affected by the individual participant's general Twitter usage, control variables such as *age*, *gender*, *education*, and *time spent on social media*, the *number of social media platforms used*, and *previous social bot knowledge*. It was chosen to create mock-up profiles instead of real Twitter profiles to control other factors influencing profile perceptions such as engagement metrics (number of likes), language, profile descriptions, or follower count. All profiles were pre-tested.

A repeated-measures ANOVA confirmed that individuals could generally differentiate between human and social bot profiles. A significant interaction of the factor 'profile' with the control variable age indicated that the younger the participants were, the worse they could differentiate between a human and a social bot profile. The central hypothesis for congruency was not supported. Attitude (in-)congruency did not significantly change whether participants rated a profile as a social bot or a human. However, interaction effects of congruency with the three control variables *age*, *previous social bot knowledge*, and *time spent on social media* supported the motivated reasoning hypothesis. Younger participants who had encountered social bots before and spent more time on social media rated attitude-congruent profiles as more human and attitude-incongruent profiles as more bot-like. As hypothesized, the effect of attitude-(in)congruency was mediated by the perceived credibility of a profile. Hence, attitude-congruent profiles were rated as more credible, leading participants to perceive attitude-congruent profiles as more human.

These results generally support the motivated-reasoning hypothesis for a sub-group of users: those who are the most familiar with the social media environment. The effects of familiarity, in turn, can be explained by previous findings showing an effect of familiarity inducing processing fluency. Processing fluency, as in the ease with which an intuitive response

is produced, can increase feelings of rightness and confidence as well as reliance on heuristics (Thompson et al., 2013). In turn, reliance on heuristics has been connected to stronger effects of motivated reasoning (Pennycook & Rand, 2019; Tappin et al., 2017). In addition, results might also indicate a different usage of the term ‘social bot’. In popular media, it has been discussed that calling someone a ‘social bot’ in social media can be a pejorative term to indicate dislike or disagreement without actually thinking that an account is a social bot.

3.3 Study 3: “I agree with you, bot!” How users (dis)engage with social bots on Twitter

Results from Study 2 in this cumulus, as well as previous studies on social bot communication (Yan et al., 2020), confirmed the motivated reasoning hypothesis: Partisanship-congruent accounts were perceived as more human and partisanship-incongruent accounts as more social-bot like. In the next step, this finding was then implemented to understand how these differences in perception affect the persuasive impact of social bots.

Because social bots have been accused of stirring election outcomes (Bastos & Mercea, 2019; Ferrara et al., 2016; Schäfer et al., 2017), posing health hazards (Allem et al., 2017; Broniatowski et al., 2018), and spread of inauthentic content (Wang et al., 2018), it is vital to understand how social bots exert influence. While previous studies have examined the influence of social bots on social networks by bots, for example, pushing hashtags or starting astroturfing campaigns, less is known about the influence of social bots on users. In contrast to network influence, social influence proposes that different actors on a network exert influence through engagement. Transferring this to the effects of social bots, a social-influence hypothesis proposes that social bots spread their agenda by attracting users to engage with them. In this paper, it is hypothesized that the success of the social influence of bots depends on two factors: the displayed level of humanness of the bot and the displayed attitude/partisanship of the bot account.

Relying on insights from previous human-computer interaction (Edwards et al., 2014; Edwards et al., 2015), it is proposed that users are more willing to engage with others who appear to be human users. In other words, the more human-like social bots are, the more likely they will be to engage users. Drawing on motivated reasoning, it is suggested that users prefer to encounter like-minded users. In addition, a research question asks whether both factors (level of humanness & partisanship) affect each other. Assessing users’ motivations to engage helps to understand why differences are observed.

In a 3 (level of humanness: low, medium, high) x 2 (congruency: congruent, incongruent) within-subject online experiment, N = 223 U.S. American Twitter users were recruited through the online panel Prolific. All users viewed overall 18 different Twitter profiles, varying in the level of humanness and partisanship. After viewing a profile, participants were asked whether they would like to engage (follow, retweet, comment, and quoted retweet) the profile, how they would react if the profile would engage with them (follow, retweet, comment, and quoted retweet), why they choose to answer in a certain way (engagement motivations) and whether they thought the profile was a social bot or a human user.

Results indicated a strong effect of humanness and partisanship. Confirming the hypothesis, profiles with low levels of humanness were the least likely to receive engagement. The more human-like a profile became, the higher the chances users would engage. However, this was only true for profiles that shared the users' partisanship (congruent profiles). Independent of their level of humanness, incongruent profiles were very unlikely to receive any engagement. In addition, it was found that the effect of congruency was most pronounced for highly human-like profiles, increasing the likelihood of engagement. The effect of congruency on engagement likelihood could partly be explained by how users perceived the accounts. The more congruent a profile was, the more it was perceived as a human profile and the more it was likely to engage users.

These results conclude that the influential impact of social bots through direct engagement is probably comparatively low. Social bots must show very high levels of humanness, successfully disguising their automated nature, to receive engagement. We already know from previous research that these accounts are relatively scarce (Assenmacher et al., 2020). In contrast to a popular narrative that motivated reasoning makes users more vulnerable, these results also indicate that motivated reasoning has a protective function.

3.4 Study 4: I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption

In study 4, participants were asked to evaluate the credibility of (mis-)information and trust in the source. It was expected that identity threat would lead to lower credibility and trust ratings, whereas identity affirmation would lead to higher credibility and trust ratings. It was also expected that this is accompanied by affective reactions (higher ratings of anger/anxiety for identity-threatening (mis-)information and higher ratings of enthusiasm for identity-affirming (mis-)information). Moreover, Study 4 also included a manipulation of identity-salience. Relying on social-identity theory (Tajfel & Turner, 1979, 1986) and self-

categorization theory (Turner et al., 1987), previous studies could show that identity salience can be manipulated through contextual cues and reduce stereotype susceptibility (Shih et al., 1999), stereotype threat (McGlone & Aronson, 2006) and increase policy support (Unsworth & Fielding, 2014). Hence, it was hypothesized that shifting identity salience from a threatened to an unthreatened identity would increase credibility and trust ratings in information (increased acceptance of factual information), also leading to lower experience of anger and anxiety. Similarly, shifting identity salience from an affirming identity to a non-affirming identity was hypothesized to reduce credibility and trust in misinformation (decrease acceptance of misinformation), accompanied by a decreased experience of enthusiasm.

Hypotheses were tested in two experimental studies with overall $N = 353$ participants. The purpose of the first study was to increase acceptance of factual information, whereas the purpose of the second study was to decrease acceptance of misinformation. The setup in both studies was similar. Before being presented with a news article, one-half of participants received an identity-salience manipulation. The Trust in News Media scale by Kohring and Matthes (2007) was used to assess credibility and trust. Similar to study 1, emotions were measured through self-reports.

Results of both studies supported the identity-protection hypothesis: participants in the first study condition (increase factual news acceptance) who were threatened by the factual news evaluated the article significantly worse than participants who were not threatened. Similarly, participants in the second study condition (decrease in misinformation acceptance) whose identity was affirmed gave higher credibility and trust rating for the misinformative news story. Unlike expected, this was for the first study condition not reflected in the experience emotion of anger and anxiety. Both threatened and unthreatened participants experienced similar levels of anger and anxiety. For both study conditions, identity-salience manipulation did not lead to expected credibility and trust rating changes. However, in the second study condition, after receiving the identity-salience manipulation, participants experienced less identity-affirmation-related enthusiasm.

3.5 Study 5: The role of emotions and identity-protection cognition when processing (mis)information

The research objective of this paper is to contribute to the growing knowledge about misinformation by scrutinizing the effects of identity-protection cognition and subsequent protection-related emotions. In a theory-driven approach, insights from the theory of identity-protection cognition are combined with theories of identity threat, identity affirmation, and

emotional reactions to identity threat/affirmation. Moreover, the moderating role of cognitive sophistication is tested. It is predicted that cognitive sophistication skills drive the reasoning outcome for neutral stimuli (no identity threat). For polarizing, identity threatening/affirming stimuli, the reasoning outcome should be driven by identity-protection/affirmation goals and mediated by emotional reactions of anger and anxiety (for identity-threat) and enthusiasm (for identity affirmation). Similarly, cognitive sophistication is expected to moderate this relationship for identity-threatening/affirming stimuli.

For this purpose, an online experiment using a convenience sample of 463 (304 female) German citizens was conducted. Assessing individual cognitive sophistication skills through a numeracy scale, participants were shown three different fictitious scenarios (neutral/polarized I/polarized II) in a randomized order. Each scenario consisted of a math task, requiring participants to draw inferential conclusions from numerical data. After viewing each scenario, participants were asked to indicate which two mutually exclusive statements represent the correct conclusion. Immediately after each scenario, participants were instructed to self-report their emotional reactions, elicited by the task. Finally, participants were asked to report their political identity, basic demographic data, and read a debriefing statement.

The results supported the hypothesis for neutral stimuli (no identity threat/affirmation). Participants' cognitive sophistication skills predicted the correct response. However, the results for the polarized scenarios were less clear. Political identity predicted only in one scenario the reasoning outcome. The reasoning outcome in the second polarizing scenario was, again, predicted by cognitive sophistication. Consequently, the moderating role of cognitive sophistication was only found in the first polarizing scenario but not in the second.

Furthermore, the mediating role of emotions was only found for the first polarizing scenario. More precisely, it was found that only anxiety, but neither anger nor enthusiasm, mediated the relationship of political identity and the reasoning outcome. Interestingly, although not predicted by political identity, the affective response of enthusiasm was nevertheless significant in predicting the reasoning outcome in all conditions. Because emotions were not consistently related to political identities, we assume that elicitation was somehow the result of political identities and the task content, which in turn elicited enthusiasm. This became evident when inspecting the emotional reactions of the control group. According to identity-protection cognition, the control condition should not induce identity-threat or identity-affirmation and subsequent emotions. Nevertheless, mean ratings of anger, anxiety, and enthusiasm were significantly greater than zero, even in the control condition.

To sum up, neither the political identification nor cognitive sophistication could fully explain the data but rather the interaction of both. Other than expected, political identity was not as decisive in predicting the reasoning outcome, but, instead, emotional reactions determined responses.

4. DISCUSSION

The aims of this thesis were, on the one hand, to establish empirical effects of motivated reasoning on misinformation on social media, and, on the other hand, to gain a deeper understanding of the driving mechanisms of motivated reasoning by investigating the role of emotions and identity. Previous research could already confirm that, due to motivated reasoning, individuals are more likely to assess attitude-congruent misinformation as accurate information and are, in turn, more likely to assess attitude-incongruent misinformation as inaccurate (Anthony & Moulding, 2019; Ecker et al., 2014; Kahne & Bowyer, 2017; Kuklinski et al., 2000; Nyhan & Reifler, 2010). In this thesis, I go beyond this notion of veracity judgments and examine how motivated reasoning shapes misinformation circulation by investigating hyper-partisan news-sharing on Twitter (Study 1) and users' perceptions and engagement with malicious social bots (Study 2 & 3).

Besides identifying empirical effect of motivated reasoning on misinformation sharing, I investigated the underlying psychological processes of motivated reasoning. In particular, I examined an identity-protection cognition model of motivated reasoning and empirically tested the contribution of threat and affirmation-related emotions to motivated reasoning. In doing so, I argue throughout this thesis (Chapter 2.5.1. & 2.5.2) for an identity-centric model of motivated reasoning.

In the following sub chapters, the core results of the five studies included in the cumulus of this thesis are discussed with respect to the two research foci. Building on and extending these results, I derive theoretical implications, followed by a section describing the limitations of my research and future research possibilities. In the final chapters of this thesis, I derive practical implications as well as a closing chapter, comprising a general conclusion.

4.1 Overview of the findings

4.1.1 Motivated reasoning and misinformation on social media

Many researchers have pointed to motivated reasoning to understand the unprecedented spread of misinformation online, which suggests that attitude-congruent misinformation is less likely to be questioned and scrutinized and more likely to be believed and shared. Previous

research found support for the proposed effects of motivated reasoning (Anthony & Moulding, 2019; Ecker et al., 2014; Kahne & Bowyer, 2017; Kuklinski et al., 2000; Nyhan & Reifler, 2010): Attitude-congruent misinformation was perceived as more accurate as compared to attitude-incongruent misinformation which was perceived as less accurate, indicating biased veracity judgments. I extended these findings by examining how motivated reasoning influences how misinformation is shared on social media. In doing so, I focused on hyper-partisan news and social bots.

Through observational data, Study 1 confirmed the motivated reasoning hypothesis for hyper-partisan news-sharing. To gauge Twitter users' attitudes, profile descriptions were leveraged and semi-automatically clustered into attitude categories. Using logistic regression analyses, it was found that Trump supporters were more likely to share hyper-partisan news about Trump than any other user group. Similarly, patriots, Infowars-supporter, and self-declared Christians were more likely than any other user group to share hyper-partisan news matching their attitude. In contrast, users supporting the military, holding pro-gun attitudes, or endorsing conspiracy theories did not show clear sharing patterns. One explanation of the null results for these users is that users fell into multiple attitude categories. For example, users supporting and endorsing the military were also likely to support former president Donald Trump, displayed patriotic views, and endorsed weapons. Assuming, as I argue in this thesis (Chapter 2.5.1), that attitudes become connected to identities, these results are in line with social identity theory (see Chapter 2.5.2) which suggests that individuals hold more than one (social) identity. Moreover, previous studies on social identity theory suggest that not all identities are equally important for individuals and might not be similarly salient at all times (Tajfel & Turner, 1986). Hence, supporting the military was either not as important to users as compared, for example, to supporting Donald Trump, or attitudes about the military did not become as salient as other attitudes.

Results of Study 1, supporting the motivated reasoning hypothesis for hyper-partisan news-sharing, are commensurate with previous results on news in general (An et al., 2013) and recent experimental results concerning misinformation sharing on Facebook and WhatsApp (Bauer & Clemm von Hohenberg, 2020). Due to motivated reasoning, users become more vulnerable to share (mis)information which are in line with their attitude/partisanship but are also protected from sharing (mis)information if it was not in line with their attitude/partisanship.

In Study 2 and Study 3, these results were extended to sharing by automated social media accounts, so-called social bots. While previous investigations have studied if and how social bots themselves share misinformation (Bessi & Ferrara, 2016; Forelle et al., 2015; Wang

et al., 2018), Study 2 and Study 3 examined how users perceive and interact with these accounts. For both studies, the underlying assumption was that social bots exert influence not only through network effects (Cheng et al., 2020; Keijzer & Mäs, 2021; Ross et al., 2019) but also by engaging with human users who, in turn, promote misinformation shared by social bots. The guiding hypothesis was that attitude-congruent social bot accounts are perceived as more human-like than attitude-incongruent accounts that are perceived as more bot-like. Study 2 supported this hypothesis showing that the effect of attitude-congruency was partly due to biased credibility perceptions. Extending these results, Study 3 could confirm that attitude-congruent accounts are not only perceived as more human-like but are also more likely to engage users. This effect was particularly strong for highly human-like accounts, which can disguise their automated nature. In contrast, attitude-incongruent accounts were most likely ignored, independent of their level of humanness. These results confirm not only confirm the twofold contribution of motivated reasoning (increased vulnerability but also increased protection) but also indicate that social media users do not blindly interact with any account which shares their political partisanship. Users differentiate between credible and uncredible accounts.

To conclude, the results of all three studies investigating the effects of motivated reasoning on misinformation sharing on social media reveal that motivated reasoning increases vulnerability to misinformation on the one hand but also increases protection against misinformation on the other hand. However, the effects of motivated reasoning are decreased if accounts are perceived as less credible.

4.1.2 Connecting identity-protection cognition and emotions

Based on the assumption that identity-protection cognition originates in the experience of identity-threat or identity-affirmation, I argue throughout the present thesis that the experience of threat or affirmation should elicit emotional reactions. In Study 4, I tested this overall assumption, measuring participants' experience of anger, anxiety, and enthusiasm after reading (mis)information. Results of Study 4a indicated that if the content of news threatened participants' identity, they evaluated true news worse. However, the experienced levels of anger and anxiety of threatened identities were similar to non-threatened participants. Moreover, while identity-threatened participants did not increase their evaluations, when receiving an identity salience manipulation their levels of experienced anger decreased significantly. In contrast, no changes for experienced levels of anxiety were found. In Study 4b, identity-affirmed participants rated false news not only higher, they also reported significantly higher

levels of experienced enthusiasm than non-affirmed participants. However, unlike Study 4a, the identity salience manipulation did not affect experienced levels of enthusiasm.

While in Study 4, emotional reactions were another outcome variable, in Study 5, it was hypothesized that emotions mediate the relationship of identity-protection cognition and subsequent information processing. Because identity-protection was not found in all condition of Study 5 (see more in the next Chapter 4.1.3), testing the mediation hypothesis was limited to the condition in which identity-protection cognition was found. This was the case for the condition which indicated a crime rate decrease upon the intake of refugees. Here, having a relatively conservative view and a politically right-leaning identity was associated with a decreased likelihood of a correct response. Likewise, holding such an identity was associated with higher levels of anxiety. In turn, higher levels of anxiety decreased the likelihood of a correct response. This result supports the mediation hypothesis, and indicated that emotion are incorporated via a ‘how-do-I-feel-about-it’ heuristic (Schwarz & Clore, 1983).

For the remaining conditions, only in some cases were the experienced emotions associated with the political identity. Interestingly, for both conditions of the crime rate scenario⁵, being more conservative and right leaning resulted in higher levels of anxiety. This contradicts predictions made by identity-protection cognition. In the condition ‘crime rate increase’, which is congruent to conservative and right-leaning narratives, and which was pretest to support this claim, no identity threat should have been present, and, hence, levels of anxiety should have been low.

To help understand these divergent findings, emotional reactions of the control group were also assessed. According to identity-protection cognition, the control condition should not induce identity-threat or identity-affirmation and subsequent emotional reactions. Nevertheless, mean ratings of anger in the control condition were significantly higher than mean ratings of anger in the refugee intake condition and similarly high to the mean ratings of anger for the Diesel ban. While mean ratings of anxiety and enthusiasm in the control condition were lower than in either of the polarizing conditions, ratings were significantly greater than zero. Considering the content of the control condition, a skin rash that either increases or decreases after applying a crème, it is not surprising that the content induced an emotional reaction. After all, a failed treatment that worsens a condition is likely to induce either reactions of anger or anxiety. In contrast, a successful treatment that improves a condition is likely to induce relief or enthusiasm.

⁵ Crime rate increase and crimes rate decrease as a consequence of refugee intake.

Hence, in light of the findings for the control condition, interpreting the results of Study 5, it becomes evident that the hypothesized and reported emotional reactions due to identity-protection cognition are possibly conflated with emotional reactions due to the content of the information. Results of previous studies face similar problems. For example, Suhay and Erisen (2018) did not include a control condition; thus, possible differentiations of *identity emotions* and what I label as *content emotions* cannot be made.

In addition to the mixed findings concerning the mediating role of emotional reactions, there was no consistent pattern of elicitation concerning all three emotions. While in Study 4 elicitation was mostly found for anger and enthusiasm, in Study 5, the effects of anger were reduced, and elicitation of anxiety and enthusiasm was more pronounced. However, these findings might be explained by different message frames (see also Chapter 4.4).

Taken together, the findings of both studies yield mixed results. While all studies show that emotions are involved in identity-protection cognition, clear inferences about the precise role of emotions cannot be made. It remains uncertain, for example, under which circumstances anger as compared to anxiety is induced. Moreover, results of Study 5 suggest a differentiation between identity emotions as a reaction to identity threat/affirmation and content emotions as a reaction to the semantic content of a message.

4.1.3 Motivated reasoning as identity-protection?

Previous studies have already connected motivated reasoning to identity-based motivation (Van Bavel & Pereira, 2018) and identity-protection cognition (Kahan, 2017; Kahan et al., 2013; Nauroth et al., 2017). Building on these findings, I argue in Chapter 2.5.1. and 2.5.2 for an identity-centric model of motivated reasoning, equating (political) attitudes, opinions, and preferences with (social) identities. Central to this identity-centric model of motivated reasoning is the notion of identity threat and identity affirmation which results in subsequent emotional reactions (Chapter 2.5.4).

Results of Study 4 could support the predictions of an identity-centric model: Threatened identities gave lower ratings of trust, accuracy, and journalistic assessment (Study 4a), whereas affirmed identities provided higher ratings of trust, accuracy, and journalistic assessment (Study 4b). This echoes the findings from Study 1-3, showing that the effects of motivated reasoning on (mis)information acceptance are twofold. Threatened identities in Study 4a rated *true* news worse than non-threatened identities. In turn, affirmed identities in Study 4b rated *false* news better than non-affirmed identities. In addition to the overall effect of identity-protection cognition, in Study 4, I also tested the effect of an identity salience

manipulation to decrease the effect of identity-protection cognition. Mean evaluations changed in the hypothesized direction. However, the differences were not significant.

Moreover, the mean changes also affected those identities which were “protected” against misinformation (incongruency conditions). For example, omnivores in Study 4b, who rejected misinformation favoring vegetarians, seemed to have increased their evaluations of the misinformation after the identity salience manipulation. As a side note, it might be mentioned here that, while no differences between acquired identities and innate identities were assumed, the acquired identity vegetarian (Study 4b) resulted in a stronger effect than the effect of the innate identity gender (Study 4a).

In Study 5, identity-protection cognition was overall less evident as compared to Study 4. The joint measure of political leaning and ideology, which constituted the latent variable political identity, predicted the reasoning outcome only in one of two polarized conditions. Results from logistic regressions revealed that if the cover story conveyed that the data represented the relationship of refugee intake and crime rates, participants’ political identity and not participants’ numeracy predicted responses. Responses for the second polarized condition, the Diesel-car driving ban, were predicted by the participants’ numerical skills, similar to the non-polarized condition (skin rash).

At least two different explanations can account for the inconsistent results for political identity in Study 5. First, the second polarizing condition, the Diesel-car driving ban, was not polarizing enough and, consequently, more similar to the control condition (skin rash). Given that the polarizing potential was pretested and showed the second-highest levels of polarization (after refugee intake), this explanation is rather unlikely. Second, the non-significant results for political identity are a consequence of low attitude-identity alignment. While I have argued in Chapter 2.5.1 that recent advances from the fields of political science and political psychology propose that partisan-ideological sorting has brought partisan identities, ideologies, and subsequent issue positions into alignment (see also Abramowitz, 2010; Kozlowski & Murphy, 2021), these results stem mostly on findings from the USA.

However, ideologies and partisanship might align less with issue positions in different political contexts, such as Germany. Hence, in Study 5, it is likely that the selected identity (political identity) did not align well with the issue position (driving ban). This assumption is supported by previous research showing that Germans generally show reduced levels of partisan alignment (Dassonneville et al., 2014), except for refugee intake and general immigration attitudes (Mader & Schoen, 2019), which supports the significant findings for the refugee intake condition of Study 5. Hence, measuring an individual’s political partisanship or ideology does

not necessarily assess an individual's issue position. This understanding is in line with previous methodological considerations by Washburn and Skitka (2017). The authors point out that "knowing that a person is liberal or conservative [...] will always be an imperfect predictor of the person's position on any given issue" (p. 2). This signifies the importance of the cultural and political context and the intricate relationship of identities, such as partisanship and ideology (Huddy, 2001), and opinions, attitudes, preferences, and beliefs. Hence, the results of Study 5 also raise the question of how identities and attitudes relate to each other.

In line with previous results on cognitive sophistication (Lind et al., 2018; Pennycook & Rand, 2019), findings in Study 5 also indicated that individuals with greater cognitive sophistication skills showed less bias in the refugee intake condition. This suggests that cognitive sophistication skills mitigate the effects of misinformation.

To conclude, understanding motivated reasoning as identity-protection cognition revealed mixed findings. Results of Study 4 support identity-protection cognition and indicate the difficulty of reducing its detrimental impact on individuals' misinformation evaluations. In contrast, the results of Study 5 are less clear. Here, predictions by identity-protection cognition were only supported for one out of two conditions. Moreover, the results of Study 5 make evident that cultural and political contexts shape how motivated reasoning can be assessed. Correlations of identities such as partisanship and political ideology with certain opinions, attitudes, preferences, and beliefs are context dependent. Hence, to understand how identities and attitudes relate to each other, third variables connecting the constructs of identity and attitudes, such as identity/attitude strength, need to be introduced.

4.2 Theoretical implications

The five studies included in the cumulus of this thesis aimed to answer two overarching research questions: (1) How does motivated reasoning affect misinformation sharing on social media, and (2) how can emotions and identity explain motivated reasoning? Besides the immediate results introduced in the previous section, the results of the studies also provide several theoretical contributions concerning motivated reasoning.

Summarizing these contributions, I suggest a refined model of identity-protection cognition (see Figure 4). Following Kahan (2016, 2017), central to the model remain identity threat or identity affirmation as a consequence of incoming (mis)information (Figure 4, path a). In turn, identity threat or affirmation bias subsequent processing on different levels such as perceptions, evaluations, and reasoning (Figure 4, path b). I add to this model several factors which influence these processes: content emotions and identity emotions (see Chapter 4.2.1),

identity strength (see Chapter 4.2.2), credibility cues (see Chapter 4.2.3), and processing style (see Chapter 4.2.4). Before I introduce all added elements in relation to the results of this thesis in greater detail in the following sections, in the following paragraphs, the resulting model is briefly described.

Depending on a person's identity, incoming (mis)information elicits either an identity threat if the (mis)information is incongruent to the identity, or an identity affirmation if the (mis)information is congruent to the identity (Figure 4, path a). In turn, the identity threat/affirmation induces an emotional response such as anger and anxiety as a reaction to threat and enthusiasm as a reaction to affirmation (Figure 4, path c). In the model, these emotional reactions are labeled identity emotions. In addition to these identity emotions, the semantic content of the (mis)information can also induce emotional reactions, affecting identity-protection cognition (Figure 4, path d). Both identity emotions and content emotions are combined through weighted averaging into one unified emotional response (Figure 4, paths e_1 & e_2). The unified emotional response, following the affect-as-information hypothesis (Clore et al., 2001) and the affect heuristic (Slovic et al., 2007), functions as additional information (Figure 4, path f), guiding subsequent processing by increasing the utility of the information's content and increasing the utility of reasoning itself (Blanchette & Caparos, 2013).

Besides the emotions evoked by the semantic content of the (mis)information, it is suggested that identity strength moderates the relationship of (mis)information and identity threat/affirmation (Figure 4, path g). Similarly, credibility cues of the incoming

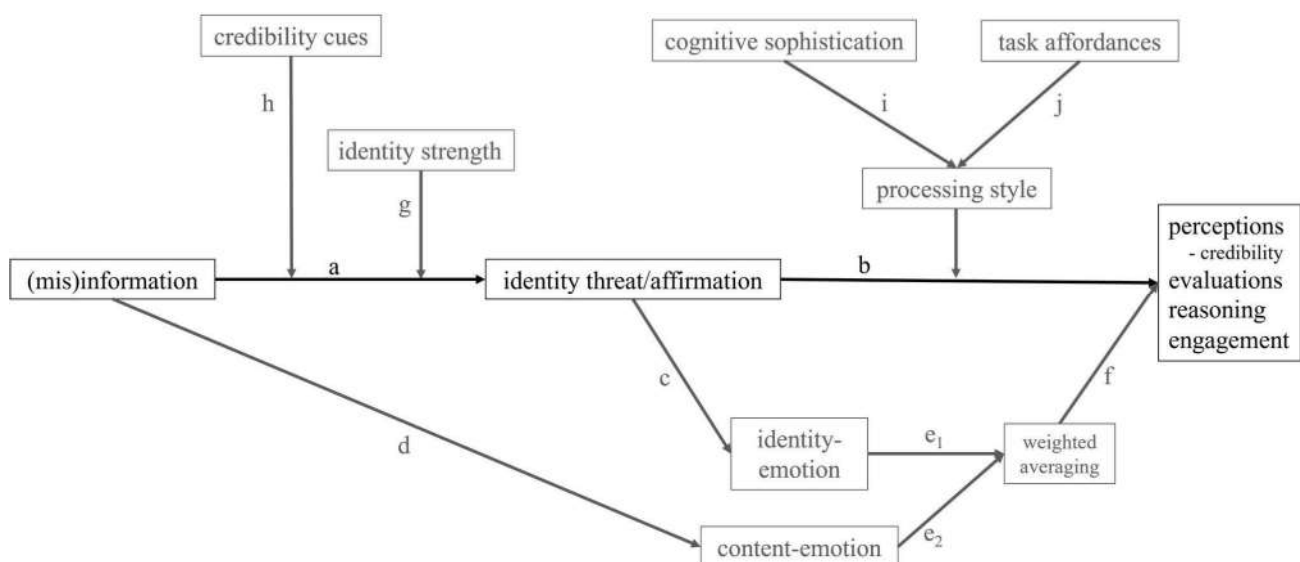


Figure 4: An updated model of identity-protection cognition. Black elements depict the original conceptualization of identity-protection cognition by Kahn (2016; 2017) (paths a & b). Orange elements depict additions made based on the results of this thesis (paths c-j).

(mis)information such as signs of automation moderate the threat/affirmation experience (Figure 4, path h). Also included in the updated model of identity-protection cognition are cognitive sophistication (Figure 4, path i) and task affordances (Figure 4, path j) which both contribute by affecting the processing style.

4.2.1 Content-emotions and identity-emotions

Throughout this thesis and Study 4 and 5, I propose that the elicitation of the emotions anger, anxiety, and enthusiasm due to identity threat or affirmation is inherent to identity-protection cognition. In doing so, I follow previous results by Suhay and Erisen (2018), who found that anger, anxiety and enthusiasm mediated motivated reasoning (see Chapter 2.5.3). However, testing these assumptions in Study 4 and Study 5 yielded mixed results (see Chapter 4.1.2). In both studies, it becomes evident that emotions contribute to identity-protection cognition, supporting previous findings. However, due to mixed findings in both studies, clear inferences about the precise role of emotions cannot be made. Especially, results from Study 5 indicate that while it was intended to measure emotional reactions elicited by identity-protection cognition, emotions elicited by the content likely diluted these measures. This became particularly evident when inspecting the control group's results, in which participants indicated similar levels of emotional reactions compared to the two experimental groups. Consequently, it must be concluded that the elicitation of emotions was confounded with the elicitation of content emotions which neither Suhay and Erisen (2018) nor I accounted for.

Previous results scrutinizing mood induction procedures (MIP) support this hypothesis by providing empirical evidence that emotions can effectively be induced through different techniques such as films, images, faces, sounds/voices, music, imagery and recall, words and bodily movements and postures (Quigley et al., 2014; Westermann et al., 1996; Velten, 1968). While most of these findings refer to face-to-face interventions, Verheyen and Göritz (2009) found evidence supporting similar effects for web-based online experiments with plain texts. Although not intended, especially the stimulus material of Study 5, potentially induced emotions of anxiety or anger by describing increases in crime rates and decreases in air quality—two issues likely to directly affect the physical integrity of participants. Similarly, experimental stimuli selected by Suhay and Erisen (2018) could induced emotions such as anxiety due to potential physical and psychological harm (Suhay & Erisen, 2018, Appendix A: “Women who have abortions often suffer negative physical consequences and mental anguish.”), economic threat (Suhay & Erisen, 2018, Appendix A: “Illegal immigrants are an economic burden on the country.”), and anger due to perceived injustice (Suhay & Erisen, 2018,

Appendix A: “People who are breaking the law by being in the country illegally should be punished, not rewarded, for their unlawful behavior.”).

To conclude, as results of Study 5 indicate, two possible sources can evoke an emotional response: identity threat/affirmation (Figure 5, path c) and the semantic content of the (mis)information (Figure 5, path d). While emotions elicited through the stimulus material are not inherent to identity-protection cognition, they can, nevertheless, impact identity-protection cognition. However, does an individual experience two emotional responses simultaneously or are both responses joined into one?

Previous research suggests that, if two or more emotion arousing stimuli are simultaneously present or quickly follow each other, all emotions are subsequently integrated into one unified emotional response (Asutay et al., 2019; Asutay et al., 2020; Cunningham et al., 2013; Kuppens & Verduyn, 2017). For example, Asutay and colleagues (2019) presented participants with a sequence of affectively laden pictures that varied between pleasant and unpleasant and three different levels of arousal. After being presented to each sequence, participants reported their experienced affective states. In the next step, the authors tested the impact of the temporal order of the stimuli through multi-level modeling. Results obtained by Asutay et al. (2019) suggest that affective responses are integrated via weighted averaging, “in which higher weights are assigned to more recent and potentially more potent stimuli” (p. 172).

Transferring these insights to the results of Study 5, I suggest that both emotional experiences should be integrated into one uniform emotional experience via an averaging process (Figure 5, path e₁, e₂, and f). The result of such an averaging process is determined, as

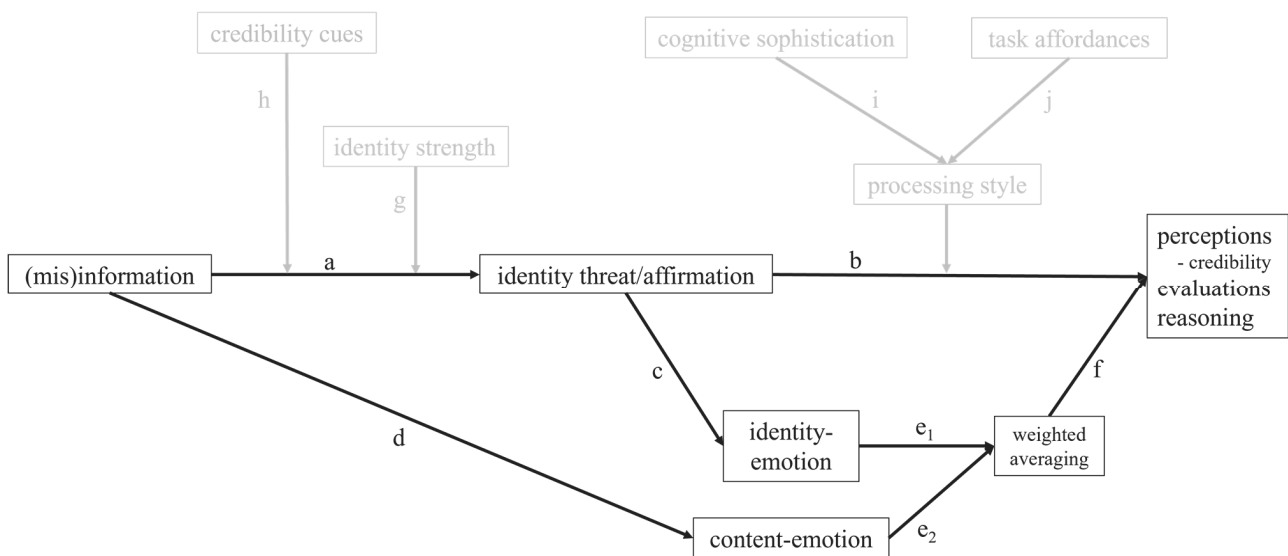


Figure 5: Content emotion and identity emotion mediating identity-protection cognition through weighted averaging (highlighted in black print).

Asutay et al. (2019) suggest, by the more potent stimulus. For example, if the semantic content of (mis)information triggers a stronger emotional response (e.g., fear), this response might outweigh a cooccurring emotional response (e.g., enthusiasm) induced by an identity affirmation. Concretely, results obtained in Study 5 (see Chapter 4.1.2) could be interpreted in a way that, reading (mis)information about increased crime rates, conservatives/right leaning individuals might have reacted with both fear induced by the content and enthusiasm induced by the identity affirmation.

Which of the two emotional responses, content emotion or identity emotion, dominates might result from, for example, situational factors such as credibility cues of the message (see also Chapter 4.2.3). If the message comes from a highly credible source, the negative consequences might be highlighted by increased credibility, and the fear response might dominate the overall emotional response. Once uniformed, the emotional reaction guides subsequent responses, following the affect heuristic (Slovic et al., 2007), as found in Study 5.

To conclude, based on findings of Study 5 (emotional reactions of the control group), previous empirical findings (e.g., Asutay et al., 2019), and predictions by the affect heuristic (Slovic et al., 2007), I suggest that content-related emotions affect identity-emotions through weighted averaging processes, mediating the effect of threat/affirmation induced emotions on the response.

Because misinformation have been found to be tailored to inflict certain emotions in the news consumers (Brady et al., 2017; Freiling et al., 2021; Sanchez & Dunning, 2021; Wang et al., 2020), for example through framing of the message (Lecheler et al., 2015), it becomes crucial to incorporate content-related emotions in future investigations of identity-protection cognition. Differentiating such content-related from identity-protection-related emotions is, hence, a challenge for future studies on identity-protection cognition (see Chapter 4.4).

4.2.2 Identity and identity-constituting attitudes in identity-protection cognition

Across studies 4 and 5, the hypothesized identity threat/affirmation induced bias occurred inconsistently. Results of Study 4, which targeted gender identities and dietary identities, followed the predicted pattern of identity-protection cognition (see Chapter 4.1.3). In contrast, results of Study 5, targeting political attitudes, which were treated as identities, yielded mixed results (see Chapter 4.1.3). Only for some cases, it seems, that political attitudes can be equated with identities.

To make sense of these mixed findings, I suggest that a third variable likely defines the association of identities and political attitudes. One possible candidate to relate identities and political attitudes is *identity strength* which describes “the importance or psychological

attachment that individuals place on their identities” (Settles, 2004, p. 487). Consequently, stronger identifiers are more likely to hold strong attitudes concerning the specific identity than weaker identifiers. Especially social and group identities can give rise to certain attitudes (Mason, 2015; Spears, 2021). Similarly, strong attitudes can also give rise to strong identities, with strong attitudes becoming an expression of identities (Bliuc et al., 2007; Jones, 1999; Musgrove & McGarty, 2008). For example, as a member of a specific political group, it might be expected to hold a specific political opinion.

In addition to the added connection of identities and attitudes, identity strength is also likely to moderate identity-protection cognition. Leeper and Slothuus (2014) suggest that some identities are stronger than other identities through higher levels of importance. In turn, I assert that stronger identities are more likely to elicit identity threat/affirmation than weaker identities, moderating the (mis)information threat/affirmation link. Previous empirical findings support this view. For example, Visser et al. (2016) found that participants defended important and strongly held attitudes more than weaker attitudes. Similarly, Taber and Lodge (2006) found that identity strength moderated the effects of motivated skepticism.

While previous studies understand identity/attitude strength as a moderator of motivated reasoning in general (Taber & Lodge, 2006; Visser et al. 2016), I understand identity strength as a moderator of the threat/affirmation perception, with strong identities increasing threat/affirmation perceptions and weak identities decreasing such (see 6, path g). I see indirect support for this hypothesis in results from Study 5. The mediation analysis showed that political identity and emotional reactions were positively correlated, meaning that the more a person

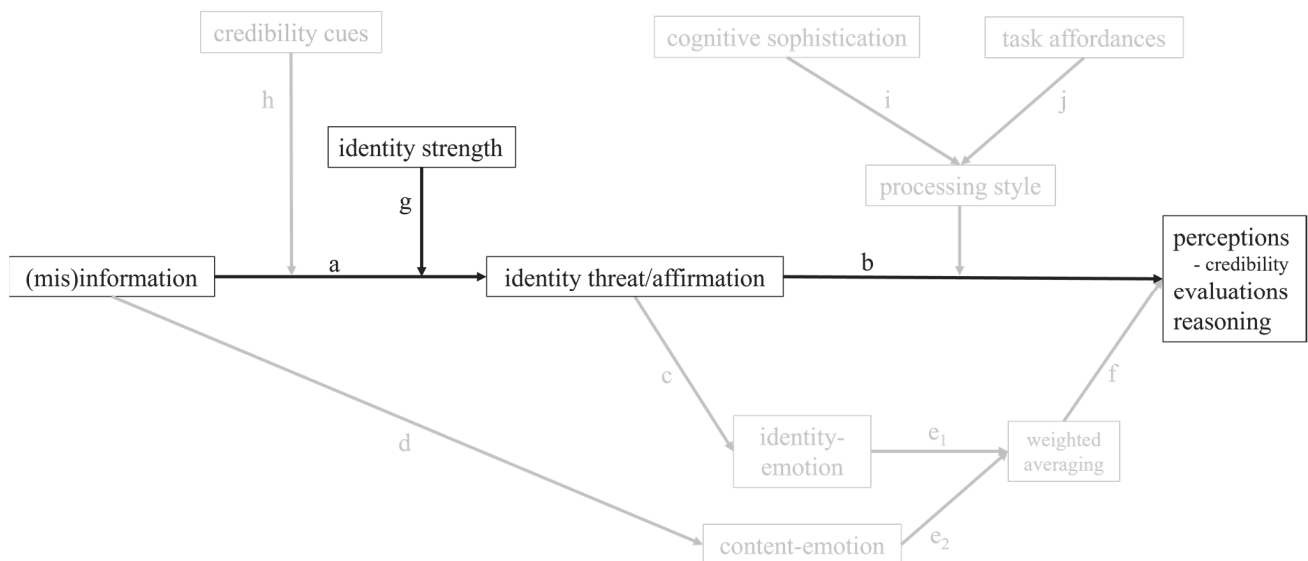


Figure 6: Identity strength moderating the (mis)information identity threat/affirmation link (highlighted in black print).

held a political identity, the greater the emotional reaction. If the emotional reaction results from identity threat/affirmation, this would imply the suggested relationship of identity strength and threat/affirmation perception. Previous findings also support this hypothesized relationship, showing that increased identity strength increased perceived identity threat (Branscombe et al., 1999; De Hoog, 2012).

To conclude, based on the mixed results concerning the role of attitudes and identities in Study 4 and 5, in this section, I derive that identity strength likely affects the manifestation of identity-protection cognition by (1) increasing or decreasing the relevance of an identity, and, consequentially, (2) increasing or decreasing identity threat/affirmation.

4.2.3 Credibility cues and identity-protection cognition

In Study 2, it was demonstrated that the biased perceptions of the Twitter profiles could partly be explained through biased credibility judgments. It must be noted that, with the causal link of credibility judgements and profile perceptions assumed only theoretically but not experimentally manipulated, it is also plausible that biased perceptions lead to biased credibility judgements. While the notion of causality poses interesting future challenges, results of Study 2 undoubtedly reveal that credibility perceptions are affected by identity-protection cognition.

These results align with prior research examining confirmatory heuristics in credibility judgments (Metzger & Flanagin, 2013; Sundar, 2008; Sundar et al., 2009; Winter et al., 2016). For example, investigating the credibility of online sources through focus group data, Metzger et al. (2010) found that, instead of deliberate information processing, participants commonly relied on different heuristics such as “social confirmation of personal opinion” (p. 423), which indicates that users are more likely to view attitude-congruent sources as credible. In turn, participants were more likely to view attitude-incongruent sources as incredible. Findings of Study 2 echo these results, indicating that attitude-congruent social bots were perceived as more credible than attitude-incongruent social bots. This indicates that credibility perceptions are affected by identity-protection cognition, similar to evaluations and reasoning (see Figure 7, end of path b).

Results of Study 3 point to less evident effects of credibility. In Study 3, the two factors, the humanness of an account and the displayed partisanship an account, interacted. For incongruent accounts, the level of humanness did not matter. All bot accounts, from low to highly human-like, were mostly likely ignored. In contrast, level of humanness mattered for congruent accounts: low humanness significantly decreased participants’ willingness to engage with the account. Consequently, participants did not blindly engage in identity-protection but incorporated stimulus cues into their engagement intentions.

This is in line with previous research which found that participants employ *expectancy violations heuristics* to assess the credibility of information and sources (Metzger et al., 2010; Metzger & Flanagin, 2013). In Metzger et al. (2010), for example, participants reported perceiving websites as less credible which displayed low professionalism, indicated by typos or grammatical errors, poor site design, and poor visual appearance.

This connects to results found in Study 3. The low humanness of the Twitter accounts, operationalized through repetitive behavior that indicated automation, could have been interpreted as low professionalism, increasing the likelihood of participants disengaging with the accounts. For incongruent accounts, it was enough to display an identity and respective attitudes contrary to an users' identity to elicit an identity threat which, in turn, lead to blatant disengagement. For congruent accounts, displaying a congruent identity was not enough to elicit an identity affirmation, leading to increased engagement intentions. Accounts also had to indicate high levels of humanness.

Based on these results, I propose that (mis)information that display signs of incredibility should be more likely to be discounted because they elicit less identity affirmation or identity threat and, hence, lead to less identity-protection cognition (see Figure 7, path h). The effects of different credibility cues on the elicitation of threat and affirmation pose interesting future research questions.

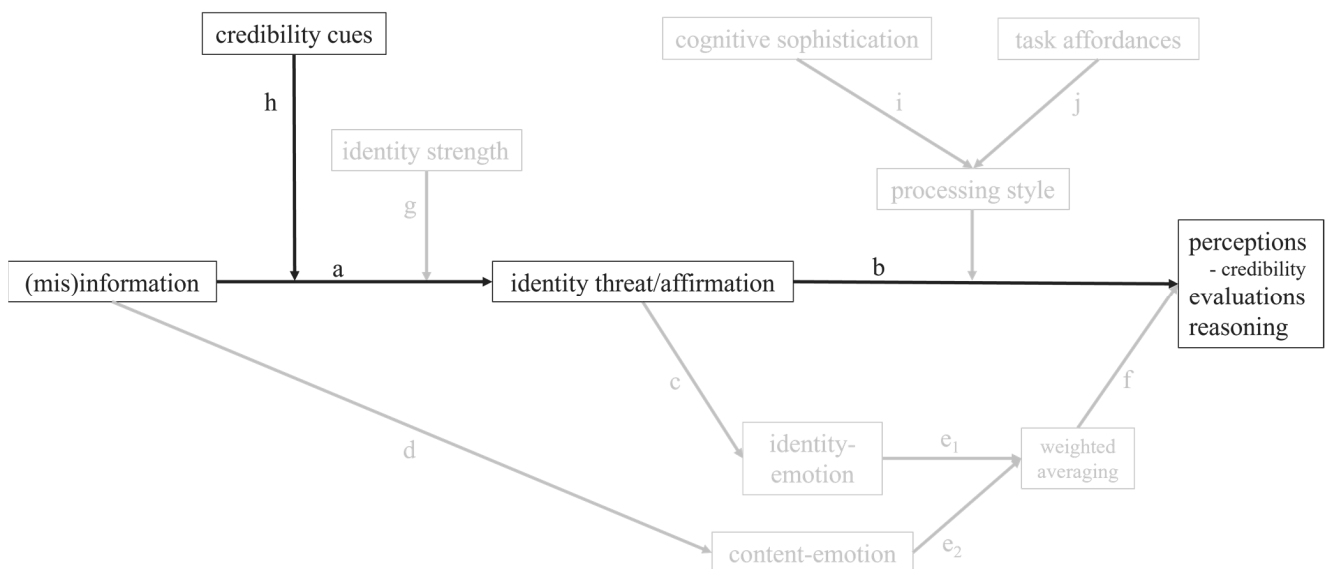


Figure 7: Credibility as a result of identity-protection cognition (end of path b) and credibility cues moderating the (mis)information identity threat/affirmation link (path h) (highlighted in black print).

4.2.4 The influence of processing style on identity-protection cognition

Study 5 scrutinized the moderating influences of cognitive sophistication. For one experimental condition (crime decrease), results indicate that cognitive sophistication affected identity-protection cognition so that political identity was less likely to affect the reasoning outcomes for individuals with higher cognitive sophistication skills. The result supports one strand of previous research (Lind et al., 2018; Pennycook & Rand, 2019; Vegetti & Mancosu, 2020), showing that individuals with higher cognitive sophistication skills were less likely to engage in identity-protection cognition. The underlying assumption of these studies is that preferences supporting and protecting prior attitudes and identities result from heuristic processing. Consequently, individuals who are less inclined to deliberate and are more likely to rely on heuristic processing are worse at differentiating between false and true information (Pennycook & Rand, 2019; Vegetti & Mancosu, 2020) and are more likely to arrive at predefined conclusions (Lind et al. 2018). In contrast, Kahan et al. (2017) found that cognitive abilities increase the effects of motivated reasoning, indicating that individuals with greater cognitive ability are better able to rationalize identity and attitudes consistent information to be accurate (see Chapter 2.5.5), which was, however, not supported by results of Study 5.

Moreover, results of Study 5 for the conditions crime increase and both Diesel ban conditions indicated non-significant results for cognitive sophistication, which is also supported by previous findings, indicating no correlation between motivated reasoning and cognitive abilities (Stanovich & West, 2008a, b, 2007). Stanovich and West (2007) explain significant results for cognitive abilities and null results by referring to the experimental design. The authors found that correlations are usually found in studies employing within-subjects designs, whereas between-subjects designs usually find null results. Consequently, Stanovich and West (2007) suggest that within-subjects designs likely raise participants' awareness for variables of interest. In turn, increased awareness for variables of interest functions as a cue to decontextualization and detachment of one's current perspective and a cue for more deliberative processing. However, under "naturalistic reasoning situations, participants of high cognitive ability may be no more likely to recognize the need for decontextualization than participants of low cognitive ability" (p. 240/241). Hence, with Study 5 employing a within-subjects design, results of cognitive sophistication may have been the result of the experimental set-up.

Although the exact impact of cognitive sophistication needs to be investigated in greater detail in future studies, following the results of Study 5, I included cognitive sophistication into the updated model of identity-protection cognition. Results of Study 5 and previous findings suggest that individuals with a predisposition for higher cognitive sophistication skills are more

likely to process information systematically, decreasing the effect of identity threat/affirmation (see Figure 8, path i).

If deliberation affects identity-protection cognition as Study 5 and previous studies suggest, task affordances might also cue deliberation, affecting identity-protection cognition. Research by Stanovich and West (2008a,b) suggests that tasks differ in the extent to which they promote cognitive decoupling (detachment of one's current identity/attitude). In that case, Studies 2, 3, and 4 might have been more prone to induce identity-protection cognition as all three studies asked participants to evaluate either Twitter profiles or information presented to them. In contrast, in Study 5, participants were confronted with a complex inference task that might have been more likely to promote deliberate processing (Stanovich & West, 2008a, b). Hence, task affordances are also included in identity-protection cognition (see Figure 8, path j).

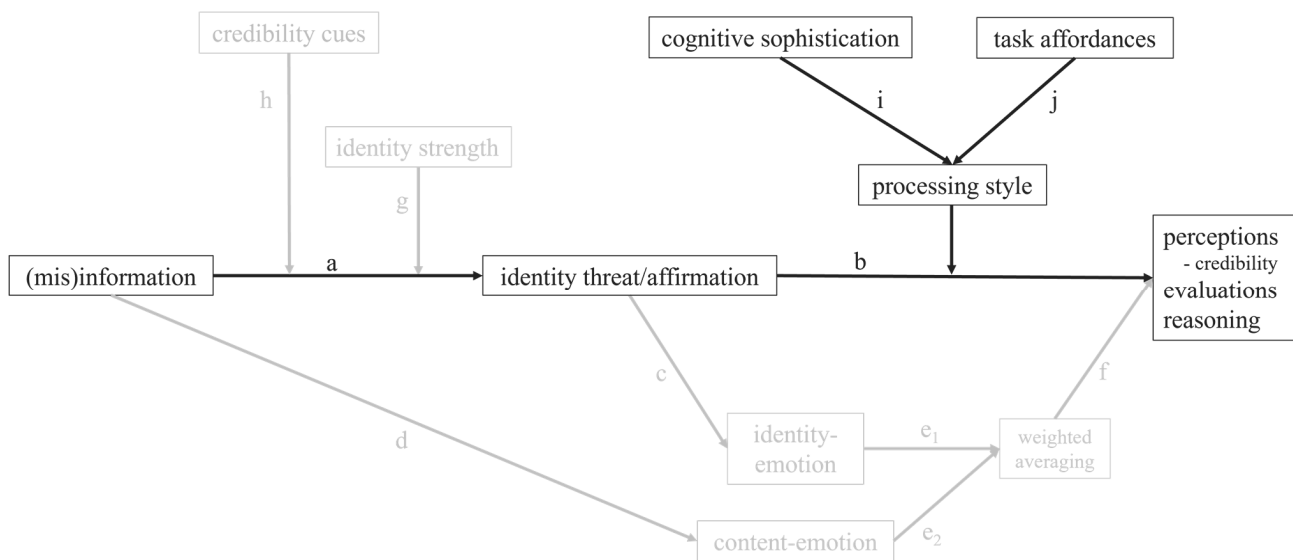


Figure 8: Effects of cognitive sophistication and task affordances on the processing style (highlighted in black print).

4.3 Practical implications

Motivated reasoning affects misinformation by obscuring what is perceived as accurate and false (Clayton et al., 2019; Flynn et al., 2017; Pennycook & Rand, 2019; Vegetti & Mancosu, 2020). In Study 1-3, I extend this finding to misinformation sharing. Study 1 shows that motivated reasoning directly affects which hyper-partisan news are shared within a network. Furthermore, Studies 2 and 3 indicate that motivated reasoning affects how users perceive and engage with misinformation sharing entities: social bots. Similar to Study 1, it is found that motivated reasoning influences how these automated accounts are perceived

concerning their humanness, credibility and how likely human users are to engage with social bots.

All three studies indicate that the impact of motivated reasoning is twofold: On the one hand, users become more vulnerable to sharing misinformation and perceiving malicious social bots as credible sources of information. This is the case when misinformation confirms previous beliefs or is shared by alleged members of one's (political) in-group (Anthony & Moulding, 2019; Ecker et al., 2014; Kahne & Bowyer, 2017; Kuklinski et al., 2000; Nyhan & Reifler, 2010). On the other hand, motivated reasoning also functions as a protection against misinformation sharing. This is found when misinformation contradicts previous beliefs or is shared by members of one's (political) out-group. Consequently, the common framing of motivated reasoning causing individuals to believe and share misinformation is valid but also one-sided. Arguing, for example, like Kahne and Bowyer (2017) that "in the presence of misinformation, directional motivated reasoning has unambiguously negative implications for democratic deliberation." (p. 9) falls short of acknowledging motivated reasoning's positive effects. It is equally valid to state that motivated reasoning causes individuals *not* to believe and share misinformation. As I will elaborate in this section, this differentiation is important because possible interventions aiming to reduce motivated reasoning might unintentionally also reduce the protective effect of motivated reasoning.

Generally, because of its two-fold effect on misinformation, there is no easy way around motivated reasoning. Quick to mind might come the idea to reduce motivated reasoning through interventions. However, before implementing possible interventions to reduce motivated reasoning, it is important to ask (1) for whom should it be reduced, (2) under which circumstances should motivated reasoning be reduced, and (3) how should it be reduced?

First, the results of this thesis suggest that interventions intended to reduce motivated reasoning should be applied with great care concerning the targeted individuals. Otherwise, researchers might unwillingly also decrease the protective qualities of motivated reasoning for some individuals. With limitations, this could be observed in Study 4. Here, the manipulation was intended to reduce the salience of a threatened identity (Study 4a) and to reduce the salience of an affirming identity (Study 4b). After the identity salience intervention, mean differences between threatened and non-threatened (Study 4a) and affirmed and non-affirmed (Study 4b) participants (marginally significantly) decreased, especially for the group which received misinformation.

Following the idea of an intervention, it would be necessary to identify individuals who would be more vulnerable to misinformation due to motivated reasoning. However, this poses

great challenges, as it requires knowledge about both the individual's stance as well as the misinformative content. In cases of known polarizing issues, identifying these individuals might be less complicated, as Study 1 indicates. In Study 1, some Twitter users' profile descriptions successfully served as a proxy for individuals' stance. This allowed identifying which individuals were more likely to share attitude-congruent hyper-partisan news. While the individual's stance might be inferred, an intervention also requires detection of misinformation, which poses entirely new challenges such as a lack of gold-standard to train algorithms, multi-lingual problems, cross-platform detection, and time constraints (Meel & Vishwakarma, 2020).

In contrast to interventions aiming at misinformation, interventions aiming to increase accurate information acceptance and sharing might be easier to implement, especially for known controversies. For example, Lewandowsky et al. (2012) and colleagues suggest presenting initially threatening information (e.g., this could also be a debunking message) in an affirming manner by “focusing on opportunities and potential benefits rather than risks and threats” (p. 123).

Second, understanding under which circumstances to reduce motivated reasoning through interventions raises mostly philosophical and ethical questions. After all, categorizing something as false or misleading implies that the actual truth is known, leading epistemological questions and questions concerning power dynamics.

To answer both, for the context of misinformation, I suggest that two sub-classes of misinformation emerge, varying in the degree to which questions of epistemology and power can be applied: (1) episodic misinformation and (2) semantic misinformation. With episodic or historical misinformation, I refer to any false claims about the occurrences of specific episodes of the past. This can refer to the sequence of events or statements of public figures such as quotations taken out of context. For such episodic misinformation, questions of truth can be answered and debunked through, for example, (eye) witnesses of events and photo or video footage. A prominent example of episodic misinformation was the 2017 inauguration ceremony of Donald Trump⁶. After the event, disputes over the number of attendees of the ceremony compared to the 2009 inauguration of Barack Obama arose. Donald Trump claimed the size of the audience at his ceremony exceeded the numbers of Obama's ceremony—a false claim which transit data and photographic footage revealed. However, in the future, advances in deep-learning technologies can make it more difficult to differentiate true from false claims by synthetically fabricating/manipulating audio and video sources (Agarwal et al., 2020).

⁶<https://www.theguardian.com/us-news/2017/jan/22/donald-trump-kellyanne-conway-inauguration-alternative-facts>

Nevertheless, for interventions to reduce motivated reasoning, the truth can more easily be identified and does not involve the question of epistemology.

In contrast to episodic misinformation, for semantic misinformation epistemology and power are more relevant. In that, I refer to semantic misinformation as *knowledge* or *empirical evidence*. However, contemporary theories in philosophy such as postmodern constructivism and critical theory suggest that there is no objective knowledge, i.e., objective truth, as knowledge is always mediated through subjective perceptions. In times of misinformation and ‘post-truth’, these theories have faced harsh criticism, being accused of supporting the erosion of facts (Flatscher & Seitz, 2018; Latour, 2004).

To know facts, for empirical research in psychological science, Camina and colleagues (2020) investigated the operationalization of beliefs. In doing so, the authors define knowledge as something “that is supposedly true in some objective sense” (p. 331), referring to agreed-upon objective foundations such as evidence-based discourse. Following Camina et al.’s (2020) epistemological stance and definition of knowledge, I refer to semantic misinformation as any false claims contradicting any contemporary evidence-based discourse of that point in time. This makes truth, inevitably, tentative⁷ but also allows to identify misinformative claims about, for example, anthropogenic climate change (Druckman & McGrath, 2019; Farrell, 2019), vaccinations (Lewandowsky et al., 2012), or genetically modified food (Valenzuela et al., 2019). Additionally, in Michel Foucault’s (2008) work, the author adds to the tentative nature of knowledge that, on the one hand, knowledge is shaped by power, i.e., shaped by institutions such as political systems or societal norms, and that, on the other hand, knowledge induces power. Hence, implementing misinformation corrections inherently means to employ and to gain power.

With the tentative nature of knowledge and its entrenchment in power in mind, what does this imply for interventions (such as the reduction of motivated reasoning) aiming to decrease beliefs in misinformation? I argue that it implies to carefully construct interventions to counter false claims about evidence-based facts. Applying this position to the five studies of this thesis, the studies fall into two categories: In Study 1, Study 4, and Study 5, interventions would be adequate as all three studies deal with evidence-based facts. Study 1 might be the least adequate of the three, as it investigates hyper-partisan news, which frequently oscillates between facts and attitudes (Faris et al., 2017). In contrast, Study 2 and Study 3 investigate how users perceive other *users* (in this case, automated accounts) in social networks. While the

⁷At some point in time, for example, it was considered true that the sun orbits the earth. Consequently, statements negating this claim would fall into the category of semantic misinformation. Following Copernicus’ discovery, it is today, however, common knowledge that the earth orbits the sun.

effects of motivated reasoning might make users more susceptible to misinformation sharing accounts, it can hardly be legitimized to intervene in how individuals connect with others on social media.

After discussing for whom motivated reasoning should be reduced and under which circumstances it should be reduced, I turn to the question, *how* should it be reduced? Previous literature suggests various strategies to overcome identity-protection cognition, which I will connect in the following paragraphs to implications derived from the updated model of identity-protection cognition.

The intervention probed in Study 4 yielded mixed results. First, different interventions have been pursued to change or relax individuals' prior attitudes or identity salience by instructing participants to be open-minded or to put their feelings aside (Martel et al., 2019; Taber & Lodge, 2006), affirming participants' identities (Bayes et al., 2020; Cohen et al., 2000), or instructing participants to change their perspective (Lilly, 2012). Another idea suggests shifting identity salience from a threatened (affirmed) identity to a non-threatened (non-affirmed) identity to reduce identity-protection cognition. Although such identity shift interventions and open-mindedness instructions or identity affirmations pose possible solutions to identity-protection cognition in the lab, applying such interventions on a large scale *outside* the lab will be (technologically) challenging.

One possible solution to translate in-lab studies to a larger context outside the lab comes from Pennycook et al. (2021). Under the premise that increased attention to accuracy would reduce misinformation sharing, the authors selected 5379 Twitter users who had previously shared news from untrustworthy sources. These users received a private message asking them to assess the accuracy of a non-partisan news item. This simple accuracy prompt significantly reduced misinformation sharing. While Pennycook et al. (2021) assume that the accuracy prime increased general attention for the accuracy of news, in the context of identity-protection cognition, it might also be that the accuracy prime evoked an accuracy motivation challenging the perceived identity threat/affirmation. However, future studies should explore in greater detail how competing motivations affect identity-protection cognition.

Second, it has been suggested that individuals with higher media literacy skills better differentiate between accurate and false information. While Kahne and Bowyer (2017), for example, found correlational prove supporting the hypothesis that increased media literacy increased news discernment, results of Study 2 reported in this cumulus showed no effect of media literacy increasing the discernment of human versus social bot accounts. Going beyond correlational prove, studies in which participants underwent a media literacy training increased

participants' skills to differentiate accurate versus false news. However, these interventions did not aim to reduce identity-protection per se but, for example, informed users more generally how content is produced or informed participants about untrustworthy cues of misinformation (Hameleers, 2020; Vraga et al., 2020).

One of the few studies, targeting particularly identity-protection cognition, comes from Sivek (2018). In her essay "Both facts and feelings: Emotion and news literacy", Sivek argues that existing media literacy training programs, while addressing factual issues such as questionable URLs or missing sources, hardly address the role of emotion in falling for misinformation. Taking the example of Checkology⁸, Sivek (2018) demonstrates that the checklist "encourages users to think first about emotions elicited by stories" (p. 130) but does not assist users in dealing with emotions elicited. Sivek's argument is in line with studies discussed throughout this thesis (Chapter 2.5.3) as well as the findings regarding the role of emotions in identity-protection cognition of Study 4 and 5 (Chapter 2.5.4). According to Sivek (2018), possible ways to include emotional reactions are mindfulness techniques, increasing general emotional awareness, and training programs that inform participants about thinking processes such as identity-protection cognition.

To conclude, in this chapter, I discussed how the results of this thesis could be used to inform interventions aiming to reduce identity-protection cognition. I elaborated how such interventions pose not only technical challenges but also pose ethical questions.

4.4 Limitations and future studies

After extending the previous model of identity-protection cognition and elaborating possible practical implications in the two previous sections, I acknowledge the theoretical and methodological limitations of the results discussed and conclusions drawn.

4.4.1 Emotions

As already discussed in Section 4.2.1, the emotional reactions assessed in Study 4 and 5 might have been elicited by the semantic content of the (mis)information as well as the assumed identity threat/affirmation. To carefully dissect these two sources' emotional reactions will be a challenge for future studies. In addition, the assessment of emotional reactions in Study 4 and 5 was limited to self-report measures of emotions. Because self-report measures of emotions comprise their limitations (see Marcus et al., 2006), such as they require participants to reflect on their emotional state, diluting possible emotional reactions, another possibility for future studies is to assess emotional reactions through psychophysiological measures. For

⁸ A non-partisan, educational nonprofit news literacy project based in the USA: <https://checkology.org/>

example, Boyer (2021) used skin conductance level and facial electromyography measures to approximate valence and arousal elicited through TV news items. The inclusion of such psychophysiological measures has also been found to significantly increase explainable variance of statistical models (Asutay et al., 2019).

Beyond measurement challenges of emotions, alternative theoretical explanations for emotional reactions and alternative implications of emotions on information processing need to be discussed. The updated model of identity-protection cognition, introduced in Chapter 4.2, follows an appraisal theory of emotion (Smith & Lazarus, 1993). The respective emotions aroused are the result of the appraised identity threat/affirmation, “inform[ing] the individual that the event is relevant and may inform the individual of his/her perceived coping ability” (Harmon-Jones, 2010, p. 190). For the case of identity-protection cognition, the coping ability manifests as counter-arguing, dismissing, miscrediting, or downplaying of (mis)information that threatens an identity and acceptance without further scrutiny of identity affirming (mis)information. While some emotional theorists assume that the appraisal of the (mis)information is unconscious to the individual (Clore et al., 2001; Harmon-Jones, 2010; Lazarus, 1995), others propose that emotional reactions are a result of conscious appraisal (Smith & Lane, 2015). Suppose emotional reactions follow a cognitive, conscious appraisal of the incoming (mis)information. In that case, the emotional reaction might be simply a *result* of identity-protection cognition, similar to counterarguing or downplaying, instead of a mediator (similar to Study 4). While the problem of conscious versus unconscious appraisal poses an interesting endeavor for future research, I suggest to integrate affective responses at the preconscious level (see also Lodge and Taber's, 2013, and Redlawsk's, 2002, hot cognition hypothesis), eliciting a conflict signal which is, in turn, (consciously or unconsciously) appraised as an identity threat/affirmation. In a second step, as a reaction to identity threat/affirmation, discrete emotions such as anger, anxiety, and enthusiasm are elicited, which guide information processing following the affect-as-information hypothesis (Clore et al., 2001) and the affect heuristic (Slovic et al., 2007) (see also Chapter 4.2.1)

Besides alternative theoretical explanations for the elicitation of emotions, at least two alternative implications of emotions can be proposed. First, in affective intelligence theory, Marcus et al. (2000) assume that anxiety motivates individuals to allocate greater attention to a stimulus, resulting in more careful information seeking and processing and less reliance on habit. In contrast, according to Marcus et al. (2000), enthusiasm and anger trigger reliance on heuristic processing strategies such as reliance on prior attitudes (see also: Chapter 2.5.4). Incorporating affective intelligence theory, findings by Weeks (2015) suggest that anger

facilitates the effects of motivated reasoning. In contrast, anxiety alleviates its effects—establishing emotions as a moderator of motivated reasoning and, consequently, identity-protection cognition.

Second, previous research investigating the effects of emotions on information processing found that emotions, positive just as negative, decreased cognitive performance (Blanchette & Richards, 2003; Cheung-Blunden & Ju, 2016; Jung et al., 2014). The underlying theoretical assumption of these studies is that emotions negatively affect reasoning by taking up needed working memory capacities which was also empirically found (Viau-Quesnel et al., 2019). However, this conceptualization of emotion does not support the results of Study 5, in which anxiety and enthusiasm both decreased and increased (depending on individuals' stance) performances. Concludingly, future studies might assess in greater depth the effect of emotions on processing.

So far, the updated model of identity-protection cognition has also been silent on when to expect anger as a reaction to identity-threat compared to anxiety. At the most, it was introduced in the theoretical part of this thesis that anger is likely to be elicited due to a perceived violation of one's standards (Carver & Harmon-Jones, 2009). In contrast, anxiety would result from lacking personal control and increased uncertainty (Eysenck et al., 2007) (see Chapter 2.5.4). Previous research suggests that anger and anxiety can be a result of the message frame (Nabi, 2003; Lecheler et al., 2015). For example, prior studies investigating news reporting on immigrants could show that reporting often centered either on reports eliciting fear by highlighting economic or individual consequences (Boomgaarden, 2007) or reports eliciting anger by highlighting undermining in-group norms (Verkuyten, 2018). Hence, predicting which emotion, anger, anxiety or a mix of both is elicited would result from how the identity threat is framed.

4.4.2 Identity threat/affirmation

Similar to previous studies on motivated reasoning in general (e.g., Bolsen et al., 2014; Kunda, 1987; Pennycook & Rand, 2019) and identity-protection in specific (Kahan et al., 2017; Lind et al., 2018), in all studies of the cumulus, identity threat/affirmation was *assumed* when participants were confronted with incongruent/congruent information. This assumption limits my findings in two ways: First, methodologically, identity threat/affirmation was neither measured through validated scales nor experimentally induced. However, only experimental induction would ascertain identity threat/affirmation to be the cause of biased perception/reasoning. Previous studies employing such identity threat/affirmation have, for example, asked participants to write texts about fearful/happy events of their past (e.g., Weeks,

2015). The evoked identity threat/affirmation is, in turn, incidental and not integral to the afterwards presented (mis)information, moderating and not mediating downstream cognition via an emotion-induced allocation of attention (Gable et al., 2015).

Second, limiting findings theoretically to confront participants with either identity incongruent or congruent (mis)information implies that identities/identity-constituting attitudes existed a priori. However, misinformation such as health misinformation (Suarez-Lledo & Alvarez-Galvez, 2021) are not necessarily polarizing in nature. For example, cancer information circulated on social media might not target any political identity. However, previous research has found that roughly one-third of cancer information circulated on Facebook was false (Gage-Bouchard et al., 2018). Instead of identity-protection mechanisms, reasoning about such non-polarizing misinformation might be driven by the emotions elicited by the semantic content of the information or credibility cues of the message (paths e & h in the model).

In addition, and especially in light of Study 4, differences in identity acquisition might have also shaped the manifestation of identity-protection cognition. Generally, identities can be acquired, such as one's political identity or occupational identity. However, some identities are ascribed, such as one's gender or ethnicity. Examining the difference between acquired and ascribed identities, Turner et al. (1984) found, for example, that group commitment was stronger when identities were acquired as compared to ascribed. The authors explain this finding by referring to cognitive dissonance theory (Festinger, 1957): Individuals who chose (acquire) their identity need to compensate for possible negative aspects of the chosen identity via enhanced justification that one's choice was good. Results of Study 4a and 4b support this finding. The effect of identity-protection cognition was larger for the acquired identity vegetarians than for the ascribed identity of women. Transferring these considerations to the context of misinformation, which has been closely connected to politics and political identities, Jerit and Zhao (2020) suggest that acquired political identities might exacerbate the effects of identity-protection cognition, making individuals even more vulnerable to congruent political misinformation. Consequently, future research investigating the role of identity acquisition poses interesting research avenues and important contributions in combatting misinformation.

4.4.3 Individual differences

Besides individuals' inclination for systematic processing (cognitive sophistication), which I included in the updated model of identity-protection cognition, other individual dispositions are likely to affect identity-protection cognition. For example, the dispositional readiness to seek out new and potentially threatening stimuli, which has been described as

Actively Open-Minded Thinking (AOT) (Haran et al., 2013), is likely to affect identity-protection in a way that individuals with higher levels of AOT should experience less identity threat/affirmation, resulting in reduced bias. First insights support this view (Bronstein et al., 2019). Similarly, as a counterpart of AOT, which describes openness, the epistemic need for cognitive closure describes individual levels of closed-mindedness (Webster & Kruglanski, 1994) and could affect identity-protection in a way that individuals with higher levels of cognitive closure feel more aversive towards identity incongruent (mis)information.

Moreover, it was recently found that grandiose and vulnerable narcissists engage significantly less in systematic reasoning, possibly increasing the predisposition for identity-protection cognition (Littrell et al., 2020). In addition, it has been shown that vulnerable narcissists react stronger to identity threats (Czarna et al., 2018), possibly increasing identity-protection cognition biases. In contrast, a buffer mitigating identity-protection cognition has been identified in scientific curiosity (Kahan et al., 2017). Individuals with higher levels of science curiosity displayed lower levels of identity-protection cognition.

4.4.4 Alternative explanations

While motivated reasoning and identity-protection cognition have become more and more popular in recent years (Yeo et al., 2015), two alternative approaches can similarly explain the observed favoring of identity-congruent (mis)information and opposition to identity-incongruent (mis)information. First, the observed bias can result from an implicit affect regulation strategy as a response to a psychological threat. In the clinical-empirical model of emotion regulation, Westen and Blagov (2007) suggest “that emotions are evolved response tendencies that reinforce behavioral and mental processes that are pleasurable and select against those that are aversive” (p. 374). This emotion-regulation understanding is not inherently different from an identity-protection framework. In contrast to identity-protection cognition, which originates in cognitive dissonance theory, an emotion-regulation perspective to motivated reasoning draws on empirical results from neuroimaging studies (Westen et al., 2006). In their fMRI study, Westen and colleagues (2006) confronted participants with identity-incongruent information. They observed “activations in the lateral and medial orbital PFC, ACC, insula, and the posterior cingulate and contiguous precuneus and parietal cortex” (p. 1955). The results indicated that motivated reasoning displayed significantly different activation compared to scenarios when participants were asked to reason about neutral stimuli. According to the authors, the activation of the left ventral lateral frontal cortex points to an implicit affect regulation strategy. However, the authors also recognize that their study does not

allow to conclude an exact timeline, sequencing implicit affective regulation before subsequent cognitive responses.

A second alternative to why one can observe biased responses is the incomplete updating of mental models (Swire & Ecker, 2018), an explanation that dates back to early misinformation investigations (Seifert, 2002). While small changes to one's mental model can be integrated (Bailey & Zacks, 2015), this approach suggest that, once misinformation is coherently incorporated in a person's mental model, the person is more likely to dismiss any information which would require a global updating of one's model (Kurby & Zacks, 2012; Johnson & Seifert, 1994). This continued influence effect can be overcome when misinformation is replaced with an alternative explanation to replace the previous causal structure supporting the misinformation. However, such a process is likely to be effortful, requiring cognitive and motivational resources (Lewandowsky et al., 2012).

4.4.5 Motivations and processing style

The early importance of motivational states is reflected in the nomenclature of the observed bias: *motivated* reasoning. In her seminal work, Kunda (1990) opens her argumentation with the statement that “[t]he notion that goals or motives affect reasoning has a long and controversial history in social psychology” (p. 480). Later in her work, she argues that two motivations (accuracy motivation & directional motivation), decide which reasoning strategies individuals employ (see also Chapter 2.3.1). Similarly, Chaiken and colleagues (1996) argue for three different motivational states which drive information processing. While Chaiken et al. (1996) also discuss accuracy motivation, the authors dissect directional motivation into defense motivation and impression formation. The authors aptly describe defense motivation as “the desire to hold attitudes and beliefs that are congruent with existing self-definitional attitudes and beliefs” (p. 557). In contrast, impression motivated individuals are driven by the social situation and express judgments tailored to the audience present or imagined.

In light of this, the mechanisms suggested in identity-protection cognition adhere to the directional defense motivation description. Similar to identity-protection cognition, Chaiken et al. (1996) propose that a defense motivation results from a threat to the self or self-definitional attitudes, values, and beliefs. Consequently, I suggest placing the defense motivation in the model of identity-protection cognition after the elicitation of identity threat.

However, which motivation is evoked by an identity affirmation? More provocatively speaking, can one even speak of *identity-protection* cognition when an identity is affirmed by incoming (mis)information, and can one assume that similar processes are at play compared to

identity-threatening (mis)information? Previous literature on identity-protection is relatively silent in this regard. At the most, (Dunning, 2015) could show that individuals who perceive a threat to their identity react with the need to reaffirm their identity (see also: Sherman & Cohen, 2006).

One of the rare studies investigating possible underlying processing differences between motivated reasoning in the case of incongruent information compared to motivated reasoning in the case of congruent information comes from Jain and Maheswaran (2000). Building on previous results suggesting that congruent information is examined less critically than incongruent information (Ditto & Lopez, 1992) and that individuals engage in more extensive search to justify prior attitudes (Kruglanski, 1980), the authors hypothesized that incongruent information results in more systematic and elaborated processing as compared to congruent information. The results support the author's hypothesis: Participants engaged in more systematic processing when faced with incongruent information.

Taking these findings into account for identity-protection cognition would mean that identity-affirming (mis)information is passed on with less scrutiny. In contrast, identity-threatening (mis)information are followed by a defense motivation and leads to more systematic processing to counter the identity threat, which has also been termed rationalization (Cushman, 2020). However, this suggestion contradicts recent findings, which indicated that more systematic processing reduces identity-protection cognition (Bago et al., 2020; Pennycook & Rand, 2019, 2021) (see also Chapter 2.5.5). Similar to Study 5 of this cumulus, these studies assessed individual predispositions to engage in systematic processing and cognitive sophistication, showing that individuals who are more inclined to process information systematically exhibited less bias.

How can systematic processing increase bias in one case and decrease bias in another? I suggest that one way to explain these contradicting results is to ask under which condition individuals with higher cognitive sophistication skills employ these skills to arrive at an accurate conclusion and under which condition the same individuals employ their skills to rationalize (mis)information to arrive at a pre-defined conclusion. In other words, one needs to take into account the individual predisposition as well as the situational cues that might trigger a defense motivation. A second possible explanation suggests that the individual predisposition to systematically process (mis)information is conceptually different from the individual predisposition to suppress (intuitive) prepotent responses. To carefully dissect how individual predispositions, motivations, and situational differences impact identity-protection cognition raises important questions for future studies.

4.5 Conclusion

The overarching aim of the studies in this thesis is to gain a deeper understanding of the relationship between misinformation on social media and motivated reasoning. This thesis shows that motivated reasoning affects misinformation processing and diffusion on different levels, such as the sharing of misinformation, perceptions of and engagements with fraudulent entities, evaluations of credibility and journalistic practices, and reasoning about (mis)information. The five studies conducted for this thesis demonstrate similar effects of motivated reasoning on misinformation circulation: congruent (mis)information is more readily trusted, accepted, and shared than incongruent (mis)information which is met with less trust, more rejection, and fewer shares. Consequently, the effects of motivated reasoning are twofold: Motivated reasoning makes individuals either more receptive or resistant to misinformation.

The similar effects of motivated reasoning on misinformation circulation are explained by an identity threat/affirmation and threat/affirmation induced emotional reactions. However, it is also found that threat/affirmation induced emotional reactions are likely confounded with emotions elicited by the semantic content of the (mis)information. Moreover, credibility cues of the (mis)information, individuals' cognitive sophistication abilities, and task affordances moderate motivated reasoning. To incorporate these findings, a refined model of motivated reasoning as identity-protection cognition is introduced.

Overall, the results of this thesis contribute to a better understanding of how motivated reasoning affects misinformation on social media and elucidate motivated reasoning's underlying psychological processes.

5. REFERENCES

- Abramowitz, A. (2010). *The disappearing center: Engaged citizens, polarization, and American democracy*. Yale University Press. <https://doi.org/10.12987/9780300162882>
- Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020). Detecting deep-fake videos from appearance and behavior. *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. <https://doi.org/10.1109/WIFS49906.2020.9360904>
- 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>
- Allem, J. P., Ferrara, E., Uppu, S. P., Cruz, T. B., & Unger, J. B. (2017). E-Cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR Public Health and Surveillance*, *3*(4), e98. <https://doi.org/10.2196/publichealth.8641>
- An, J., Quercia, D., & Crowcroft, J. (2013). Fragmented social media: A look into selective exposure to political news. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, *1*, 51–52. <https://doi.org/10.1145/2487788.2487807>
- Anthony, A., & Moulding, R. (2019). Breaking the news: Belief in fake news and conspiracist beliefs. *Australian Journal of Psychology*, *71*(2), 154–162. <https://doi.org/10.1111/ajpy.12233>
- Aronson, E. (1968). Dissonance theory: Progress and problems. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 5–27). Rand McNaily.
- Aronson, E., Chase, T., Helmreich, R., & Ruhnke, R. (1974). A two-factor theory of dissonance reduction: The effect of feeling stupid or feeling awful on opinion change. *International Journal for Research and Communication*, *3*, 59–74.
- Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. *Social Media and Society*, *6*(3), 1–14. <https://doi.org/10.1177/2056305120939264>
- Asutay, E., Genevsky, A., Barrett, L. F., Hamilton, J. P., Slovic, P., & Västfjäll, D. (2019). Affective calculus: The construction of affect through information integration over time. *Emotion*, *21*(1), 159–174. <https://doi.org/10.1037/emo0000681>
- Asutay, E., Genevsky, A., Hamilton, J. P., & Västfjäll, D. (2020). Affective context and its uncertainty drive momentary affective experience. *Emotion*, 1–32. <https://doi.org/10.1037/emo0000912>
- Bacon, F. (n.d.). *The new organon and related writings*. Liberal Arts Press.
- 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>
- Bailey, H. R., & Zacks, J. M. (2015). Situation model updating in young and older adults: Global versus incremental mechanisms. *Psychology and Aging*, *30*(2), 232–244. <https://doi.org/10.1177/0165025419874125>

- 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>
- Bamberg, M. (2011). Who am I? Narration and its contribution to self and identity. *Theory & Psychology*, 21(1), 3–24. <https://doi.org/10.1177/0959354309355852>
- Bankert, A., Huddy, L., & Rosema, M. (2017). Measuring partisanship as a social identity in multi-party systems. *Political Behavior*, 39(1), 109–132. <https://doi.org/10.1007/s11109-016-9349-5>
- Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1), 38–54. <https://doi.org/10.1177/0894439317734157>
- Bauer, P. C., & Clemm von Hohenberg, B. (2020). Believing and sharing information by fake sources: An experiment. *Political Communication*, 00(00), 1–25. <https://doi.org/10.1080/10584609.2020.1840462>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.
- Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21(11), 1–17. <https://doi.org/10.5210/fm.v21i11.7090>
- Bisgaard, M. (2015). Bias will find a way: Economic perceptions, attributions of blame, and partisan-motivated reasoning during crisis. *The Journal of Politics*, 77(3), 849–860. <https://doi.org/10.1086/681591>
- Blanchette, I., & Caparos, S. (2013). When emotions improve reasoning: The possible roles of relevance and utility. *Thinking and Reasoning*, 19(3–4), 399–413. <https://doi.org/10.1080/13546783.2013.791642>
- Blanchette, I., Caparos, S., & Bastien, T. (2018). Emotion and reasoning. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge international handbook series. The Routledge international handbook of thinking and reasoning* (pp. 57–70). Routledge/ Taylor & Francis Group.
- Blanchette, I., & Richards, A. (2003). The effect of emotion on conditional reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 151–156.
- Bliuc, A. M., McGarty, C., Reynolds, K., & Muntele, D. (2007). Opinion-based group membership as a predictor of commitment to political action. *European Journal of Social Psychology*, 37(1), 19–32. <https://doi.org/10.1002/ejsp.334>
- Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, 36(2), 235–262. <https://doi.org/10.1007/s11109-013-9238-0>
- Boomgaarden, H. G. (2007). *Framing the others: News and ethnic prejudice*. University of Amsterdam, The Netherlands.
- Boyer, M. M. (2021). Aroused argumentation: How the news exacerbates motivated reasoning. *The International Journal of Press/Politics*, 1–24.

<https://doi.org/10.1177/19401612211010577>

- Brader, T., Valentino, N. A., & Suhay, E. (2008). What triggers public opposition to immigration? Anxiety, group cues, and immigration threat. *American Journal of Political Science*, 52(4), 959–978. <https://doi.org/10.1111/j.1540-5907.2008.00353.x>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35–58). Blackwell.
- Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, 71, 499–515. <https://doi.org/10.1146/annurev-psych-010419-050807>
- Broniatowski, D. A., Jamison, A. M., Qi, S. H., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10), 1378–1384. <https://doi.org/10.2105/AJPH.2018.304567>
- 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>
- Burkhardt, J. M. (2017). History of fake news. In *Combating fake news in the digital age* (Vol. 53, Issue 8, pp. 5–8). Library Technology Reports.
- Camina, E., Bernacer, J., & Guell, F. (2020). Belief operationalization for empirical research in psychological sciences. *Foundations of Science*, 26, 325–340. <https://doi.org/10.1007/s10699-020-09722-9>
- Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin*, 135(2), 183–204. <https://doi.org/10.1037/a0013965>
- Chaiken, S. (1987). The heuristic model of persuasion. In M. Zanna, J. Olson, & C. Herman (Eds.), *Social influence: The Ontario symposium* (5th ed., pp. 3–39). Lawrence Erlbaum.
- Chaiken, S., Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In P. M. G. & J. A. Bargh (Ed.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 553–578). The Guilford Press.
- Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. Uleman & J. Bargh (Eds.), *Unintended thought* (pp. 212–252). Guilford Press.
- Cheng, C., Luo, Y., & Yu, C. (2020). Dynamic mechanism of social bots interfering with public opinion in network. *Physica A: Statistical Mechanics and Its Applications*, 551, 124163. <https://doi.org/10.1016/j.physa.2020.124163>
- Cheung-Blunden, V., & Ju, J. (2016). Anxiety as a barrier to information processing in the event of a cyberattack. *Political Psychology*, 37(3), 387–400. <https://doi.org/10.1111/pops.12264>
- Clayton, K., Davis, J., Hinckley, K., & Horiuchi, Y. (2019). Partisan motivated reasoning and

misinformation in the media: Is news from ideologically uncongenial sources more suspicious? *Japanese Journal of Political Science*, 20(3), 129–142.
<https://doi.org/10.1017/S1468109919000082>

- Clemm von Hohenberg, B. (2019). *An ocean of possible truth: Biased processing of news on social media*. <https://ssrn.com/abstract=3281038>
- Clore, G. L., Gasper, K., & Garvin, E. (2001). Affect as information. In J. P. Forgas (Ed.), *Handbook of Affect and Social Cognition* (pp. 121–144). Lawrence Erlbaum Associates.
<https://doi.org/10.4135/9781412956253.n8>
- Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26(9), 1151–1164. <https://doi.org/10.1177/01461672002611011>
- Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80, 56–71.
<https://doi.org/10.1016/j.dss.2015.09.003>
- Cunningham, W. A., Dunfield, K. A., & Stillman, P. E. (2013). Emotional states from affective dynamics. *Emotion Review*, 5(4), 344–355.
<https://doi.org/10.1177/1754073913489749>
- Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, 43(28), 1–59.
<https://doi.org/10.1017/S0140525X19001730>
- Czarna, A. Z., Zajenkowski, M., & Dufner, M. (2018). How does it feel to be a narcissist? Narcissism and emotions. In A. Hermann, A. Brunell, & J. Foster (Eds.), *Handbook of trait narcissism* (pp. 255–263). Springer, Cham. https://doi.org/10.1007/978-3-319-92171-6_27
- Dassonneville, R., Hooghe, M., & Vanhoutte, B. (2014). Partisan dealignment in Germany: A rejoinder to Russell Dalton. *German Politics*, 23(1–2), 145–155.
<https://doi.org/10.1080/09644008.2014.916694>
- De Hoog, N. (2012). Processing of social identity threats: A defense motivation perspective. *Social Psychology*, 44, 361–372. <https://doi.org/10.1027/1864-9335/a000133>
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14, 238–257. <https://doi.org/10.1177/1088868309352251>
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3), 554–559.
<https://doi.org/10.1073/pnas.1517441113>
- DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. American Psychological Association. <https://doi.org/10.1037/11503-000>
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4), 568–584. <https://doi.org/10.1037/0022-3514.63.4.568>
- Doosje, B., Spears, R., & Ellemers, N. (2002). Social identity as both cause and effect: The development of group identification in response to anticipated and actual changes in the intergroup status hierarchy. *British Journal of Social Psychology*, 41(1), 57–76.

<https://doi.org/10.1348/014466602165054>

- Druckman, J. N., Leeper, T. J., & Slothuus, R. (2016). Motivated responses to political communications: Framing, party cues, and science information. In H. Lavine & C. S. Taber (Eds.), *The Feeling, Thinking Citizens* (1st ed.). Routledge.
<https://doi.org/10.4324/9781351215947>
- Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119.
<https://doi.org/10.1038/s41558-018-0360-1>
- Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J. W., de Gelder, B., Deonna, J., de Sousa, R., ... Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*.
<https://doi.org/10.1038/s41562-021-01130-8>
- Dunning, D. (2015). Motivational theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 108–131). Guilford Press.
- Ecker, U. K. H., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory and Cognition*, *42*(2), 292–304.
<https://doi.org/10.3758/s13421-013-0358-x>
- Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, *33*, 372–376.
<https://doi.org/10.1016/j.chb.2013.08.013>
- Edwards, C., Edwards, A., Spence, P. R., & Westerman, D. (2015). Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies*, *67*(2), 1–12. <https://doi.org/10.1080/10510974.2015.1121899>
- Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. <https://doi.org/10.1080/23808985.2019.1602782>
- Eilders, C. (2006). News factors and news decisions. Theoretical and methodological advances in Germany. *Communications*, *31*(1), 5–24.
<https://doi.org/10.1515/COMMUN.2006.002>
- Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353.
<https://doi.org/10.1037/1528-3542.7.2.336>
- Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 U.S. presidential election. *Berkman Klein Center for Internet & Society Research Paper*.
<http://nrs.harvard.edu/urn-3:HUL.InstRepos:33759251%0AThis>
- Farrell, J. (2019). The growth of climate change misinformation in US philanthropy: Evidence from natural language processing. *Environmental Research Letters*, *14*(3).
<https://doi.org/10.1088/1748-9326/aaf939>
- Feldman-Barrett, L. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social, Cognitive and Affective Neuroscience*, *12*(1),

1–23. <https://doi.org/10.1093/scan/nsw154>

- Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8). <https://doi.org/10.5210/fm.v22i8.8005>
- Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104. <https://doi.org/10.1145/2818717>
- Figenschou, T. U., & Ihlebæk, K. A. (2019). Challenging journalistic authority. Media criticism in far-right alternative media. *Journalism Studies*, 20(9), 1221–1237. <https://doi.org/10.1080/1461670X.2018.1500868>
- Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism and Mass Communication Quarterly*, 77(3), 515–540. <https://doi.org/10.1177/107769900007700304>
- Flatscher, M., & Seitz, S. (2018). Latour, Foucault und das Postfaktische: Zur Rolle und Funktion von Kritik im Zeitalter der “Wahrheitskrise.” *Le Foucauldien*, 4(1), 1–30. <https://doi.org/10.16995/lefou.46>
- Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Advances in Political Psychology*, 38(S1), 127–150. <https://doi.org/10.1111/pops.12394>
- Forelle, M. C., Howard, P. N., Monroy-Hernandez, A., & Savage, S. (2015). Political bots and the manipulation of public opinion in venezuela. *SSRN Electronic Journal*, 1–8. <https://doi.org/10.2139/ssrn.2635800>
- Foucault, M. (2008). Power/knowledge. In S. Seidman & J. C. Alexander (Eds.), *The new social theory reader*. Routledge.
- Freiling, I., Krause, N. M., Scheufele, D. A., & Brossard, D. (2021). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media and Society*. <https://doi.org/10.1177/14614448211011451>
- Frijda, N. H. (1993). Moods, emotion episodes, and emotions. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 381–403). Guilford Press.
- Gable, P. A., Poole, B. D., & Harmon-Jones, E. (2015). Anger perceptually and conceptually narrows cognitive scope. *Journal of Personality and Social Psychology*, 109(1), 163–174. <https://doi.org/10.1037/a0039226>
- Gage-Bouchard, E. A., LaValley, S., Warunek, M., Beaupin, L. K., & Mollica, M. (2018). Is cancer information exchanged on social media scientifically accurate? *Journal of Cancer Education*, 33(6), 1328–1332. <https://doi.org/10.1007/s13187-017-1254-z>
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of Peace Research*, 2(1), 64–91. <https://doi.org/10.1177/002234336500200104>
- Gilad, B., Kaish, S., & Loeb, P. D. (1987). Cognitive dissonance and utility maximization. A general framework. *Journal of Economic Behavior and Organization*, 8(1), 61–73. [https://doi.org/10.1016/0167-2681\(87\)90021-7](https://doi.org/10.1016/0167-2681(87)90021-7)
- Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can’t believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221–233. <https://doi.org/10.1037/0022-3514.65.2.221>

- Guo, C., Cao, J., Zhang, X., Shu, K., & Yu, M. (2019). *Exploiting emotions for fake news detection on social media*. <https://deepai.org/publication/exploiting-emotions-for-fake-news-detection-on-social-media>
- Hameleers, M. (2020). 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*, 0(0), 1–17. <https://doi.org/10.1080/1369118X.2020.1764603>
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, 8(3), 188–201.
- Harmon-Jones, E. (2010). A cognitive dissonance theory perspective on the role of emotion in the maintenance and change of beliefs and attitudes. *Emotions and Beliefs*, 185–211. <https://doi.org/10.1017/cbo9780511659904.008>
- Hartmann, T., & Tanis, M. (2013). Examining the hostile media effect as an intergroup phenomenon: The role of ingroup identification and status. *Journal of Communication*, 63(3), 535–555. <https://doi.org/10.1111/jcom.12031>
- Holt, K., Figenschou, T. U., & Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, 7(7), 860–869. <https://doi.org/10.1080/21670811.2019.1625715>
- Horne, B. D., & Adali, S. (2017). *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news*. 759–766. <http://arxiv.org/abs/1703.09398>
- Hornsey, M. J. (2008). Social identity theory and self-categorization theory: A historical review. *Social and Personality Psychology Compass*, 2(1), 204–222. <https://doi.org/10.1111/j.1751-9004.2007.00066.x>
- Howard, P. N., & Kollanyi, B. (2016). *Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum*. <https://doi.org/10.2139/ssrn.2798311>
- Howe, L. C., & Krosnick, J. A. (2017). Attitude strength. *Annual Review of Psychology*, 68, 327–351. <https://doi.org/10.1146/annurev-psych-122414-033600>
- Huddy, L. (2001). From social to political identity: A critical examination of social identity theory. *Political Psychology*, 22(1), 127–156. <https://doi.org/10.1111/0162-895X.00230>
- Huddy, L., Feldman, S., Taber, C. S., & Lahav, G. (2005). Threat, anxiety, and support of antiterrorism policies. *American Journal of Political Science*, 49(3), 593–608. <https://doi.org/10.1111/j.1540-5907.2005.00144.x>
- Jain, S. P., & Maheswaran, D. (2000). Motivated reasoning: A depth-of-processing perspective. *Journal of Consumer Research*, 26(4), 358–371. <https://doi.org/10.1086/209568>
- Jerit, J., & Zhao, Y. (2020). Political misinformation. *Annual Review of Political Science*, 23, 77–94. <https://doi.org/10.1146/annurev-polisci-050718-032814>
- Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1420–1436. <https://doi.org/10.1037/0278-7393.20.6.1420>

- Jones, P. (1999). Beliefs and identities. In J. Horton & S. Mendus (Eds.), *Tolerance, identity, and difference* (pp. 65–86). Palgrave Macmillan.
https://doi.org/10.1057/9780333983379_4
- Jones, S. C., & Regan, D. T. (1974). Ability evaluation through social comparison. *Journal of Experimental Social Psychology*, 133–146. [https://doi.org/10.1016/0022-1031\(74\)90062-6](https://doi.org/10.1016/0022-1031(74)90062-6)
- Juneja, P., & Mitra, T. (2021). Auditing e-commerce platforms for algorithmically curated vaccine misinformation. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–27. <https://doi.org/10.1145/3411764.3445250>
- Jung, N., Wranke, C., Hamburger, K., & Knauff, M. (2014). How emotions affect logical reasoning: evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Frontiers in Psychology*, 5(570), 1–12.
<https://doi.org/10.3389/fpsyg.2014.00570>
- Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgement and Decision Making*, 8(4), 407–424. <https://doi.org/10.2139/ssrn.2182588>
- Kahan, D. M. (2017). *Misconceptions, misinformation, and the logic of identity-protective cognition*. <https://ssrn.com/abstract=2973067>
- Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, 4(3), 465–505. <https://doi.org/10.1111/j.1740-1461.2007.00097.x>
- 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>
- Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science curiosity and political information processing. *Advances in Political Psychology*, 38(S1), 179–199. <https://doi.org/10.1111/pops.12396>
- Kahan, D. M., Peters, E., Dawson, E., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1), 54–86.
<https://doi.org/10.1017/bpp.2016.2>
- Kahne, J., & Bowyer, B. (2017). Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal*, 54(1), 3–34. <https://doi.org/10.3102/0002831216679817>
- Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Nature Publishing Group*, 6(39589), 1–11. <https://doi.org/10.1038/srep39589>
- Karataş, A., & Şahin, S. (2017). A review on social bot detection techniques and research directions. *Proceedings of the International Conference on Security and Cryptology Conference Turkey*, 156–161.
- Keijzer, M. A., & Mäs, M. (2021). The strength of weak bots. *Online Social Networks and Media*, 21. <https://doi.org/10.1016/j.osnem.2020.100106>
- Keller, F. B., Schoch, D., Stier, S., & Yang, J. H. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2), 256–

280. <https://doi.org/10.1080/10584609.2019.1661888>

- Kim, H. K., Ahn, J., Atkinson, L., & Kahlor, L. A. (2020). Effects of COVID-19 Misinformation on Information Seeking, Avoidance, and Processing: A Multicountry Comparative Study. *Science Communication*, 42(5), 568–915. <https://doi.org/10.1177/1075547020959670>
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228. <https://doi.org/10.1037/0033-295X.94.2.211>
- Kozlowski, A. C., & Murphy, J. P. (2021). Issue alignment and partisanship in the American public: Revisiting the ‘partisans without constraint’ thesis. *Social Science Research*, 94, 102498. <https://doi.org/10.1016/j.ssresearch.2020.102498>
- Kruglanski, A. W. (1980). Lay epistemology process and contents. *Psychological Review*, 87(1), 70–87. <https://doi.org/10.1037/0033-295X.87.1.70>
- Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impression primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, 19, 448–468. [https://doi.org/10.1016/0022-1031\(83\)90022-7](https://doi.org/10.1016/0022-1031(83)90022-7)
- Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., Rich, R. F., Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3), 790–816. <https://doi.org/10.1111/0022-3816.00033>
- Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media and Society*, 1(2). <https://doi.org/10.1177/2056305115610141>
- Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53(4), 636–647. <https://doi.org/10.1037/0022-3514.53.4.636>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, 17, 22–26. <https://doi.org/10.1016/j.copsyc.2017.06.004>
- Kurby, C. A., & Zacks, J. M. (2012). Starting from scratch and building brick by brick in comprehension. *Memory and Cognition*, 40(5), 812–826. <https://doi.org/10.3758/s13421-011-0179-8>
- Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, 30(2), 225–248. <https://doi.org/10.1086/421123>
- Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.
- Lazarus, R. S. (1995). Vexing research problems inherent in cognitive-mediational theories of emotion—and some solutions. *Psychological Inquiry*, 6(3), 183–196. https://doi.org/10.1207/s15327965pli0603_1
- Lazer, D. (2020). Studying human attention on the Internet. *Proceedings of the National Academy of Sciences of the United States of America*, 117(1), 21–22. <https://doi.org/10.1073/pnas.1919348117>

- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Lecheler, S., Bos, L., & Vliegenthart, R. (2015). The mediating role of emotions: News framing effects on opinions about immigration. *Journalism and Mass Communication Quarterly*, 92(4), 812–838. <https://doi.org/10.1177/1077699015596338>
- Leeper, T. J., & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Advances in Political Psychology*, 35(1), 129–156. <https://doi.org/10.1111/pops.12164>
- Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion*, 14(4), 473–493. <https://doi.org/10.1080/026999300402763>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/1529100612451018>
- Lind, T., Erlandsson, A., Västfjäll, D., & Tinghög, G. (2018). Motivated reasoning when assessing the effects of refugee intake. *Behavioural Public Policy*, 1–24. <https://doi.org/10.1017/bpp.2018.41>
- Littrell, S., Fugelsang, J., & Risko, E. F. (2020). Overconfidently underthinking: Narcissism negatively predicts cognitive reflection. *Thinking and Reasoning*, 26(3), 352–380. <https://doi.org/10.1080/13546783.2019.1633404>
- Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. Cognition, Choice, and the Bounds of Rationality. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of Reason* (pp. 183–213). Cambridge University Press. <https://doi.org/10.1017/cbo9780511805813.009>
- Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.
- Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109. <https://doi.org/10.1001/archfami.4.5.463>
- Mader, M., & Schoen, H. (2019). The European refugee crisis, party competition, and voters' responses in Germany. *West European Politics*, 42(1), 67–90. <https://doi.org/10.1080/01402382.2018.1490484>
- Maitner, A. T., Mackie, D. M., Claypool, H. M., & Crisp, R. J. (2010). Identity salience moderates processing of group-relevant information. *Journal of Experimental Social Psychology*, 46(2), 441–444. <https://doi.org/10.1016/j.jesp.2009.11.010>
- Marcus, G. E., MacKuen, M. B., Wolak, J., & Keele, L. (2006). The measure and mismeasure of emotion. In D. P. Redlawsk (Ed.), *Feeling politics: Emotion in political information processing* (pp. 31–46). Palgrave Macmillan. <https://doi.org/10.1057/9781403983114>
- Marcus, G. E., Neuman, W. R., & MacKuen, M. B. (2000). *Affective intelligence and political judgement*. University of Chicago Press.
- Marques, J. M., Yzerbyt, V. Y., & Leyens, J.-P. (1988). The black sheep effect: Judgmental

- extremity towards ingroup members in inter-and intra-group situations. *European Journal of Social Psychology*, 18(3), 1–16. <https://doi.org/10.1002/ejsp.2420180308>
- Mason, L. (2015). “I disrespectfully agree”: The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1), 128–145. <https://doi.org/10.1111/ajps.12089>
- Mason, Lillian. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.
- McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, 27(5), 486–493. <https://doi.org/10.1016/j.appdev.2006.06.003>
- McKasy, M. (2020). A discrete emotion with discrete effects: effects of anger on depth of information processing. *Cognitive Processing*, 21(4), 555–573. <https://doi.org/10.1007/s10339-020-00982-8>
- Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153, 112986. <https://doi.org/10.1016/j.eswa.2019.112986>
- Mercier, H. (2016). Confirmation bias - myside bias. In R. Pohl (Ed.), *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory* (2nd ed., pp. 99–114). Psychology Press. <https://doi.org/10.4324/9781315696935>
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- Moravec, P., Minas, R., & Dennis, A. R. (2018). Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper*, 18(87), 1–36. <https://doi.org/10.2139/ssrn.3269541>
- 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>
- Musgrove, L., & McGarty, C. (2008). Opinion-based group membership as a predictor of collective emotional responses and support for pro- and anti-war action. *Social Psychology*, 39(1), 37–47. <https://doi.org/10.1027/1864-9335.39.1.37>
- Nabi, R. L. (1999). A cognitive-functional model for the effects of discrete negative emotions on information processing, attitude change, and recall. *Communication Theory*, 9(3), 292–320. <https://doi.org/10.1111/j.1468-2885.1999.tb00172.x>
- Nabi, R. L. (2003). Exploring the framing effects of emotion: Do discrete emotions differentially influence information accessibility, information seeking, and policy preference? *Communication Research*, 30(2), 224–247. <https://doi.org/10.1177/0093650202250881>

- Nauroth, P., Gollwitzer, M., Kozuchowski, H., Bender, J., & Rothmund, T. (2017). The effects of social identity threat and social identity affirmation on laypersons' perception of scientists. *Public Understanding of Science, 26*(7), 754–770. <https://doi.org/10.1177/0963662516631289>
- Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters institute digital news report*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_1.pdf
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. <https://doi.org/https://doi.org/10.1037/2F1089-2680.2.2.175>
- Nielsen, R. K., & Graves, L. (2017). News you don't believe: Audience perspectives on fake news. In *Reuters Institute for the Study of Journalism*.
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Östgaard, E. (1965). Factors influencing the flow of news 2 (1): 39–63. *Journal of Peace Research, 2*(1), 39–63. <https://doi.org/10.1177/002234336500200103>
- Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and misinformation: A dangerous combination? *BMJ (Online), 349*(6178), 14–15. <https://doi.org/10.1136/bmj.g6178>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(April). <https://doi.org/10.1038/s41586-021-03344-2>
- 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*(September), 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Peterson, E., & Allamong, M. B. (2021). The influence of unknown media on public opinion: Evidence from local and foreign news sources. In *Preprint* (Vol. 00, Issue 0).
- Peterson, E., & Iyengar, S. (2021). Partisan gaps in political information and information-seeking behavior: Motivated reasoning or cheerleading? *American Journal of Political Science, 65*(1), 133–147. <https://doi.org/10.1111/ajps.12535>
- Postmes, T. (2015). Psychology: Climate change and group dynamics. *Nature Climate Change, 5*(3), 195–196. <https://doi.org/10.1038/nclimate2537>
- Quandt, T. (2018). Dark participation. *Media and Communication, 6*(4), 36–48. <https://doi.org/10.17645/mac.v6i4.1519>
- Quandt, T., Frischlich, L., Boberg, S., & Schatto-Eckrodt, T. (2019). Fake news. In T. P. Vos & F. Hanusch (Eds.), *The International Encyclopedia of Journalism Studies* (pp. 1–6). John Wiley & Sons. <https://doi.org/10.1002/9781118841570.iejs0128>
- Quigley, K. S., Lindquist, K. A., & Barrett, L. F. (2014). Inducing and measuring emotion and affect: Tips, tricks, and secrets. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 220–252). Cambridge University Press. <https://doi.org/10.1017/cbo9780511996481.014>
- Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L.

- (2016). The measles vaccination narrative in twitter: A quantitative analysis. *JMIR Public Health Surveillance*, 2(1), 1–15. <https://doi.org/10.2196/publichealth.5059>
- Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PLoS ONE*, 15(10). <https://doi.org/10.1371/journal.pone.0241045>
- Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. In *Journal of Politics* (Vol. 64, Issue 4, pp. 1021–1044). The University of Chicago Press on behalf of the Southern Political Science Association.
- Redlawsk, D. P. (2006). *Feeling politics. Emotion in political information processing* (D. P. Redlawsk (ed.)). Palgrave Macmillan.
- Ribeiro, M. H., Calais, P. H., Almeida, V. A. F., & Meira, W. (2017). “Everything I disagree with is #FakeNews”: Correlating political polarization and spread of misinformation. In *Proceedings of DATA SCIENCE + JOURNALISM@KDD 2017, Halifax, Canada, August 2017*, 8. https://doi.org/10.475/123_4
- Roberts, S. T. (2018). Digital detritus: “Error” and the logic of opacity in social media content moderation. *First Monday*, 23(3), 1–13. <https://doi.org/10.5210/fm.v23i3.8283>
- Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4), 394–412. <https://doi.org/10.1080/0960085X.2018.1560920>
- Sanchez, C., & Dunning, D. (2021). Cognitive and emotional correlates of belief in political misinformation: Who endorses partisan misbeliefs? *Emotion*. <https://doi.org/10.1037/emo0000948>
- Schäfer, F., Evert, S., & Heinrich, P. (2017). Japan’s 2014 general election: Political bots, right-wing internet activism, and prime minister Shinzō Abe’s hidden nationalist agenda. *Big Data*, 5(4), 294–309. <https://doi.org/10.1089/big.2017.0049>
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. <https://doi.org/10.1177/0539018405058216>
- Schwarz, N., & Clore, G. L. (1983). How do I feel about it? The information function of affective states. In K. Fiedler & J. P. Forgas (Eds.), *Affect, cognition and social behavior: New evidence and integrative attempts* (pp. 44–63). C.J. Hogrefe.
- Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *Psychology of Learning and Motivation*, 41, 265–292. [https://doi.org/10.1016/S0079-7421\(02\)80009-3](https://doi.org/10.1016/S0079-7421(02)80009-3)
- Settles, I. H. (2004). When multiple identities interfere: The role of identity centrality. *Personality and Social Psychology Bulletin*, 30(4), 487–500. <https://doi.org/10.1177/0146167203261885>
- Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, 38, 183–242. [https://doi.org/10.1016/S0065-2601\(06\)38004-5](https://doi.org/10.1016/S0065-2601(06)38004-5)
- Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10(1), 80–83.

<https://doi.org/10.1111/1467-9280.00111>

- Sivek, S. C. (2018). Both facts and feelings: Emotion and news literacy. *Journal of Media Literacy Education, 10*(2), 123–138. <https://doi.org/10.23860/jmle-2018-10-2-7>
- Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177*(3), 1333–1352. <https://doi.org/10.1016/j.ejor.2005.04.006>
- Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and the emotions. *Cognition and Emotion, 7*(3/4), 233–269. <https://doi.org/10.1080/02699939308409189>
- Smith, J. R., & Hogg, M. A. (2008). Social identity and attitudes. In W. Crano & R. Prislin (Eds.), *Attitudes and attitude Change* (pp. 337–360). Psychology Press. <https://doi.org/10.4324/9780203838068>
- Smith, R., & Lane, R. D. (2015). The neural basis of one's own conscious and unconscious emotional states. *Neuroscience and Biobehavioral Reviews, 57*, 1–29. <https://doi.org/10.1016/j.neubiorev.2015.08.003>
- Spears, R. (2021). Social influence and group identity. *Annual Review of Psychology, 72*(15), 15.1-15.24. <https://doi.org/10.1146/annurev-psych-070620-111818>
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*(3), 225–247. <https://doi.org/10.1080/13546780600780796>
- Stanovich, K. E., & West, R. F. (2008a). On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking and Reasoning, 14*(2), 129–167. <https://doi.org/10.1080/13546780701679764>
- Stanovich, K. E., & West, R. F. (2008b). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672–695. <https://doi.org/10.1037/0022-3514.94.4.672>
- Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 261–302). Academic Press.
- Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology and Behavior, 106*(1), 5–15. <https://doi.org/10.1016/j.physbeh.2011.06.004>
- Stocking, G., & Sumida, N. (2018). Social media bots draw public's attention and concern. In *Pew Research Center*. <https://www.pewresearch.org/journalism/2018/10/15/social-media-bots-draw-publics-attention-and-concern/>
- Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated reasoning and public opinion. *Journal of Health Politics, Policy and Law, 36*(6), 935–944. <https://doi.org/10.1215/03616878-1460524>
- Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research, 23*(1). <https://doi.org/10.2196/17187>
- Suhay, E., & Erisen, C. (2018). The role of anger in the biased assimilation of political information. *Political Psychology, 39*(4), 793–810. <https://doi.org/10.1111/pops.12463>

- Sundar, S. Shyam, Xu, Q., & Oeldorf-Hirsch, A. (2009). Authority vs. peer: How interface cues influence users. *Conference on Human Factors in Computing Systems - Proceedings*, 4231–4236. <https://doi.org/10.1145/1520340.1520645>
- Sundar, Shyam S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility*, 73–100. <https://doi.org/10.1162/dmal.9780262562324.073>
- Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17(2), 202–227. <https://doi.org/10.1111/j.1467-9760.2008.00325.x>
- Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In B. Southwell, E. A. Thorson, & L. Sheble. (Eds.), *Misinformation and Mass Audiences* (pp. 195–211). University of Texas Press.
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- Taber, C. S., & Lodge, M. (2016). The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Advances in Political Psychology*, 37(1), 61–85. <https://doi.org/10.1111/pops.12321>
- Tajfel, H., & Turner, J. (1979). An intergrative theory of intergroup conflict. In M. J. Hatch & M. Schultz (Eds.), *Organizational identity: A reader* (pp. 33–47). Oxford University Press.
- Tajfel, H., & Turner, J. (1986). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Key Readings in Social Psychology. Political Psychology* (pp. 276–293). Psychology Press.
- Tambuscio, M., Ruffo, G., Flammini, A., & Menczer, F. (2015). Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*, 977–982. <https://doi.org/10.1145/2740908.2742572>
- Tandoc, E. C., & Johnson, E. (2016). Most students get breaking news first from Twitter. *Newspaper Research Journal*, 37(2), 153–166. <https://doi.org/10.1177/0739532916648961>
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining “fake news”: A typology of scholarly definitions. *Digital Journalism*, 6(3), 1–17. <https://doi.org/10.1080/21670811.2017.1360143>
- Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, 34, 81–87. <https://doi.org/10.1016/j.cobeha.2020.01.003>
- Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, 146(8), 1143–1149. <https://doi.org/10.1037/xge0000298>
- Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, 48, 227–236. <https://doi.org/10.2307/3033683>

- Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, 57(388–398).
<https://doi.org/10.1037/0022-3514.57.3.388>
- Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, 52, 700–709.
<https://doi.org/10.1037/0022-3514.52.4.700>
- Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, 128(2), 237–251.
<https://doi.org/10.1016/j.cognition.2012.09.012>
- Trevors, G. J. (2019). Psychological tribes and processes: Understanding why and how misinformation persists. In P. Kendeou, D. Robinson, & M. T. McCrudden (Eds.), *Misinformation and Fake News in Education*. Information Age Publishing.
- Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism and Mass Communication Quarterly*, 94(1), 38–60. <https://doi.org/10.1177/1077699016654682>
- Tripodi, F. (2018). Searching for alternative facts: Analyzing scriptural inference in conservative news practices. *Data & Society Research Institute*, 1–64.
https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Searching-for-Alternative-Facts_FINAL-5.pdf
- Tsang, S. J. (2020). Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism & Mass Communication Quarterly*, 1–18. <https://doi.org/10.1177/1077699020952129>
- Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovicj, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. In *Political polarization, and political disinformation: a review of the scientific literature*. <https://doi.org/10.2139/ssrn.3144139>
- Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, 20(5), 520–535.
<https://doi.org/10.1111/jcc4.12127>
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Blackwell.
- Turner, J. C., Hogg, M. A., Turner, P. J., & Smith, P. M. (1984). Failure and defeat as determinants of group cohesiveness. *British Journal of Social Psychology*, 23, 97–111.
<https://doi.org/10.1111/j.2044-8309.1984.tb00619.x>
- Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and collective: Cognition and social context. *Personality and Social Psychology Bulletin*, 20(5), 454–463. <https://doi.org/10.1177/0146167294205002>
- Tussyadiah, I. P., Kausar, D. R., & Soesilo, P. K. M. (2018). The effect of engagement in online social network on susceptibility to influence. *Journal of Hospitality and Tourism Research*, 42(2), 201–223. <https://doi.org/10.1177/1096348015584441>
- Unsworth, K. L., & Fielding, K. S. (2014). It's political: How the salience of one's political identity changes climate change beliefs and policy support. *Global Environmental*

Change, 27(1), 131–137. <https://doi.org/10.1016/j.gloenvcha.2014.05.002>

- Valenzuela, S., Halpern, D., Katz, J. E., & Miranda, J. P. (2019). The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, 7(6), 802–823. <https://doi.org/10.1080/21670811.2019.1623701>
- 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>
- 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>
- Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4), 473–482. [https://doi.org/10.1016/0005-7967\(68\)90028-4](https://doi.org/10.1016/0005-7967(68)90028-4)
- Verheyen, C., & Göritz, A. S. (2009). Plain texts as an online mood-induction procedure. *Social Psychology*, 40(1), 6–15. <https://doi.org/10.1027/1864-9335.40.1.6>
- Verkuyten, M. (2018). The benefits of studying immigration for social psychology. *European Journal of Social Psychology*, 48(3), 225–239. <https://doi.org/10.1002/ejsp.2354>
- Viau-Quesnel, C., Savary, M., & Blanchette, I. (2019). Reasoning and concurrent timing: A study of the mechanisms underlying the effect of emotion on reasoning. *Cognition and Emotion*, 33(5), 1020–1030. <https://doi.org/10.1080/02699931.2018.1535427>
- Visser, P. S., Krosnick, J. A., & Norris, C. J. (2016). Attitude importance and attitude-relevant knowledge. In J. A. Krosnick, A. I-Chant, & T. H. Stark (Eds.), *Political Psychology* (pp. 205–247). Psychology Press.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 1151, 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Vraga, E. K., Bode, L., & Tully, M. (2020). Creating news literacy messages to enhance expert corrections of misinformation on Twitter. *Communication Research*, 1–23. <https://doi.org/10.1177/0093650219898094>
- Wang, P., Angarita, R., & Renna, I. (2018). Is this the era of misinformation yet? Combining social bots and fake news to deceive the masses. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 1557–1561. <https://doi.org/10.1145/3184558.3191610>
- Wang, R., He, Y., Xu, J., & Zhang, H. (2020). Fake news or bad news? Toward an emotion-driven cognitive dissonance model of misinformation diffusion. *Asian Journal of Communication*, 30(5), 317–342. <https://doi.org/10.1080/01292986.2020.1811737>
- Wardle, C., & Derakhshan, H. (2018). Thinking about ‘information disorder’: Formats of misinformation, disinformation, and mal-information. In C. Ireton & J. Posetti (Eds.), *Journalism, “fake news” & disinformation-UNESCO* (pp. 43–54). https://en.unesco.org/sites/default/files/f._jfdn_handbook_module_2.pdf
- Washburn, A. N., & Skitka, L. J. (2017). Science denial across the political divide: Liberals and conservatives are similarly motivated to deny attitude-inconsistent science. *Social Psychological and Personality Science*, 9(8), 972–980. <https://doi.org/10.1177/1948550617731500>

- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6), 1049–1062. <https://doi.org/10.1037/0022-3514.67.6.1049>
- 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>
- Westen, D., & Blagov, P. S. (2007). A clinical-empirical model of emotion regulation: From defense and motivated reasoning to emotional constraint satisfaction. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 373–392). The Guilford Press.
- Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 U.S. Presidential election. *Journal of Cognitive Neuroscience*, 18(11), 1947–1958. <https://doi.org/10.1162/jocn.2006.18.11.1947>
- Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, 26(4), 557–580. [https://doi.org/10.1002/\(SICI\)1099-0992\(199607\)26:4<557::AID-EJSP769>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1099-0992(199607)26:4<557::AID-EJSP769>3.0.CO;2-4)
- Winter, S., Brückner, C., & Krämer, N. C. (2015). They came, they liked, they commented: Social influence on facebook news channels. *Cyberpsychology, Behavior, and Social Networking*, 18(8), 431–436. <https://doi.org/10.1089/cyber.2015.0005>
- Winter, S., Metzger, M. J., & Flanagin, A. J. (2016). Selective use of news cues: A multiple-motive perspective on information selection in social media environments. *Journal of Communication*, 66(4), 669–693. <https://doi.org/10.1111/jcom.12241>
- Yan, H. Y., Yang, K. C., Menczer, F., & Shanahan, J. (2020). Asymmetrical perceptions of partisan political bots. *New Media and Society*, 00(0), 1–22. <https://doi.org/10.1177/1461444820942744>
- Yeo, S. K., Cacciatore, M. A., & Scheufele, D. A. (2015). News selectivity and beyond: Motivated reasoning in a changing media environment. In O. Jandura, T. Petersen, & A. Schielicke (Eds.), *Publizistik und gesellschaftliche Verantwortung* (pp. 83–104). Springer Fachmedien. <https://doi.org/10.1007/978-3-658-04704-7>

ARTICLE 1

Shareworthiness and Motivated Reasoning in Hyper-Partisan News Sharing Behavior on Twitter

Magdalena Wischnewski^a, Axel Bruns^b and Tobias Keller^c

^aSocial Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany

^bDigital Media Research Centre, Queensland University of Technology, Brisbane, QLD, Australia

^cgfs.bern, Bern, Switzerland

This is an original manuscript of an article published by Taylor & Francis in Digital Journalism Vol. 9(5), available online:

<https://www.tandfonline.com/doi/full/10.1080/21670811.2021.1903960>

Author Note

Magdalena Wischnewski: <https://orcid.org/0000-0001-6377-0940>

Axel Bruns: <https://orcid.org/0000-0002-3943-133X>

Tobias Keller: <https://orcid.org/0000-0001-5263-4812>

Correspondence concerning this article should be addressed to Magdalena Wischnewski: magdalena.wischnewski@uni-due.de

Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter

Magdalena Wischnewski*

Social Psychology: Media and Communication, University of Duisburg-Essen, Germany,
magdalena.wischnewski@uni-due.de, ORCID ID: <https://orcid.org/0000-0001-6377-0940>, Twitter ID:
@wischnewski_m

Axel Bruns

Digital Media Research Centre, Queensland University of Technology, Australia, a.bruns@qut.edu.au,
ORCID ID: <https://orcid.org/0000-0002-3943-133X>, Twitter ID: @snurb_dot_info

Tobias Keller

gfs.bern, Switzerland tobias.keller@gfsbern.ch, ORCID ID: <https://orcid.org/0000-0001-5263-4812>,
Twitter ID: @Tobias_Keller

* Corresponding Author

Word count text: 8266

Acknowledgment and Funding Information

Thanks to Ina Rentemeister for her support in the coding process, as well as QUT Digital Media Research Centre staff Brenda Moon, Ehsan Dehghan, Timothy Graham, and Daniel Angus for their support of this work.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, by the Research Training Group "User-Centred Social Media" at the University of Duisburg-Essen, and by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant No. 823866. At QUT, this research was also supported by the Australian Research Council Discovery project *DP200101317 Evaluating the Challenge of 'Fake News' and Other Malinformation* and by the Swiss National Science Foundation grant No. P2ZHP1_184082.

Shareworthiness and motivated reasoning in hyper-partisan news sharing behavior on Twitter

Abstract

While news sharing by ordinary social media users has received growing attention, hyper-partisan news sharing, which has been closely associated with misinformation circulation, has received less attention. In this study, we investigate hyper-partisan news sharing from two perspectives: (1) the features that make hyper-partisan news share-worthy, as well as (2) the user motivations that drive the sharing process. We scrutinize one week's content from *Infowars.com* as it was shared on Twitter. Through both manual coding of news content and semi-automated clustering of Twitter account descriptions, we find that human interest and conflict in news stories drive the sharing process from a content perspective. Concerning the user perspective, we find partial support for a sharing hypothesis based on motivated reasoning, which indicates that users are more likely to share hyper-partisan news stories if these align with their own political opinions.

Keywords: news sharing, shareworthiness, motivated reasoning, Infowars, hyper-partisan news, popularity cues

Introduction

Digital media, and social media as a distinct subset within this category, have become an ever more crucial source of news for audiences around the world over the past decade. Indeed, the 2017 Digital News Report, published by the Reuters institute for the Study of Journalism at the University of Oxford (Newman et al., 2017), points to a substantial generational shift in news engagement practices: digital media are now the main source of news for half or more of the respondents across the 36 nations it studied who were aged 44 or under, while TV and other legacy media remained central only for older news audiences. Further, social media play an ever-increasing role as a distinct space for encountering news within this overall digital media environment; in 2017, for example, fully one third of respondents aged 18-24 received their news mainly from social media (Newman et al., 2017, p.11).

This shift away from established legacy news media and towards online news engagement results in several important changes to news consumption patterns. First, it reduces news brand loyalty: online, users are able to access and engage with a substantially broader range of news brands from around the world, and many do – leading Funt et al. (2016, n.p.) to ask, “do brands even matter anymore?” Second, the use of social media for discovering the news has also shifted news engagement dynamics to a more passive mode for many users: they are using social media as “social awareness streams” that “unbundle a news story into its individual components” (Hermida, 2012, p. 665). In this environment, users encounter news serendipitously, in the form of news items shared by their friends and connections on social media platforms such as Facebook and Twitter – and this serendipity even increases the diversity of the news sources they encounter: “those who are incidentally exposed to news on social media use more different sources of online news than non-users” (Fletcher & Nielsen, 2018, p.2459).

If the news articles that are shared by individuals in a social media user's network, as a result of those individuals' gatewatching of mainstream and niche news outlets (Bruns, 2018), are thus critically important in shaping the overall news diet of that user, then this places even greater importance on the news-sharing decisions made by these individuals. Such decisions can be understood from two broader perspectives: First, online news-sharing as a process of evaluating the importance and relevance of the news articles themselves – assessing a story's *shareworthiness* (e.g. Kilgo et al., 2020; Trilling et al., 2017). In shareworthiness, users who encounter the story as published by the news outlet consider whether the story is of relevance to their own social media followers and should therefore be shared with them on Facebook, Twitter, or other platforms. In such work, multiple success factors have been identified by comparing more and less successfully shared articles on social media.

In contrast to shareworthiness as a perspective that seeks to identify crucial characteristics of the *news* driving news sharing, the second perspective seeks to understand crucial motivations of *users* driving the news sharing process (e.g. Syn & Oh, 2015). Previous studies have already found various individual motivations for sharing news, such as impression management or information seeking (e.g. Lee & Ma, 2012). Rather than assuming specific motives that drive the sharing process, we apply motivated reasoning theory to explain news sharing. Motivated reasoning suggests that people sometimes process attitude-relevant information in a biased manner in a way that favors attitude-congruent information over attitude-incongruent information (Kunda, 1990). In the same manner, we propose that users favor and prefer to share attitude-congruent news over attitude-incongruent news. We argue that motivated reasoning as an overarching theoretical framework is appropriate since previous studies have found that individuals regularly show a bias for in-groups, especially in political communication (Druckman et al., 2016; Lodge & Taber, 2000). However, both perspectives are ultimately two sides of one coin and are, as such, interdependent and inextricably interlinked.

In this article, we combine these two perspectives on news sharing by applying both perspectives to hyper-partisan news media. We select hyper-partisan news media as studies that investigate from both perspectives (shareworthiness and sharing motivations) why material from alternative, niche, and fringe news media outlets is shared remain scarce. By contrast, sharing of news online, and the shareworthiness considerations that determine it, have received an increasing amount of scholarly attention in recent years. Such studies have largely focused on shareworthiness factors and sharing motivations for general, mainstream news content, however. This is all the more problematic in light of growing concerns about the impact of partisan and hyper-partisan news sources on political discourse, especially in deeply polarised societies such as those of the United States and United Kingdom. In their study of mainstream and social media coverage of the 2016 US presidential election campaign, for instance, Faris et al. (2017, p. 11) observe “a significant reshaping of the conservative media landscape over the past several years”, and even suggest that as a result of these shifts “the center of attention and influence for conservative media is on the far right. The center-right is of minor importance and is the least represented portion of the media spectrum” (2017, p. 10), with *Breitbart* and similar sites from the extreme right playing a particularly prominent role. Following Benkler et al. (2017), we describe such outlets as hyper-partisan: they are

sites that revive what Richard Hofstadter called “the paranoid style in American politics,”

combining decontextualized truths, repeated falsehoods, and leaps of logic to create a fundamentally misleading view of the world. (2017, n.p.)

As Benkler et al. note, this category of sites “appears to have not only successfully set the agenda for the conservative media sphere, but also strongly influenced the broader media agenda” (ibid.).

The term ‘hyper-partisan’ also enables us to move beyond simplistic evaluations of the truthfulness of the news articles that these sites publish. At issue here is not whether the news published in such sites is wholly ‘fake’ or ‘real’, nor whether ‘fake news’ is disseminated more quickly on social media platforms than ‘real news’ (Vousoughi et al., 2018); in reality, as the studies cited here have shown, the stories published by hyper-partisan news outlets often contain at least a kernel of truth, but twist their material well out of context. ‘Fake news’ is thus an inadequate, insufficiently defined term; what is more important is that hyper-partisan news content is implicated in the dissemination of mis-, dis-, and malinformation, in the definition provided by Wardle and Derakhshan (2017, p. 20), and has been blamed for disrupting elections (Shin et al., 2018) and eroding societal trust (Turcotte et al., 2015). As a result, the need to understand *what* hyper-partisan content is shared on social media (i.e. shareworthiness) and *why* some social media users choose to share such content (i.e. individual motives) becomes all the more pressing.

To complete the picture of hyper-partisan news sharing, we examine (1) news characteristics that increase or decrease the shareworthiness of hyper-partisan content, and (2) user characteristics that provide the motivations for this sharing process. We do this in two steps: first, we ask which characteristics of hyper-partisan news make articles more or less likely to be shared on social media. Second, employing motivated reasoning theories from cognitive and social psychology, we hypothesise that articles which support users’ prior beliefs and attitudes are more likely to be shared. Hence, we pursue the following two research questions:

1. What makes hyper-partisan news more or less shareworthy on social media?
2. How can motivated reasoning explain hyper-partisan news sharing?

To answer these research questions, we investigate one week’s content from *Infowars.com*, a well-known hyper-partisan news outlet, as it was shared on Twitter⁹. Through manual coding of news content, we compare in a first step content that was published by *Infowars* and shared on Twitter with content that was published but not shared – hence, determining the shareworthiness factors. We relied on GDELT (Leetaru & Schrod, 2013), a public database that monitors global news coverage in real time, to identify all *Infowars* news articles published during the timeframe covered by our Twitter dataset (the final week of September 2019). In determining the shareworthiness factors, we specifically differentiate

⁹ A thorough justification of why we chose *Infowars* and Twitter can be found in the methodology section ‘Sampling and time frame’.

between *cross-platform* and *in-platform on-sharing*. While cross-platform sharing describes how often news were shared on Twitter in general, in-platform on-sharing assesses how often news were shared on *within* a platform (in our case Twitter). This allows for a refined interpretation of shareworthiness that distinguishes between primary and secondary sharing processes.

In addition, we collected Twitter profile descriptions from all of the accounts that shared *Infowars* URLs on Twitter, in order to use them as a proxy for understanding possible sharing motivations. Based on the profile descriptions, we grouped accounts into opinion clusters to identify if they were more likely to share opinion-congruent news stories – for instance, to determine whether self-described conservative accounts also predominantly share news articles representing such political perspectives.

Theoretical Background

Our study builds on two major theoretical foundations: shareworthiness and motivated reasoning. Respectively, these address the questions of which news stories are shared, which existing research seeks to explain by examining the inherent features of the content being shared; and of why such news stories are shared, which past studies address by exploring the likely motivations of users sharing the news. In this section, we introduce these theoretical frameworks – newsworthiness and shareworthiness on the one hand, and motivated reasoning on the other – in turn, and develop the hypotheses that guide our own research.

Which news stories are shared? From newsworthiness to shareworthiness

Today's notion of shareworthiness extends from the concept of *newsworthiness* – determining which events are most likely to become news. Developing news value theory, Östgaard (1965), as well as Galtung and Ruge (1965), introduced specific factors that impact on newsworthiness, like unexpectedness, references to individuals (human interest), or negativity. These factors have proven to influence not only journalists' selection of news stories, but also the audience's selection and preferences (Eilders, 2006). Galtung and Ruge (1965) proposed that these individual news value factors serve as a “good score of the [otherwise] elusive concept of ‘newsworthiness’” (p.71).

Translating news value theory and newsworthiness to the practice of news sharing, Trilling and colleagues (2017) arrived at the concept of *shareworthiness*, which provides a central point of focus for our article. They proposed seven factors to explain why content was shared on Facebook and Twitter. Their empirical results indicated that all seven factors – geographical distance, cultural distance, higher negativity, higher positivity, the presence of conflict and human interest (only for Facebook shares), and exclusiveness (only for Twitter shares) – predicted news sharing. This research approach has been extended by others: for example, Valenzuela et al. (2017) applied news value theory to news frames and news sharing. Through in-depth interviews, they investigated how different news frames affected sharing likelihood. Their results indicate that, contrary to Trilling et al. (2017), the presence of conflict

frames decreased news sharing, and that human-interest frames had no effect on sharing news.

Although academic interest in the shareworthiness of mainstream media content has grown in recent years, comparatively little is known about why users share *hyper-partisan* news, as a specific subset of news content that is distinct from ordinary news. Since hyper-partisan news outlets display features different from mainstream news – at the level of the identity of the news producer, the content produced, the organizational structure of the outlet, and its embedding in the wider news ecosystem (Holt, Figenschou, & Frischlich, 2019) – and position themselves outside of the traditional media system (e.g. by describing themselves as ‘alternative’ media; Figenschou & Ihlebæk, 2019), it is necessary to differentiate hyper-partisan news-sharing from traditional news-sharing: users may have different motivations for sharing such hyper-partisan content, and respond to different attributes in the content. One of the few studies that scrutinized hyper-partisan news-sharing came from Xu, Sang, and Kim (2020): using manual coding as well as computational techniques, the authors investigated how hyper-partisan news was liked and shared on Facebook. To do so, they examined articles on three levels: source, style, and content. Concerning the source, they found that, while the inclusion of the author byline in articles generally increased shares, more information about the author’s biography and more hyperlinks decreased shares. Concerning style, results indicated that more emotional content was only more likely to be liked but not shared. Also, more formal and logical language, as well as multimedia content, affected neither likes nor shares. Lastly, Xu and colleagues (2020) found that specific moral frames, such as an authority frame, increased shares.

Based on these findings, we have developed a list of previously identified factors associated with shareworthiness in both mainstream and hyper-partisan media, which we discuss in the following. This list does not encompass all previously found shareworthiness factors, but instead focuses especially on repeatedly reported news values and shareworthiness factors. To be clear: our central focus in this article is on shareworthiness (i.e., the factors that make social media users share news articles), not on newsworthiness (the factors that make journalists cover news events in the first place). There are considerable overlaps between both sets of factors, and this is unsurprising: newsworthiness factors are based on what journalists expect their audiences to be interested in, while shareworthiness factors result from the direct observation of such audience interests (as expressed in news sharing practices via social media). The two sets of factors are not entirely identical, however: journalists may cover stories that they believe audiences need to know about, whether those audiences are interested or not; audiences may share stories with limited news value if they are sufficiently amusing, surprising, or outrageous.

Proximity

In their early works, Galtung and Ruge (1965) described *proximity* as one of the news values that determine if an event is reported as news. According to the authors, the closer (culturally or geographically) an event is to the country where it is reported, the more likely it is to become news. The news value of proximity was supported by further research (Bednarek & Caple, 2017), including studies of news images (Ahva & Pantti, 2014) and among different cultures (Masterton, 2005). Moreover, the concept was successfully translated to shareworthiness (Trilling et al., 2017; Valenzuela et al., 2017). Although it has been argued

that the importance of a traditional news value like proximity would decrease in a globalized world, in “today’s global social media networks geographic proximity still matters” (Bruns, 2018, p. 136). One psychological explanation for the influence of proximity is that proximate content is more relatable for readers, which has also been found to influence commenting behavior (Weber, 2014). Proximate content might also be more likely to convey information that affects the individual personally. This is supported by the findings of Ahva and Pantti (2014), who investigated the role of proximity in today’s digital news environment. They found proximity to be used as a central means to engage audiences.

Because our investigation examines the sharing of *Infowars* content, we define proximity from the perspective of the United States. We hypothesize that:

H1: Issues that are proximate (culturally and/or geographically) are more likely to be shared. In turn, issues that are culturally/geographically distant are less likely to be shared.

Conflict

Likewise, in their list of news values, Galtung and Ruge (1965) proposed that the presence of *controversy* or *conflict* increases the likelihood of an event becoming news. A conflict is characterized by the portrayal of at least two disagreeing sides and is deemed to be of more interest to the audience than consensus (Semetko & Valkenburg, 2000). This facilitating effect of conflict has also been found for news sharing (Kim, 2015; Trilling et al., 2017; Valenzuela et al., 2017). Underlying psychological concepts that drive conflict as a news and sharing factor are sensationalism and negativity bias. Concerning the former, Ng and Zhao (2020) hypothesized that the two evolutionary needs, environmental surveillance and social involvement, cause sensational news such as conflicts to be shared more. Concerning negativity bias, psychological studies have found that negative events elicit stronger cognitive, emotional, and behavioral responses than neutral or positive events (Baumeister et al., 2001).

We therefore hypothesize that:

H2: Issues that portray a conflict or a controversy are more likely to be shared. Issues that do not contain a conflict are less likely to be shared.

Human interest

The factor *human interest*, which gives news stories a human face or an emotional angle (Neuman et al., 1992; Semetko & Valkenburg, 2000), was introduced to the discussion of news values subsequent to Galtung and Ruge’s earlier work. This factor relates to a softer style of personalized storytelling, in contrast to ‘hard news’, and is more entertainment-centered (Jebril et al., 2013). Concerning shareworthiness, Trilling and colleagues (2017) found human interest to be less important than conflict or proximity (especially for Twitter shares). However, others have found opposite effects (García-Perdomo et al., 2018). An investigation of the so-called Ice Bucket Challenge and its sharing has shown that most journalists utilized human interest in their reporting, for example (Kilgo et al., 2020). From a psychological perspective, a human-interest angle might trigger emotional arousal which, in turn, increases psychological engagement with the news story. We hypothesize that:

H3: Issues that portray an angle of human interest are more likely to be shared. Issues that do not contain an angle of human interest are less likely to be shared.

Morality

Similar to human interest, the news factor *morality* was introduced to the discussion on newsworthiness at a later stage (Neuman et al., 1992; Semetko & Valkenburg, 2000), and, although journalists use it less often due to their commitment to objective news reporting (Wasike, 2013), audiences often use morality frames to understand the news (de Vreese, 2012). In the case of shareworthiness, morality is defined as putting “the event or issue in the context of values, moral prescriptions, normative messages, and religious or cultural tenets” (Valenzuela et al., 2017, p. 809). A psychological reason for the importance of morality as a shareworthiness factor is that moral emotions, like outrage and disgust, are elicited by moralizing content. This is supported by a recent investigation which showed that moral-emotional language accelerated sharing on social media through social contagion (Brady et al., 2017). Similar to negativity bias, moral emotions have also been shown to mobilize people if the content resonates with or contradicts the individual’s value predispositions (Rubenking, 2019). We thus hypothesize that:

H4: Issues that portray moralizing content are more likely to be shared. Issues that do not portray moralizing content are less likely to be shared.

Visual content

Lastly, we want to draw attention to visual content. As *Infowars* not only publishes written news reports but also livestreams videos and shows which are “repackage[d] to fit web and social media formats” (Van den Bulck & Hyzen, 2020, p. 51), we can expect to find URLs that link to visual content. As previous studies have found, images and videos are far more likely to be shared on Twitter than in any other medium (Goel et al., 2016), although results by Xu and colleagues (2020) did not find support for this claim. This can be explained by the lower cognitive affordances and effort required in information acquisition. We therefore hypothesize that:

H5: Issues that contain mostly visual content, like videos, are more likely to be shared. Issues that are not visual reports are less likely to be shared.

Why news is shared – Motivations for news sharing

Understanding news sharing from a shareworthiness perspective neglects, however, the user’s motivations for news sharing. Though extrinsic content characteristics, like the shareworthiness factors discussed above, have certainly proven to explain news sharing patterns in general, it has also been shown that individuals have different intrinsic sharing motives that shape their specific news sharing choices. Such intrinsic motives may be of particular importance in the sharing of hyper-partisan news content, which is inherently designed to appeal to users with strong pre-existing ideological loyalties. We expect ideologically determined intrinsic user motivations to play a considerably stronger role in the

decision to share hyper-partisan news content than they would do in sharing more balanced, mainstream news reporting, where extrinsic shareworthiness factors relating to the content and substance of a story are more important. We follow this line of inquiry by connecting the logics of news sharing with the theory of motivated reasoning.

Motivated reasoning generally proposes that people sometimes process attitude-relevant information in a biased manner in a way that favors attitude-congruent information over attitude-incongruent information (Kunda, 1990). One theoretical explanation for this biased processing relates to differing motivational states. For example, Chaiken, Giner-Sorolla, and Chen (1996) suggested that individuals are not always driven by accuracy goals when processing information. Instead, individual cognition is sometimes driven by belief preservation and self-concept defense (defense motivation). Hence, the authors predict that, once defense motivation is triggered by attitude-incongruent information, individuals favor information that reinforces prior attitudes. To reduce the threatening potential of attitude-incongruent information, attitude-contradicting content can be ignored (defensive inattention) or over-critically evaluated (defensive counterarguing). Defensive counterarguing, however, is more likely to occur when content seems easy to refute (Lowin, 1969).

In line with motivated reasoning and, especially, defense motivation, we suggest that it is more likely that individuals share attitude-congruent than attitude-incongruent news. While attitude-congruent news should signal no threat to the individual, attitude-incongruent news could threaten the individual self-concept, leading to defensive inattention. Empirical findings from social and political psychology support this view. De Hoog (2013), for example, found that when people were confronted with self-threatening information, defense motivation was induced, resulting in biased information processing.

In line with motivated reasoning, ideology-based motivations that drive the spread of misinformation have been identified before (An et al., 2013; Marwick, 2018). An and colleagues (2013) found, for example, that users tend to share articles that were congruent with their prior attitudes and beliefs. This is supported by recent findings which show that attitude-congruent information is more likely to be shared, independent of source credibility (Clemm von Hohenberg, 2019). In fact, on Twitter “attitudinal congruence mattered more for known sources” (p. 33). Nevertheless, we acknowledge that there are rationales that could lead users to share attitude-incongruent news: for example, to discredit or correct such news. We hypothesize that:

H6: Individual attitudes drive the processes of news sharing in a way that those news stories that align with an individual's attitudes are more likely to be shared, whereas stories that do not align with or that are neutral to an individual's attitude are less likely to be shared.

Methodology

Sampling and timeframe

For the present study, we collected tweets from Twitter that shared URLs from *Infowars* during the last week of September 2019. We selected Twitter as a platform that is a

particularly popular social medium for news dissemination and consumption (Tandoc & Johnson, 2016), and a space that enables us to comprehensively observe the public sharing of articles from specific news outlets. We selected *Infowars*, founded by Alex Jones, as it has previously been classified as hyper-partisan (Xu et al., 2020), and has been linked to a greater network of misinformation spread (Shao et al., 2018). Further, we chose *Infowars* because amongst comparable hyper-partisan sites, it is one of the most prominent sources: the Reuters Institute *Digital News Report 2019* (Newman et al., 2019, p. 24) shows that fully 33% of its panel of US-based respondents are aware of the site. This places it second only to *Breitbart* (44%), and ahead of other key hyper-partisan outlets such as *The Blaze* (31%) and *Daily Caller* (27%). In addition, and especially also in comparison with these other prominent hyper-partisan outlets, *Infowars* serves as a particularly useful case study because its content distribution on Twitter is entirely driven by third-party accounts, rather than resulting in part from the promotional efforts of official Twitter accounts affiliated with the site: the institutional and personal accounts operated by *Infowars* and its founder Alex Jones were banned from Twitter in 2018. This means that *Infowars* content is now distributed on Twitter overwhelmingly as a result of the individual sharing decisions made by members of its audience, while content from *Breitbart* and similar sites is instead disseminated in the first place in tweets by these sites' institutional accounts, and by the retweeting of those tweets by other users. This positions *Infowars* uniquely well as an object of study for our purposes: we can be confident that the sharing decisions we observe in our data represent individual users' original decisions on whether to share any given *Infowars* article, and are not influenced by the activities of the site's institutional account or the account of its leader Alex Jones.

As the timeframe of our data collection, we selected the last week of September 2019. This covered the emergence of impeachment claims against then-US President Donald Trump, possibly resulting in higher rates of engagement on alt-right outlets such as *Infowars*. We decided to restrict the timeframe to one week of sharing. This kept the number of collected tweets small enough for comprehensive manual coding but large enough to detect trends beyond a single event, and to use computational methods in our analysis.

Data collection to answer RQ1 and H1-H5

To answer RQ1 and determine which available *Infowars* articles were shared on Twitter and which were not, we utilized the GDEL project (Global Database of Events, Language, and Tone)¹⁰. GDEL (Leetaru & Schrod, 2013) is an open data project that monitors global news coverage in real-time in over 100 languages, from print media to broadcasting and web formats, applying natural language processing, data mining, and deep learning algorithms to extract about 300 categories of events, themes, and emotions. GDEL has been used to predict social unrest (Qiao et al., 2017), study political conflict (Yonamine, 2013), or examine visual media coverage (Kwak & An, 2014). As GDEL includes hyper-partisan news outlets like *Infowars* in its dataset, it provides a useful source of information on the full range of articles published on the site.

By querying GDEL, we identified 169 *Infowars* articles that were first captured by GDEL during the period of 23 to 29 September 2019. Further, to identify which of the

¹⁰ <https://www.gdelproject.org/>

available *Infowars* news articles were actually shared on Twitter, we used the public Twitter API to capture all tweets that contained an *Infowars.com* URL (even if shortened by *t.co* or another URL shortener), and were shared in the period of 24 to 30 September 2019. We deliberately offset the Twitter dataset collection timeframe from that of the GDELT dataset by 24 hours in order to allow sufficient time for articles from *Infowars* to be shared on Twitter, and we subsequently filtered the Twitter dataset to retain only those tweets that shared an article URL in the GDELT dataset within 24 hours of GDELT's first capture of that URL.

This means that for each URL, our analysis focuses on the sharing of *Infowars* articles on Twitter within the first 24 hours of its publication (or more correctly, its capture by GDELT, which will usually have occurred shortly after publication). We introduce this limitation in order to be able to focus on the *immediate* sharing of news articles and exclude any *residual* sharing that may occur well after the initial publication of an article (including out-of-context and spam-related sharing); such filtering is justified by the fact that the vast majority of news sharing for any source tends to occur within the first hours after an article's initial publication (Bruns & Keller, 2020), and that the factors and motivations for delayed sharing are likely to diverge considerably from the shareworthiness factors and motivated reasoning involved in immediate sharing that our study investigates. With these filters applied, our final Twitter dataset consisted of 5,280 original tweets from 1,064 unique accounts, sharing 168 distinct *Infowars.com* article URLs.

Strategy of analysis to answer RQ1: Manual annotation and regression analysis

In a first step, we manually coded each of the 168 article URLs in the GDELT dataset to identify the hypothesized shareworthiness factors (see H1-H5) which constituted one category in the coding system. The codebook can be found in the supplementary material, S1 (<https://osf.io/uc6sm>). Categories were dummy-coded and not mutually exclusive. Two independent coders were trained on 10% of the dataset, achieving satisfactory intercoder reliability (Krippendorff's alpha = between 0.72 and 0.8) after three rounds of coding.

In the next step, we determined which of those articles were shared on Twitter, finding that all but one of the articles in the GDELT dataset had been shared on Twitter within 24 hours of their publication. The least retweeted articles received 4 retweets, whereas the most retweeted article received 5,427 retweets, at a median retweet count of 30 retweets per article. In the last step, we then assessed the shareworthiness of the 168 *Infowars* articles. That is, we used the shareworthiness factors, such as conflict or morality, to explain how often an article was shared on Twitter. We used two dependent variables: for each of the 168 articles we counted (a) how often they were shared on Twitter (tweet count) and calculated (b) a retweet factor that showed the amplification of initial sharing by subsequent retweeting, by dividing the total count of retweets by the count of original shares (excluding retweets).

Previous studies have included only a tweet count for articles (e.g. Trilling et al., 2017) to assess shareworthiness. The retweet factor adds to this a measure that favors those articles which were retweeted more than they were shared in original tweets. In other words, while the tweet count provides a measure of shareworthiness that considers sharing from the original publication *into* the social media platform (i.e. cross-platform sharing), the retweet factor assesses further (in-platform) on-sharing *within* the social media space (in our case Twitter). Hence, the retweet factor allows for a refined interpretation of shareworthiness that distinguishes between primary and secondary sharing processes.

In order to test which of these factors made an article more or less shareworthy, we ran two negative binomial regression models with the tweet count and retweet factor of each individual article as dependent variables and all shareworthiness factors as predictors, controlling for article length. We chose negative binomial regressions because the standard deviation was higher than the mean for both dependent variables. We ran two negative binomial regression models with the shareworthiness factors as predictors and the tweet count and retweet factor as the respective dependent measures, as well as the control variable article length (word count).

Data collection to answer RQ2 and H6

In RQ2 we hypothesize that accounts are more likely to share content which is congruent to the account's opinion or interests. Hence, we needed to gain deeper insights into the individual accounts which shared *Infowars* URLs. As the Twitter API allows us to download profile descriptions, we leveraged the 1,064 Twitter accounts' profile descriptions as a proxy for how these accounts identify themselves (using those Twitter accounts whose tweets were collected to answer RQ1 and H1-H6 – see previous section). Although we do not know if these descriptions are accurate or misleading, they still determine how other Twitter users perceive these accounts, and thus serve as a useful indicator of their public persona. Accounts that did not have a description (2%) were omitted from the analysis.

Strategy of analysis to answer RQ2: Automated annotation, clustering process, and logistic regression analysis

We used a semi-automated approach to cluster the collected profiles based on their descriptions (see Spierings et al., 2018; Keller 2020). We first created keyword lists to group accounts by their descriptions automatically, resulting in seven lists: (1) *Trump*, containing pro-Donald Trump keywords such as #MAGA, #KAG, or "Trump"; (2) *Patriot*, containing keywords such as #AmericaFirst, "PatriotsUnite", or "Nationalists"; (3) *Infowars*, containing keywords such as "Infowarrior", @RealAlexJones, or "InfoArmy"; (4) *Christian*, containing keywords such as "Believer", "Christ" or "Bible"; (5) *Military*, containing keywords such as "Veteran", "Served", or "Army"; (6) *Pro-Gun*, containing keywords such as "NRA" or "2A"; and (7) *Conspiracy*, containing keywords such as "WWG1WGA", "QANON, or "Conspiracy" (for a complete list of keywords, see supplementary material S2: <https://osf.io/uc6sm>). The automated analysis ended after several runs which neither improved the number of identified accounts, nor the correct classification of accounts (validated manually). Of the 1,043 accounts which had a profile description, 475 did not fall into at least one of the seven lists (45%).

To test validity more thoroughly, we took a random sample of 50 accounts, including accounts from each list, and conducted a manual analysis with the same criteria as for the automated analysis. We received good results in terms of accuracy (>0.8), precision (>0.7), recall (0.8), and F1-score (>0.8) for each category. Hence, the clustering resulted in seven clusters which were converted into dummy-coded categories to describe each profile. Categories were not mutually exclusive, allowing profiles to fall into more than one category.

To answer RQ2, we also needed to classify shared news articles into the same cluster categories as used for the profiles. Hence, two independent coders were trained on 10% of the dataset. The codebook for this can be found in the supplementary material S3 (<https://osf.io/uc6sm>). After two rounds of categorization, intercoder reliability was

satisfactory for the overall categories (Krippendorff's alpha = .62 - .72), and the remaining data were coded. Similar to the profile categories, article categories were not mutually exclusive, and articles could fall into more than one category.

Finally, we conducted six logistic regressions to test whether accounts with specific interests were more likely to share news on particular topics than others, as suggested by motivated reasoning theory (H6). For all regression analyses, we entered the article category as a dependent variable and the respective profile cluster as an independent variable, while controlling for all other profile clusters.

Results

RQ 1: Which hyper-partisan news are successfully shared on Twitter?

As both data sets were almost of the same size (GDELT = 169, Twitter = 168), we expected to find only one article that needed to be omitted. Our expectations were met: of the 169 individual URLs collected by GDELT, we found that 168 were shared on Twitter within 24 hours. One of these URLs did not link to an *Infowars* article at the point of coding and had to be omitted from the analysis, leaving 167 articles. The coding process for shareworthiness factors found that most of these articles contained culturally or geographically proximate content, from a US perspective (82%). Roughly every other article contained a conflict (56%) or human-interest content (51%), while moralizing content (29%), as well as visual content such as pictures or videos (36%), were present in roughly one third of all articles. The average length of articles was 465 words.

[INSERT TABLES 1 AND 2]

Results for the tweet count variable, meaning how often an URL was tweeted or retweeted on Twitter (Tables 1 and 2), showed that three shareworthiness factors significantly influenced whether an article was shared or not (proximity, conflict, and human interest). Concerning proximity, H1, we hypothesized that proximate content would increase shareworthiness. In our data, however, we found the opposite effect. Content that thematized issues closer to the USA was less likely to be shared. By contrast, as hypothesized in H2 and H3, the factors conflict and human interest increased the likelihood of being shared. Moreover, we found no support for the hypotheses that moralizing (H4) or visual content (H5) had a significant influence on shareworthiness.

Interestingly, when examining the results for the retweet factor, we found that only the shareworthiness factor human interest reached significance, indicating that if an article had a human angle, it was more likely to receive substantial secondary amplification through retweeting. This general lack of correlation may indicate that the in-platform on-sharing of *Infowars* content through retweets is driven far more strongly by factors relating to the platform (Twitter) than the source (*Infowars*) – for instance, by the identity and the follower base of the Twitter account initially sharing the URL, or the choice of hashtags and other markers used in the original tweet.

While these results indicate which extrinsic content factors in *Infowars* articles increased their shareworthiness, we were also interested in how intrinsic account identities

influenced this sharing process. In the next section, we report results for the clustering analysis.

RQ2: Can motivated sharing practices explain news sharing?

We categorized all account profiles and shared articles using the seven clusters described in the methodology section. We found that the biggest cluster of accounts consisted of Trump supporters ($n = 240$), followed by Christian profiles ($n = 161$), patriotic profiles ($n = 145$), Pro-Gun profiles ($n = 75$), military profiles ($n = 71$), conspiracy profiles ($n = 59$), and *Infowars* supporters ($n = 25$). Because profiles could fall into more than one category, we wanted to know if specific profile categories were associated with each other. To do so, we calculated correlations of profiles using Cramer's V for dichotomous variables. Results are presented in Table 3. According to Cohen's (1988) interpretation of Cramer's V , we found that almost all profile clusters had a medium to high correlation with the Trump cluster. This suggested that, while different accounts showed different specific interests, they connected through their overall support for Donald Trump. One exception was the cluster "*Infowars*", representing self-declared *Infowars*-followers, which correlated substantially only with the cluster "Conspiracy" (self-declared conspiracy theorists such as QAnon followers).

[INSERT TABLE 3]

Likewise, of the 167 articles, most fell into the Trump cluster ($n = 70$), followed by conspiracy articles ($n = 38$), *Infowars* articles ($n = 35$), articles with patriotic themes ($n = 18$), Christianity ($n = 9$), the military ($n = 6$), or Pro-Gun articles ($n = 4$). Similar to the profile clusters, articles could be categorized in more than one category. To find possible co-occurring patterns, we calculated Cramer's V for the article clusters. Results are reported in Table 4. Compared to the profile clusters, associations between these article categories were less pronounced. The strongest co-occurrence was found for articles on Christianity and Pro-Gun articles ($V = 0.32$), followed by *Infowars* and conspiracy articles ($V = 0.31$).

[INSERT TABLE 4]

As suggested by the theory of motivated reasoning, we used six logistic regressions to test whether accounts with specific interests were more likely to share news on particular topics than others (H6). Results indicate partial support for H6 (see Table 5). For articles discussing (a) Trump, (b) *Infowars*, and (c) Christianity, we found that the corresponding profile clusters were, as hypothesized, more likely to share these articles. However, for articles discussing *Infowars* or Christianity, we found a different profile cluster to be more likely to share these articles (the military cluster for *Infowars* and the conspiracy cluster for Christianity). For Pro-Gun articles or conspiracy theories, we found that seemingly unrelated profile clusters were more likely to predict sharing. Pro-Gun articles were more likely to be shared by Christian profiles, and conspiracy articles more likely to be shared by Pro-Gun profiles. We can explain the association between Christian profiles and the sharing of Pro-Gun articles by content co-occurrence (see Table 4): we found that Pro-Gun articles often also contained content related to Christianity ($V = 0.32$).

Further, we found no association between any of the profile clusters with articles that discussed the military. Most surprisingly, we found the opposite effect of what we hypothesized for articles discussing patriotic content. These were, in fact, shared significantly less by accounts from the patriot cluster—the only significant negative effect we found. One possible explanation for this finding may be that accounts describing themselves as patriots use the term in such a generic way that it may lose its specificity in relation to news articles on *Infowars* that are *per se* conservative. In other words, a self-described patriot may not behave very differently to other clusters because *Infowars* news articles speak to all of these clusters.

[INSERT TABLE 5]

Discussion and conclusion

In this study, we have compared *Infowars.com* article URLs, as collected by the GDELT project, with a dataset of *Infowars.com* URLs shared on Twitter and collected by us. Overall, we found that almost all *Infowars* articles were shared on Twitter within 24 hours. This is in line with previous findings by Trilling and colleagues (2017), who found that only a marginal proportion of news articles was not shared (on either Twitter or Facebook).

Through manual coding of the articles and negative binomial regression models, we found that three shareworthiness factors (proximity, conflict, and human interest) significantly predicted sharing, if sharing is operationalized as original tweets and retweets. However, contrary to what we expected in H1, proximity decreased the likelihood of sharing. We also found that most articles contained content proximate to the USA (roughly 82%). We suggest that *Infowars* articles that did not cover US-issues became more salient simply because most content does cover the USA. In turn, increased salience could have resulted in higher shares of non-USA content. We connect the other two factors, conflict and human interest, with results found for RQ2, where we investigated users' motivations of sharing. We found that by far the biggest group of accounts endorsed Donald Trump. In addition, we see a connection to the selected timeframe of our data collection – one week after the first calls of impeachment against Donald Trump.

Moreover, moralizing content, as well as visual content, had no significant influence on sharing likelihood. The results for moralizing content seem to contradict previous findings (Valenzuela et al., 2017; Xu et al., 2020). However, Valenzuela and colleagues (2017) as well as Xu and colleagues (2020) both investigated moralizing *frames* rather than *content*. Differences could stem from this conceptualization; additionally, the hyper-partisan nature of *Infowars* content as compared to the mainstream news observed by other studies may also mean that moralizing aspects in the content are less unusual, and therefore less significant. We found that every third article included moralizing content. Concerning visual content, our findings are in line with Xu et al. (2020), who found that multimedia content played an insignificant role in news sharing (but studied Facebook rather than Twitter).

Interestingly, when we used our retweet factor (the ratio between retweets and original shares) to investigate the shareworthiness attributes affecting the on-sharing of URLs

within the platform, through retweeting, only one shareworthiness factor, human interest, remained significant. This also indicates that findings are dependent on how researchers define and operationalize shareworthiness, and specifically that a distinction between *cross-platform sharing* (in our case, from the *Infowars* site to Twitter) and *in-platform on-sharing* (through the retweeting of tweets containing *Infowars* URLs) is critical in understanding the factors that influence shareworthiness. On Twitter, certainly, but most likely also on many other social media platforms, in-platform on-sharing processes may be affected more strongly by platform features and affordances (the status and follower base of the accounts posting URLs to the platform; the injection of posts into popular hashtags and communities; or the amplification of content by trending topic and newsfeed algorithms) than by the content of the stories themselves. This perspective is also in line with the observation that many social media users engaging in the in-platform on-sharing of content may not themselves click on and read the original source story: in other words, their decision to on-share will be influenced not by the shareworthiness attributes embedded in the article itself, which they may never encounter, but only by those attributes that can be gleaned from the tweet sharing its URL.

Moreover, we asked not only which content features made hyper-partisan news more shareworthy, but also whether specific accounts were more likely to share particular news. According to motivated reasoning theory, we expected that users were more likely to share news that was congruent with their personal views (H6). Our results showed that, as hypothesized, some news stories were more likely to be shared by accounts that aligned with the themes addressed in the news story, supporting the motivated sharing hypothesis. This was especially pronounced for accounts that endorsed Donald Trump, but also for those with strong affinities for *Infowars* and Christianity. However, the picture was not always as clear. For some article themes (e.g. for articles discussing the military), we did not find any strong association with any one profile cluster, whereas for others we found associations with clusters on divergent topics (e.g. articles discussing weapons, which were more likely to be shared by accounts that endorsed Christianity). We suggest that this can be explained by the medium to strong correlation (Cramer's $V = 0.32$) between weapon and Christianity themes in articles, which meant that these two themes were often interwoven within the same article.

Our analysis in this article clearly demonstrates the benefits of combining a content-focused shareworthiness perspective with a user-focused motivated reasoning perspective. We encourage scholars to continue and extend this dual approach and to thus incorporate both perspectives into future research about mainstream as well as hyper-partisan news sharing behaviors. It is evident from our observations here that the benefit of both perspectives is greater than the sum of its parts: an analysis that addresses only the content factors that determine shareworthiness ignores the considerable diversity of interests, opinions, and ideologies that is likely to exist amongst the social media users involved in sharing any given news story; conversely, an analysis that builds exclusively on the motivated reasoning undertaken by users as they share news stories ignores that such reasoning will unfold in vastly divergent ways for different types of content. In other words, shareworthiness and motivated reasoning are two sides of one coin, interdependent, and thus inextricably interlinked – and our approach here provides a model for a more systematic exploration of that linkage.

Limitations

While our selection of only one outlet (*Infowars*) and our timeframe of one week of articles, on a single social media platform, limits the generalizability of our findings, the approach we have taken here provides a useful model for further research. Future studies would need to apply this approach to a larger number of hyper-partisan media outlets, over a longer period of time, and for multiple social media platforms, and/or could compare sharing patterns for these hyper-partisan news sites with their more mainstream counterparts. In some studies, variables such as the time of publication and the author of the article were also found to influence sharing practices (Xu et al. 2020), and these factors could also be incorporated into further analysis.

For our cluster analysis, initial profile clusters were selected based on our understanding of the data and might, therefore, display a bias. Moreover, we saw that 45% of collected accounts did not fall into one of these clusters, inducing a selection bias. Likewise, we cannot vouch for the accuracy and reliability of the self-descriptions provided by Twitter accounts. An account description might not necessarily represent an individual's attitude and can, therefore, only be treated as a proxy; more reliable information on attitudes could be gained only from direct interviews or surveys of all users, but this would introduce significant new methodological challenges. As we have shown by assessing cluster correlations, moreover, it is possible that profiles align with multiple clusters, and that this gives rise to a hierarchy of clusters and sub-clusters. For example, almost all account descriptions in our dataset aligned with the pro-Trump cluster – except for those of explicit *Infowars* supporters, which correlated more closely with conspiracy theorists. This suggests that these accounts might fall into two higher-level categories: Trump supporters and non-supporters. Further studies may attempt to use computational machine learning techniques, building on the profile descriptions as well as the tweets posted by accounts, to generate a more detailed range of account clusters, for instance, and/or they could follow Bruns, Moon, Münch, and Sadkowsky (2017) in determining account clusters based on patterns in the accounts' follower/followee networks.

Future studies

One observation from the manual coding of the tweets was that only a fraction of the tweets added further text beyond the original article headline. This may indicate the use of manual social sharing functions embedded on the article page itself; however, it could also result from automated news sharing. Automated sharing is often associated with automated or semi-automated computer programs, so-called social bots; however, such functionality is not limited only to (malicious or deceptive) bot accounts: services like *IFTTT* or *d/vr.it* enable (benign) automated social media posts on otherwise human-run accounts as well. The use of such services may be indicated by the use of specific URL shorteners (e.g. *ift.tt*), or by a service signature in the tweet metadata; future studies of shareworthiness factors should account, as far as possible, for such automatic sharing of news.

Overall, our findings concerning hypotheses H1-6 show that the hyper-partisan news outlet *Infowars* increases its shares through angles of human interest and conflict, while proximity, which has previously been found to increase shares, reduces shareworthiness. We also find that changing the operationalization of shareworthiness impacted on the outcome. We therefore suggest a differentiated approach to shareworthiness that distinguishes between the factors involved in *cross-platform sharing* and *in-platform on-sharing*,

respectively. Future studies should also investigate if the importance of human-interest and conflict factors is typical for hyper-partisan news sites like *Infowars*, or if our results are unique to this specific outlet and timeframe.

Likewise, our results concerning a motivated reasoning approach to news-sharing are promising. We find that three of the major clusters (self-declared Trump supporters, Christians, and *Infowars* supporters), were indeed more likely to share news content congruent with these worldviews. As both approaches yielded promising results, we argue for an integrative model of news sharing, which considers both the shareworthiness factors in articles and the underlying attitudes of users.

References

- Ahva, L., & Pantti, M. (2014). Proximity as a journalistic keyword in the digital era: A study on the “closeness” of amateur news images. *Digital Journalism*, 2(3), 322–333. <https://doi.org/10.1080/21670811.2014.895505>
- An, J., Quercia, D., & Crowcroft, J. (2013). Fragmented social media: A look into selective exposure to political news. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, (1), 51–52.
- Aneez, Z., Neyazi, T. A., Kalogeropoulos, A., & Nielsen, R. K. (2019). India Digital News Report. *Reuters Institute*. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-03/India_DNR_FINAL.pdf
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037//1089-2680.5.4.323>
- Beauvois, J.-L., & Joule, R.-V.-. (1996). *A radical dissonance theory*. London: Taylor & Francis.
- Bednarek, M., & Caple, H. (2017). *The discourse of news values: How news organizations create newsworthiness*. New York: Oxford University Press.
- Benkler, Y., Faris, R., Roberts, H., & Zuckerman, E. (2017, March 3). Study: Breitbart-Led Right-Wing Media Ecosystem Altered Broader Media Agenda. *Columbia Journalism Review*. <http://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- Bruns, A. (2018). *Gatewatching and news curation: Journalism, social media, and the public sphere*. New York: Peter Lang Publishing.
- Bruns, A., Moon, B., Münch, F., & Sadkowsky, T. (2017). The Australian Twittersphere in 2016: Mapping the Follower/Followee Network. *Social Media + Society*, 3(4), 1–15. <https://doi.org/10.1177/2056305117748162>
- Bruns, A., & Keller, T. R. (2020). News Diffusion on Twitter: Comparing the Dissemination Careers for Mainstream and Marginal News. Paper presented at the Social Media & Society 2020 conference, online, 22 July 2020. <https://www.youtube.com/watch?v=pCKpDkC8iqI>

- Chaiken, S. (1987). The heuristic model of persuasion. In M. Zanna, J. Olson, & C. Herman (Eds.), *Social influence: The Ontario symposium* (pp. 3–39). Hillsdale, NJ: Lawrence Erlbaum.
- Chaiken, Shelly, Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The Psychology of action: Linking cognition and motivation to behavior* (pp. 553–578). New York, NY: Guilford Press.
- Clemm von Hohenberg, B. (2019). *An ocean of possible truth: Biased processing of news on social media*. Retrieved from <https://ssrn.com/abstract=3281038>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.
- de Hoog, N. (2013). Processing of social identity threats. A defense motivation perspective. *Social Psychology*, 44, 361–372. <https://doi.org/10.1027/1864-9335/a000133>
- de Vreese, C. H. (2012). New avenues for framing research. *American Behavioral Scientist*, 56(3), 365–375. <https://doi.org/10.1177/0002764211426331>
- Druckman, J. N., Leeper, T. J., & Slothuus, R. (2016). Motivated responses to political communications: Framing, party cues, and science information. In H. Lavine & C. S. Taber (Eds.), *The Feeling, Thinking Citizens* (1st ed., pp. 125–150). Routledge. <https://doi.org/10.4324/9781351215947>
- Eilders, C. (2006). News factors and news decisions. Theoretical and methodological advances in Germany. *Communications*, 31(1), 5–24. <https://doi.org/10.1515/COMMUN.2006.002>
- Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 U.S. Presidential election, 1–13. Retrieved from <https://cyber.harvard.edu/publications/2017/08/mediacloud>
- Festinger, L. (1976). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Figenschou, T. U., & Ihlebæk, K. A. (2019). Media criticism from the far-right: Attacking from many angles. *Journalism Practice*, 13(8), 901–905. <https://doi.org/10.1080/17512786.2019.1647112>
- Fletcher, R., & Nielsen, R. K. (2018). Are People Incidentally Exposed to News on Social Media? A Comparative Analysis. *New Media & Society*, 20(7), 2450–2468. <https://doi.org/10.1177/1461444817724170>
- Funt, D., Gourarie, C., & Murtha, J. (2016, June 27). In brands we trust? The New Yorker, BuzzFeed, and the push for digital credibility. *Columbia Journalism Review*. http://www.cjr.org/special_report/newyorker_buzzfeed_trust.php
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news. *Journal of Peace Research*, 2, 64-91. <https://doi.org/10.1177/002234336500200104>
- García-Perdomo, V., Salaverría, R., Kilgo, D. K., & Harlow, S. (2018). To share or not to share: The influence of news values and topics on popular social media content in the United

- States, Brazil, and Argentina. *Journalism Studies*, 19(8), 1180–1201. <https://doi.org/10.1080/1461670X.2016.1265896>
- Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, 62(1), 180–196.
- Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 U.S. election. *Nature Human Behaviour*, 4, 472–480.
- Kwak, H., & An, J. (2016). Revealing the Hidden Patterns of News Photos: Analysis of Millions of News Photos through GDELT and Deep Learning-based Vision APIs. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1). Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14840>
- Hermida, A. (2012). Tweets and Truth: Journalism as a Discipline of Collaborative Verification. *Journalism Practice*, 6(5–6), 659–668. <https://doi.org/10.1080/17512786.2012.667269>
- Holt, K., Figenschou, T. U., & Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, 7(7), 860–869. <https://doi.org/10.1080/21670811.2019.1625715>
- Jebril, N., Vreese, C. H. De, Dalen, A. Van, & Albæk, E. (2013). The effects of human interest and conflict news frames on the dynamics of political knowledge gains: Evidence from a cross-national study. *Scandinavian Political Studies*, 36(3), 201–226. <https://doi.org/10.1111/1467-9477.12003>
- Kaiser, J., Keller, T. R., & Kleinen-von Königslöw, K. (2018). Incidental news exposure on facebook as a social experience: The influence of recommender and media cues on news selection. *Communication Research*, 1–23. <https://doi.org/10.1177/0093650218803529>
- Karnowski, V., Kümpel, A. S., Leonhard, L., & Leiner, D. J. (2017). From incidental news exposure to news engagement. How perceptions of the news post and news usage patterns influence engagement with news articles encountered on Facebook. *Computers in Human Behavior*, 76, 42–50. <https://doi.org/10.1016/j.chb.2017.06.041>
- Keller, T. R. (2020). *To whom politicians talk and listen? Mapping Swiss politician's sphere on Twitter*. *Computational Communication Research*, 2(2), 175–202. <https://doi.org/10.5117/CCR2020.2.003.KELL>
- Kilgo, D. K., Lough, K., & Riedl, M. J. (2020). Emotional appeals and news values as factors of shareworthiness in Ice Bucket Challenge coverage. *Digital Journalism*, 8(2), 267–286. <https://doi.org/10.1080/21670811.2017.1387501>
- Kim, H. S. (2015). Attracting views and going viral: How message features and news-sharing channels affect health news diffusion. *Journal of Communication*, 65(3), 512–534. <https://doi.org/10.1111/jcom.12160>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Leetaru, K., & Schrod, P. A. (2013). GDELT: Global data on events, location and tone, 1979–2012. *Annual Meeting of the International Studies Association*, (April), 1979–2012. Retrieved from <http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf>

- Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. Cognition, Choice, and the Bounds of Rationality. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of Reason* (pp. 183–213). Cambridge University Press. <https://doi.org/10.1017/cbo9780511805813.009>
- Lowin, A. (1969). Further evidence for an approach-avoidance interpretation of selective exposure. *Journal of Experimental Social Psychology*, 5(3), 265–271.
- Marwick, A. (2018). Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technology Review*, 2(2), 474–512. <https://doi.org/10.4018/ijkm.2018070101>
- Masterton, M. (2005). Asian journalists seek values worth preserving. *Asia Pacific Media Educator*, 16(16), 41–48. Retrieved from <http://ro.uow.edu.au/apme/vol1/iss16/6>
- Mitchell, A. (2015). State of the News Media. *Pew Research Center*, 151. <https://doi.org/10.1145/3132847.3132886>
- Morgan, J. S., Shafiq, M. Z., & Lampe, C. (2013). Is news sharing on twitter ideologically biased? *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 887–896. <https://doi.org/10.1145/2441776.2441877>
- Neuman, W. R., Neuman, R. W., Just, M. R., & Crigler, A. N. (1992). *Common knowledge: News and the construction of political meaning*. University of Chicago Press.
- Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). Reuters Institute Digital News Report 2017. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf
- Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019. Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_1.pdf
- Ng, Y., & Zhao, X. (2020). The human alarm system for sensational news, online news headlines, and associated generic digital footprints: A uses and gratifications approach. *Communication Research*, 47(2), 251–275. <https://doi.org/10.1177/0093650218793739>
- Olmstead, K., Mitchell, A., & Rosenstiel, T. (2011). Navigating news online: Where people go, how they get there and what lures them away. *Pew Research Center*.
- Östgaard, E. (1965). Factors influencing the flow of news. *Journal of Peace Research*, 2(1), 39–63. <https://doi.org/10.1177/002234336500200103>
- Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media*, 9, 17–29. <https://doi.org/10.1016/j.osnem.2018.11.001>
- Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting social unrest events with hidden markov models using GDELT. *Discrete Dynamics in Nature and Society*, 2017. <https://doi.org/10.1155/2017/8180272>

- Rubeking, B. (2019). Emotion, attitudes, norms and sources: Exploring sharing intent of disgusting online videos. *Computers in Human Behavior*, 96, 63–71. <https://doi.org/10.1016/j.chb.2019.02.011>
- Satariano, A., & Alba, D. (2020). Burning cell towers, out of baseless fear they spread the virus. *New York Times*. Retrieved from <https://www.nytimes.com/2020/04/10/technology/coronavirus-5g-uk.html>
- Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2), 93–109. <https://doi.org/10.1111/j.1460-2466.2000.tb02843.x>
- Shao, C., Hui, P. M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS ONE*, 13(4), 1–23. <https://doi.org/10.1371/journal.pone.0196087>
- Shearer, E., & Grieco, E. (2019). Americans are wary of the role social media sites play in delivering the news. *Pew Research Center*.
- Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83, 278–287. <https://doi.org/10.1016/j.chb.2018.02.008>
- Spierings, N., Jacobs, K., & Linders, N. (2019). Keeping an eye on the people: Who has access to MPs on Twitter? *Social Science Computer Review*, 37(2), 160–177. <https://doi.org/10.1177/0894439318763580>
- Steele, C. M., & Liu, T. J. (1981). Making the dissonant act unreflective of self: Dissonance avoidance and the expectancy of a value-affirming response. *Personality and Social Psychology Bulletin*, 7(3), 393–397. <https://doi.org/10.1177/014616728173004>
- Syn, S. Y., & Oh, S. (2015). Why do social network site users share information on Facebook and Twitter? *Journal of Information Science*, 41(5), 553–569. <https://doi.org/10.1177/0165551515585717>
- Tandoc, E. C., & Johnson, E. (2016). Most students get breaking news first from Twitter. *Newspaper Research Journal*, 37(2), 153–166. <https://doi.org/10.1177/0739532916648961>
- Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism and Mass Communication Quarterly*, 94(1), 38–60. <https://doi.org/10.1177/1077699016654682>
- Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, 20(5), 520–535. <https://doi.org/10.1111/jcc4.12127>
- Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication*, 67(5), 803–826. <https://doi.org/10.1111/jcom.12325>
- Van den Bulck, H., & Hyzen, A. (2020). Of lizards and ideological entrepreneurs: Alex Jones and Infowars in the relationship between populist nationalism and the post-global media

ecology. *International Communication Gazette*, 82(1), 42–59.
<https://doi.org/10.1177/1748048519880726>

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>

Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Report DGI(2017)09. Council of Europe. <https://shorensteincenter.org/wp-content/uploads/2017/10/Information-Disorder-Toward-an-interdisciplinary-framework.pdf>

Wasike, B. S. (2013). Framing news in 140 characters: How social media Editors frame the news and interact with audiences via Twitter. *Global Media Journal*, 6(1), 5–23.

Weber, P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New Media & Society*, 16(6), 941–957. <https://doi.org/10.1177/1461444813495165>

Xu, W. W., Sang, Y., & Kim, C. (2020). What drives hyper-partisan news sharing: Exploring the role of source, style, and content. *Digital Journalism*, 8(4), 486–505. <https://doi.org/10.1080/21670811.2020.1761264>

Yonamine, J. E. (2013). A nuanced study of political conflict using the global datasets of events location and tone (GDELT) dataset. *ProQuest Dissertations and Theses*, (August), 186.

Table 1: Results of the negative binomial regression model for the dependent variable tweet count (** $p < .001$, ** $p < .01$, * $p < .05$).

Factor	<i>B</i>	<i>S.E.</i>	<i>p</i>
Proximity	-1.01	0.26	<.0001***
Conflict	0.57	0.2	.003**
Human interest	0.77	0.19	<.0001***
Morality	0.33	0.21	.11
Visual	-0.07	0.21	.74

Table 2: Results of the negative binomial regression model for the dependent variable retweet factor (** $p < .001$, ** $p < .01$, * $p < .05$).

Factor	<i>B</i>	<i>S.E.</i>	<i>p</i>
Proximity	-0.25	0.21	.22
Conflict	0.25	0.15	.11
Human interest	0.33	0.16	.04*
Morality	0.01	0.17	.95
Visual	0.003	0.16	.98

Table 3: Correlation matrix of Cramer's V for Twitter profiles (higher values = greater correlation).

	Trump	Christian	Patriot	Pro-Gun	Military	Conspiracy	Infowars
Trump	1	0.27	0.45	0.45	0.35	0.23	0.14
Christian		1	0.09	0.13	0.15	0.07	0.02
Patriot			1	0.51	0.34	0.05	0.04
Pro-Gun				1	0.51	0.06	0.001
Military					1	0.05	0.05
Conspiracy						1	0.31
Infowars							1

Table 4: Correlation matrix of Cramer's V for *Infowars.com* articles.

	Trump	Christian	Patriot	Pro-Gun	Military	Conspiracy	Infowars
Trump	1	0.06	0.02	0.05	<0.001	0.05	0.1
Christian		1	0.23	0.32	0.19	0.13	0.15
Patriot			1	0.22	0.22	0.02	0.16
Pro-Gun				1	0.03	0.13	0.1
Military					1	0.004	0.06
Conspiracy						1	0.31
Infowars							1

Table 5: Results of the logistic regression models with articles that reported on a particular topic as dependent variable, profiles with interests matching that topic as independent variable (bolded), and the remaining profiles as control variables (** $p < .001$, ** $p < .01$, * $p < .05$).

Dependent Variable (article)	Independent and control variables (profile)	<i>B</i>	<i>S.E.</i>	<i>p</i>
Trump Articles	Trump Profile	0.27	0.07	.003***
	Patriotic Profile	-0.06	0.1	.55
	Infowars Profile	0.15	0.17	.37
	Christian Profile	0.07	0.09	.42
	Military Profile	-0.005	0.12	.97
	Pro-Gun Profile	0.04	0.12	.74
	Conspiracy Profile	0.26	0.15	.08
Patriotic Articles	Trump Profile	0.05	0.11	.63
	Patriotic Profile	-0.35	0.16	.03*
	Infowars Profile	0.23	0.24	.33
	Christian Profile	0.17	0.13	.21
	Military Profile	-0.26	0.19	.17
	Pro-Gun Profile	0.44	0.18	.01*

	Conspiracy Profile	0.15	0.21	.49
	Trump Profile	-0.07	0.09	.41
	Patriotic Profile	-0.02	0.11	.89
	Infowars Profile	0.36	0.18	.05*
Infowars Articles	Christian Profile	-0.13	0.11	.22
	Military Profile	0.32	0.14	.02*
	Pro-Gun Profile	-0.01	0.14	.95
	Conspiracy Profile	0.01	0.17	.54
	Trump Profile	-0.1	0.15	.49
	Patriotic Profile	-0.22	0.2	.27
	Infowars Profile	-0.07	0.33	.82
Christian Articles	Christian Profile	0.45	0.16	.01*
	Military Profile	-0.06	0.25	.82
	Pro-Gun Profile	-0.06	0.25	.82
	Conspiracy Profile	0.51	0.25	.04*
	Trump Profile	0.09	0.23	.71
	Patriotic Profile	-0.21	0.32	.51
Military Articles	Infowars Profile	-0.002	0.49	.99
	Christian Profile	-0.44	0.33	.18
	Military Profile	-0.44	0.42	.3
	Pro-Gun Profile	0.28	0.38	.46
	Conspiracy Profile	0.16	0.43	.71
	Trump Profile	-0.2	0.21	.34
	Patriotic Profile	-0.05	0.26	.84
Pro-Gun Articles	Infowars Profile	0.27	0.4	.51
	Christian Profile	0.48	0.22	.03*
	Military Profile	-0.07	0.32	.82
	Pro-Gun Profile	0.18	0.32	.57

	Conspiracy Profile	0.41	0.34	.22
	Trump Profile	0.002	0.08	.98
	Patriotic Profile	0.02	0.1	.81
	Infowars Profile	0.32	0.17	.06
Conspiracy Articles	Christian Profile	0.003	0.1	.97
	Military Profile	0.13	0.13	.32
	Pro-Gun Profile	-0.28	0.13	.03*
	Conspiracy Profile	-0.15	0.16	.35

ARTICLE 2

The following article is reused from:

Wischnewski, M., Bernemann, R., Ngo, T., & Krämer, N. (2021). Disagree? You must be a bot! How beliefs shape Twitter profile perceptions. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan*. New York, NY: ACM. <https://doi.org/10.1145/3411764.3445109>

Disagree? You Must be a Bot! How Beliefs Shape Twitter Profile Perceptions

ANONYMOUS AUTHOR(S)

In this online experiment, we investigate how well individuals can detect social bots on Twitter. Following motivated reasoning theory from social and cognitive psychology, our central hypothesis is that especially those accounts which are opinion-incongruent will be perceived as social bot accounts when the account is ambiguous about its nature. We also hypothesize that credibility rating mediates this relationship. We asked $N = 151$ participants to evaluate 24 Twitter accounts and decide whether the accounts were humans or social bots. Findings support our motivated reasoning hypothesis: Accounts that are opinion-incongruent are evaluated as relatively more bot-like than accounts that are opinion-congruent. Moreover, it does not matter whether the account is clearly social bot/clearly human or ambiguous about its nature. This was mediated by perceived credibility in the sense that congruent profiles were evaluated to be more credible resulting in lower perceptions as bots.

Additional Key Words and Phrases: motivated reasoning, social bots, Twitter, credibility, partisanship, bias

ACM Reference Format:

Anonymous Author(s). 2021. Disagree? You Must be a Bot! How Beliefs Shape Twitter Profile Perceptions. In *Proceedings of Yokohama '21: ACM CHI Conference on Human Factors in Computing Systems (Yokohama '21)*. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Today's social media landscape not only enables individuals to access a rich well of information and to connect with others but also allows for user-generated content to be published. However, not all activity on social media platforms originates from human users. In recent years, concerns about the influence of automated accounts, so-called social bots, has risen. Social bots on social media are generally defined as automatic or semi-automatic accounts run by computer algorithms. As they often mimic human behavior (e.g. posting content, "liking", and "retweeting"), their automated nature can go unnoticed to human users. Social bots have been associated to a plethora of different functions like copy-paste bots which post the same content multiple times to gain attention [40], amplifier accounts that boost particular sources by aggregating and repeating content [21], fake followers to boost follower counts and popularity, or online trolls which engage in malicious activity, targeting (vulnerable) users. Recent findings suggest that between 9 to 15% of active Twitter accounts were run by social bots [45].

The influence of social bots has become a public but also an academic concern. Bots have been, for example, accused of steering discussions online to promote specific ideas, spread misinformation [46], engage in political astroturfing [27], affecting the stability of financial markets [7] and endorse conspiracy theories. Moreover, influential bot activity has been found in Japan's general election in 2014 [40], the Brexit referendum in 2016 [4], the U.S. presidential election in 2016 [5], and the French general election in 2017 [17]. Besides their role in election campaigns, it was found that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

social bot activity linked to detrimental effects on public health like the promotion of e-cigarettes [1] and the promotion of anti-vaccination content [8].

Due to their detrimental effects, many efforts have been made to detect social bots online. Most solutions employ machine learning models for automated detection, which rely on account features like posting behavior, timing, or social networking structures. For a review of detection techniques, see [26]. Another methodological approach suggests identifying social bots through coordinated behavior or co-retweeting activity [27]. An example of such a strategy is to identify accounts that share identical content within a short period of time.

While these automated techniques provide one solution to detect social bots, we want to focus on the lesser attended users' side of social bot detection. In this contribution, we want to pose the following research question:

RQ: When and why do users perceive an account as a social bot?

To answer this question concerning users' perception, we rely on findings from social and cognitive psychology which points out that individuals sometimes perceive information in a biased manner. Especially, the theory of motivated reasoning suggests that information that is in line with prior-opinions and attitudes is favored over information that contradicts prior-opinions and attitudes. This results in overconfidence as well as overreliance when information is opinion-congruent but also in rejection and overcritical assessment of opinion-incongruent information. Concretely, this could result in users blindly approving of accounts only because the account represents one's own opinion, while accounts that disagree with one's own opinion are rejected or discredited as social bots.

Therefore, once we understand when and how users perceive an account as a social bot, users can be assisted through media literacy interventions as well as assisting tools. While media literacy interventions could, for example, correct erroneous user perceptions, assisting tools can support users in the identification process itself. Considering the detrimental effects of social bots, it is important that users can correctly identify social bot accounts.

As a first concern of this paper, we want to briefly summarize users' knowledge about social bots, users' engagement with social bots, users' detection abilities, and users' acceptance of social bots. Building on this as well as the theory of motivated reasoning, we present a pre-registered study, investigating users' ability to detect political social bots on Twitter. In doing so, we hypothesized that users' bot detection ability is biased in a way that opinion-congruent bots are less likely and opinion-incongruent bots more likely to be detected whereas opinion-congruent human profiles are perceived as such but opinion-incongruent human profiles become less human. Because in the past, social bots have been found to influence political events such as elections, we chose the context of political communication for our study. Similarly, we selected Twitter as a social medium as a substantial presence of social bots has previously been found on Twitter [11] and it is a popular social medium for political communication [23].

This paper makes the following contributions:

- An empirical investigation of human interaction with social bots which is grounded in social psychological theory
- Deeper understanding of when and why users perceive an account as a social bot
- Guidance for developers of assisting tools which want to support users in the detection of social bots

2 THEORETICAL BACKGROUND

2.1 What users know about, how they engage, how they detect and when they accept social bots

The acclaimed purpose of some social bots is to engage with users. Results for how users interact with social bots are, however, mixed. Building on the Computers are Social Actors (CASA) paradigm [39], Edwards and colleagues

[15], for example, wanted to know if social bots are perceived differently from human agents when communicating on Twitter. As CASA suggests, they found that individuals perceived social bots "as a credible source of information" (p. 374) and competent communicator. They also found that individuals showed no difference in intending to interact between a human agent or a social bot. These results contradict later findings by Murgia and colleagues [35] who investigated human-bot interaction on Stack Overflow, a question and answer site for developers. Their results indicate that, although communication of the human agent and social bot was identical, answers given by the social bot were less expected and received more down-votes than the human agent. This was especially prevalent when the bot made an erroneous statement, leading the authors to suggest that humans have a "low tolerance for mistakes by a bot" (p. 6). The two opposing results could be explained, however, by the varying functionality of social bots in the experiments. While Edwards and colleagues [15] used social bots to inform users, social bots in Murgia's et al. [35] study were used to engage with users by providing answers to user questions.

Both studies assume, however, that users know about social bots. Yet, are users on social media aware of social bots? Answers to that are scarce. One exception comes from the Pew Research Center [42]. The authors found that two-thirds of U.S. Americans had at least heard about social bots, although only 16% heard about them a lot, with younger individuals being more likely to have heard about them. Moreover, most people believed that social bots were used for malicious purposes (80%). The authors also asked those participants, who indicated to know about social bots, how confident they were to detect a social bot on social media. About one half of the participants were somewhat confident to confident that they would recognize a social bot account on social media, with younger individuals, again, being more confident than elder individuals. This indicates that, with 60% knowing of social bots and of those 50% feeling confident to detect social bots, only one third of U.S. Americans felt confident to detect social bots.

To support users in identifying, especially, malicious social bots, different countermeasures have been developed and tested concerning their usability. Through a user experience survey, Yang et al. [48] wanted to know how users engage with Botometer, one of the early A.I. bot detection tools available to the public as well as a commonly used tool for academic purposes [12]. If users find a Twitter account to be suspicious, they can enter the account name into Botometer which will return a probability score. Their results showed that most users found Botometer scores easy to understand (80%). Moreover, the authors found that users were equally concerned about false positives (humans erroneously classified as bots) and false negative results (bots erroneously classified as humans) a concern that has echoed recently within academia as well (see Rauchfleisch and Kaiser [37]). However, the survey by Yang et al. [48] does neither answer why and how people selected accounts they wanted to test nor how they interpreted the results. In other words, their work leaves open how people detected a suspicious account and why they made the decision to verify it on Botometer.

Concerning the acceptance of social bots, Hong and Oh [20] investigated how self-efficacy in identifying social news bots and greater prevalence of social news bots increased users' acceptance. The authors found that with increased self-efficacy in identifying social news bots, social news bots become more acceptable. Likewise, the more prevalent news from social news bots were perceived, the more acceptable they became. The authors argue, however, that self-efficacy was measured through subjective evaluations so that it was uncertain whether individuals who indicated high self-efficacy in identifying could actually identify social news bots correctly.

Concludingly, while some first attempts have been made to understand how users interact with social bots, how many users know of social bots, how they apply assisting tools such as Botometer, and when users accept social bots, it remains unknown when and why users detect social bots.

2.2 Motivated reasoning theory and credibility cues in human social bot detection

In this study, we provide a first attempt to understand the processes behind human social bot detection. In doing so, we apply the theory of motivated reasoning to understand users' detection abilities and relate this to credibility cues. Motivated reasoning theory generally proposes that incoming information is processed in a way to maintain existing attitudes and beliefs and disregard, reject or downplay opposing views. For example, a person cheering for a specific soccer club will reject any criticism against this club but happily agree that one's own team is the best in the world. While this example portrays a rather harmless consequence of motivated reasoning, engaging in motivated reasoning can become harmful easily. Replacing the soccer club in the example with a conspiracy follower might result into persons harming themselves or others.

Different underlying psychological mechanisms of motivated reasoning have been identified. Early investigations, for example, suggest that this bias is a result of different motivational states [10, 29]. Others have connected motivated reasoning to cognitive biases such as confirmation bias [36] or myside bias [33] as well as affect-driven, biased memory retrieval [43].

Although the causes for motivated reasoning are debated within social psychology, effects of it have been found in various disciplines such as science communication [41], political communication [13], information selection on social media [47], misinformation detection and processing [44] or fact-checking [14]. While motivated reasoning has been found in many domains, real-world repercussions of motivated reasoning have been mostly associated with political decision making [38]. In one study, Bisgaard [6] found, for example, that while partisans are willing to except a decrease of economic conditions, responsibilities were, however, attributed in a way to protect one's loyalties. In the context of misinformation, results from Ecker and Ang [14] indicated that, even after being retracted, misinformation had a continued influence when they were in line with previous attitudes.

In line with the existing literature on motivated reasoning, we suggest that social media users engage in motivated reasoning when perceiving social bots, resulting in an overconfidence and "blindness" towards social bots that promote favored content, and skepticism and rejection towards bots promoting opposing views. We assume, however, that this motivated social bot detection is limited to accounts which are not clearly distinguishable as human- or bot-run (ambiguous accounts). If accounts can clearly be identified as human or bot (unambiguous accounts), we suggest that users do not engage in motivated bot detection even if an account endorses an attitude contrary to one's own because individuals "draw the desired conclusion only if they can muster up the evidence necessary to support it" [29]. In other words, we assume that, if an account shows undeniable signs of automation or humanness, users cannot reason themselves into believing it is not automated or automated, respectively.

To summarize, we hypothesize:

Hypothesis 1

- a For accounts that are ambiguous whether they display a human or a bot account (ambiguous profiles), opinion-congruent accounts are more likely to be identified as human accounts.
- b For accounts that are ambiguous whether they display a human or a bot account (ambiguous profiles), opinion-incongruent accounts are more likely to be identified as bot accounts.

Hypothesis 2

- a For accounts that are clearly human (unambiguous human profiles), there is no differences between congruent and incongruent profiles concerning the correct identification as human accounts.

- b For accounts that are clearly bot (unambiguous bot profiles), there is no differences between congruent and incongruent profiles concerning the correct identification as bot accounts.

In addition to this, we argue that perceived account credibility will influence users' perception. It was found that users on social media employ a number of cues, specific to the respective platform, to arrive at a credibility evaluation [34]. Cues can concern the *source*, like perceived knowledge, passion, transparency, reliability and influence as well as the *message* or *content*, like consistency, accuracy and authenticity [25]. In the semantics of social media platforms, these cues have been translated into features like metadata, follower/followee ratio, linguistics, use of hashtags, or visual- and image-based features [24]. Another strand of research examines credibility cues related to *posting behavior*, meaning, for example, temporal patterns of tweeting and retweeting [9]. In other words, accounts were perceived as less credible when post occurred regularly and did not follow a circadian rhythm. In addition, a recent finding by [3] could show that users relied mostly on content cues such as whether an account shared random or nonsensical content as well as commercial content and poor language. However, in recent years, bots have become more sophisticated in their efforts to disguise their automated nature [2].

In connection with motivated reasoning theory, we argue that the use of the described credibility cues is biased in favor of own attitudes and beliefs. Hence, perceived credibility depends on the displayed opinion of an account. In turn, more credible accounts should be perceived as more human, whereas less credible accounts should be perceived as more bot-like. We hypothesize:

Hypothesis 3

We hypothesize that the relationship between opinion-congruency and account perception is mediated by the perceived account credibility.

All hypotheses were pre-registered under <https://osf.io/97mcr/>. Hypothesis 3 was erroneously preregistered as a moderation hypothesis which was corrected in the online registration.

3 METHOD

A report on how we determined our sample size, data exclusions, manipulations and measures as well as the dataset of this study are available online (<https://osf.io/36mkw/>). The study received ethical approval by the ethics committee of XXXX.

3.1 Sample

To test our hypotheses, we conducted a within-subject design with two independent factors, ambiguity and congruency. The factor ambiguity consisted of three levels "human", "social bot" and "ambiguous", specifying whether a Twitter account is ambiguous about its nature (neither clearly social bot nor clearly human) or not. The factor congruency consisted of the two levels, "opinion-incongruent" and "opinion-congruent". For the online experiment, we recruited participants through Prolific.co, a UK-based online survey platform similar to Amazon Mechanical Turk. The experiment was conducted in late July 2020. Participation was restricted to U.S. Americans. The final sample consisted of $N = 151$ participants (60 male, 84 female, 6 non-binary, 1 not shared) with an age range from 18 to 68 years ($M = 28.86$, $SD = 10.18$). A majority of participants held at least a High School degree ($n = 72$) or a Bachelor's Degree ($n = 54$). Most participants were Whites ($n = 86$), followed by Asian Americans ($n = 28$) and Black or African Americans ($n = 16$).

3.2 Procedure

The general task of participants was to identify whether an Twitter account represented a human-run account or a social bot. Before entering this task, we provided participants with a working definition¹ of social bots to ensure that all participants held a similar understanding. After that, we presented participants with 24 Twitter profiles, 12 of which had been previously clearly identified as representing a human or bot account, and 12 of which had been previously identified as ambiguous profiles which were neither clearly human nor bot (see section stimulus material for an in-depth description of how profiles were build). In addition, for each type, unambiguous and ambiguous, six profiles were created to represent a Democrat account and six profiles to represent a Republican account. Again, to validate this distinction, profiles were pretested.

3.3 Measures: Dependent variables and control variables

After viewing each Twitter profile, participants were asked to indicate on a continuous sliding scale from 0 to 100 in whole integers, if they thought the profile represented a bot or a human run account, with values lower than 50 indicating a bot account and values greater than 50 a human account. The stronger it was perceived as a social bot, the lower the value and the stronger it was perceived as a human account, the higher the value. This constituted the dependent variable *profile perception*.

Participants were then asked to indicate how credible they found the displayed profile. To assess credibility measures, we used the 'trustworthiness' subscale from McCroskey and Teven's [31] source credibility scale. Credibility was measured through six semantic descriptions with two antonyms on each side of a 7-point Likert scale (honest/dishonest; untrustworthy/trustworthy (R); honorable/dishonorable; moral/immoral; unethical/ethical (R); phony/genuine (R)). The sub-scale achieved a satisfactory Cronbach's $\alpha = .94$.

Because other variables might also contribute to the perception of social bots, we included several control variables in our study. These consisted of standard demographic data like age, gender, and education but also included a measure for hours spend on social media per day, number of actively used social media platforms, whether Twitter was actively used, whether participants knew about social bots as well as media literacy. To control for media literacy, we used the subscale critical consumption developed by Koc and Barut [28] which consists of eleven items that are answered on a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree. The sub-scale achieved a satisfactory Cronbach's $\alpha = .87$.

3.4 Stimulus material and independent variables

For the first factor, *ambiguity*, we manipulated the displayed Twitter profiles in a way that they were either unambiguous or ambiguous about their nature (human- or bot-run). We selected three main characteristics that have previously been found to alter individuals' credibility perception. These characteristics were timing of posting behavior, shared content, and the displayed profile picture [16, 18, 24, 32]. Concerning the timing of posting behavior, we manipulated profiles in a way that postings were either frequent (e.g. post appeared every two minutes) or infrequent. For the shared content, we manipulated whether the profile (a) tweets only, (b) retweets only, (c) shares links only, or (d) shares a mix of a, b, and c. Concerning the profile picture, we either showed a picture of a person or a picture displaying a graphic

¹Working definition that was shown to participants before entering the detection task: "Social bots are automated online accounts that communicate more or less autonomously. They typically operate on social media sites, such as Twitter. Social bots serve different functions. Some are programmed to forward, like or comment on specific topics like the weather, sports results, but also political issues. Some others are programmed by companies to disseminate advertisements within social networking sites. Some bots state openly that they are automated accounts. Other accounts do not disclose their automated nature and can hardly be differentiated from human run accounts."



Fig. 1. Example for a neutral bot profile from the pre-test (left) and the final democrat bot profile used in the study (right). Here, we manipulated the timing of posts, the content of posts (non-sensical) as well as the sharing of only original tweets (no retweets).

(non-human). All profiles were mock-ups and did not represent real accounts on Twitter. This resulted in overall 15 mock-up Twitter profiles which were pretested concerning their credibility. All mock-up profiles were created in a way that they shared non-political, neutral content (e.g. content on animals or nature) to avoid the hypothesized bias that accounts become more or less bot-like because of opinion-congruency.

Results of the pre-test indicated that three profiles were identified as social bots, six were identified as human accounts, and seven as ambiguous profiles, neither clearly bot nor clearly human. In a next step, we created unambiguous and ambiguous profiles. For the unambiguous profiles, we selected the three identified bot accounts and three of the identified human accounts to build new Twitter profiles. For each of these overall six accounts we created two version: one representing a Democrat account and one representing a Republican account. To generate ambiguous profiles, we selected six of the pretested profiles that were neither classified as clearly human nor bot and created again two version: one representing a Democrat account and one representing a Republican account. An overview of the final set of created Twitter accounts is given in Table 1. One example for an unambiguous (social bot) profile is given in Figure 1. Each profile consisted of ten posts. All created Twitter profiles are uploaded (TP1-TP24) and can be viewed under: <https://osf.io/36mkw/>.

For the second factor, congruency, we measured individual partisanship through the 4-item Partisan Identity Scale developed by Huddy et al. [22]. The factor was then calculated by adding participants' partisanship for congruent evaluations (congruent = Democratic Twitter profile for self-identified Democrats / Republican Twitter profiles for self-identified Republicans) and incongruent evaluations (incongruent = Republican Twitter profiles for self-identified Democrats / Democrat Twitter profiles for self-identified Republicans).

Table 1. Overview of all created Twitter profiles.

	Unambiguous		Ambiguous
	Bot	Human	
Democrat	3	3	6
Republican	3	3	6

Table 2. Correlation coefficients of all control variables correlated with each other. The variables *age*, *NML*, *SM platforms used*, and *time spend on SM* were continuous variables. The variables *social bot knowledge* (1 = knowing of social bots, 2 = not knowing of social bots) and *Twitter usage* (1 = using Twitter, 2 = not using Twitter) were entered as dichotomous variables (NML = New Media Literacy, SM = Social Media).

	1. Age	2. NML	3. SM Platforms used	4. Time spend on SM	5. Social bot knowledge	6. Twitter usage
1.	1	-.004	-.17*	-.22**	.01	-.03
2.		1	.14	.11	-.14	.23**
3.			1	.21**	-.15	.44**
4.				1	.07	.24**
5.					1	-.16*

** $p < .01$, * $p < .05$.

3.5 Strategy of analysis

In order to answer our hypotheses, we conducted a 2 x 3 repeated measure ANOVA. We originally pre-registered a 2 x 2 repeated measure ANOVA with only two levels of ambiguity. The distinction into three instead of two levels enabled us, however, to differentiate between human and social bot accounts. A detailed reasoning for a changed strategy of analysis can be found here (<https://osf.io/36mkw/>) as well as the results of the originally preregistered analysis.

For hypothesis 3 (hypothesizing that the relationship between congruency and account perception is mediated by perceived account credibility), we ran mediation analyses for all three levels of ambiguity, using the PROCESS macro Version 3 by Hayes [19] for SPSS. We entered a grouping variable representing the two levels of opinion-congruency as a predictor into the model, the perceived credibility ratings as a mediator and the Twitter profile perception as the outcome variable.

4 RESULTS

All analyses were conducted using R Studio Version 3.5 as well as SPSS Version 26 and pre-registered (<https://osf.io/97mcr/>).

4.1 Descriptive results

In addition to the demographic data described in the method section, we assessed five additional control variables: time spend on social media per day, number of social media platforms actively used, Twitter usage, social bot knowledge and media literacy. We found that two thirds of the participants in our sample spend between >1 and up to three hours a day on social media, while, on average, actively using 3-4 different social media platforms. Roughly two thirds reported to use Twitter actively, about 86% of the participants knew about social bots, and self-reported a mean media literacy of 3.9 ($SD = 0.58$) on a five-point Likert scale. Correlations of the control variables can be found in Table 2.

However, we wanted to know which accounts would be identified as social bots and whether participants would identify accounts in a biased manner (H1), assuming opinion-congruent accounts to be perceived as more human.

Table 3. Means and standard deviations of the dependent variable profile perception for the three levels of ambiguity. Values closer to 100 indicate higher perceived "humanness", while values closer to 0 indicate higher perceived "botness".

Ambiguity	Congruency	<i>M</i>	<i>SD</i>
Bot Profile	Congruent	39.7	20.8
	Incongruent	31.6	20.4
Human Profile	Congruent	76.6	17
	Incongruent	70.4	19.8
Ambiguous Profile	Congruent	57.4	16.2
	Incongruent	52.4	17.4

Descriptive results of the dependent variable profile perception by condition are reported in 3. A visual analysis through density distributions of the dependent variable profile perception is presented in Figure 2.

According to the visual analysis reported in Figure 2, we found the results of our pre-test confirmed. Our predefined Twitter profiles classified as social bot accounts were identified as such (means below 50), whereas accounts designed to represent humans were identified as human (means above 50). Similarly, accounts designed to be ambiguous were perceived as such (means close to 50).

4.2 Hypothesis testing through repeated measure ANOVA

As described in the section *strategy of analysis*, we conducted a 2 x 3 repeated measure ANOVA with the factors congruency (two levels) and ambiguity (three levels) to answer hypothesis H1 and H2. Because the assumption of sphericity was not met for the factor ambiguity, we used the Greenhouse-Geisser sphericity correction for this factor.

Supporting the results of the visual analysis reported in Figure 2-4, we found a significant main effect for ambiguity, $F(1.76, 150) = 7.87, p < .001, \eta_p^2 = .05$ (using Greenhouse-Geisser correction). This meant that participants evaluated profiles of the three different levels (bot, human, and ambiguous), in fact, differently. This supported our operationalization of the three different levels. In addition to this, we found a significant interaction of ambiguity with the control variable age ($F(1.79, 150) = 7.29, p = .001, \eta_p^2 = .05$). After visual analyzing this result, we found that younger individuals performed worse in differentiating between the three levels.

Concerning the factor congruency, we found no significant main effect, $F(1, 150) = 0.72, p = .39, \eta_p^2 = .005$, which indicated that, when holding profile ambiguity constant, there was no significant difference between congruent or incongruent profiles other than motivated reasoning would suggest. In turn, this implied that hypothesis H1a and H1b had to be rejected. However, when scrutinizing the results of the control variables, we noted that three variables interacted significantly with the factor congruency: social bot knowledge ($F(1, 150) = 5.52, p = .02, \eta_p^2 = .04$), age ($F(1, 150) = 13.29, p < .001, \eta_p^2 = .09$), and time spend on social media ($F(1, 150) = 4.51, p = .03, \eta_p^2 = .03$). Through visual analysis, we assessed what this interaction meant for each level of ambiguity. For all levels (human, social bot, and ambiguous profiles), we found that individuals who knew about social bots showed a greater effect of motivated reasoning. In other words, people with prior knowledge of social bots rated opinion-congruent profiles as more human

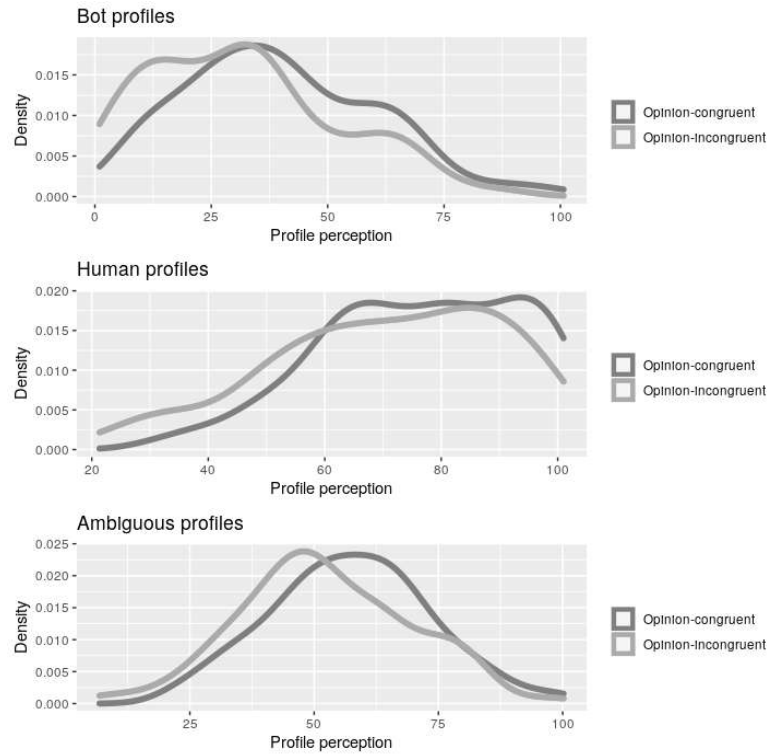


Fig. 2. Density distribution of the dependent variable *profile perception*. If a Twitter profile was perceived as a social bot, values are below 50. If the Twitter profile was perceived as a human profile, values are above 50. Red lines indicate opinion-congruent profiles, whereas blue lines indicate opinion-incongruent profiles.

and opinion-incongruent profiles as more bot-like compared to people who did not know about social bots before. Similarly, we found that the more time people spend on social media the greater the motivated reasoning effect. In other words, people who spend more time on social media rated opinion-congruent profiles as more human and opinion-incongruent profiles as more bot-like compared to people who spend less time on social media. For age, the results were less clear. Younger individuals showed more motivated reasoning only for human profiles. For bot profiles as well as ambiguous profiles age did not show a substantial effect.

Moreover, we did not find a significant interaction between congruency and ambiguity, $F(1.86, 150) = 1.46$, $p = .23$, $\eta_p^2 = .01$ (using Greenhouse-Geisser correction), including the before mentioned control variables. This meant that our overall assumption that opinion-congruency would influence profile perceptions only for ambiguous profiles was not met (the combination of H1 and H2).

4.3 Planned contrasts

Through paired sample t-test, we found a significant difference between the means of ambiguous, opinion-congruent ($M = 57.4$) and ambiguous, opinion-incongruent ($M = 52.4$) accounts, $t(150) = 3.48$, $p = .001$, $d = 0.28$. Hence, the result for ambiguous opinion-congruent accounts supported H1a which suggested that these accounts would be perceived as

human accounts (i.e. $M > 50$). H1b was, however, not supported. We hypothesized that ambiguous opinion-incongruent accounts would be perceived as social bot profiles (i.e. $M < 50$). This was not the case ($M = 52.4$). Concludingly, we found that opinion-congruent accounts were perceived as *relatively more human* than opinion-incongruent accounts, supporting the motivated reasoning hypothesis.

To test H2a and H2b, we performed a paired sample t-test for the respective hypothesis, adjusting for multiple comparisons through Bonferroni correction. We found neither support for H2a nor H2b. Means of opinion-congruent ($M = 39.7$) and opinion-incongruent ($M = 31.6$) social bots accounts were significantly different, $t(150) = 4.72$, $p < .001$, $d = .38$. Similarly, means of opinion-congruent ($M = 76.6$) and opinion-incongruent ($M = 70.4$) human accounts were significantly different as well, $t(150) = 3.77$, $p < .001$, $d = .3$. In both cases, opinion-congruent accounts were perceived as relatively more human, similarly to ambiguous accounts. However, we hypothesized that accounts that are clearly identifiable as human or bot should be perceived as such, independent of the displayed opinion in the account. We conclude that the effect of motivated reasoning is stronger than we originally anticipated as we found motivated reasoning even for accounts that were previously identified as clearly human and social bot.

4.4 Mediation analysis for credibility

We conducted overall three mediation analysis, representing the three levels of ambiguity and following the procedure described in the strategy of analysis. The outcome variable for each analysis was the perception of the Twitter profiles, whereas the two levels of congruency were the predictor variable. The mediator variable was the credibility rating of the respective level of ambiguity.

We tested the significance of this effect using bootstrapping procedures, computing 5000 bootstrapped samples with a confidence interval of 95%. The unstandardized indirect effect coefficient of credibility of congruent and incongruent social bot profiles was -7.59 with a 95% confidence interval ranging from -10.94 to -4.48 (see also Figure 3a). This supported our hypothesis that the perception of bot profiles is mediated by perceived credibility of the social bot profile. Similarly, we found a significant indirect effect of credibility of congruent and incongruent human profiles of -8.27 with a 95% confidence interval ranging from -11.12 to -5.76 (see Figure 3b) as well as a significant indirect effect of credibility of congruent and incongruent ambiguous profiles of -9.77 with a 95% confidence interval ranging from -12.64 to -7.11 (see Figure 3c). Overall, this results show that opinion-congruent profiles were perceived as more credible which lead users to rate them as more human, whereas profiles that were opinion-incongruent were perceived as less credible which lead users to rate them as more bot-like.

5 DISCUSSION

In this study, we wanted to know as to when users perceive an account to be a social bot or an human. Following motivated reasoning theory, we hypothesized that Twitter profiles which are opinion-congruent would be more likely to be identified as human accounts (H1a), whereas opinion-incongruent accounts would be more likely to be identified as social bot accounts (H1b). We also suggested that this would only occur in cases where users are unable to differentiate between social bots and human accounts (ambiguous accounts). In other words, if profiles clearly represent a human (H2a) or a social bot (H2b) account, opinion-congruency has no effect on how the profile is perceived. Additionally, we predicted that the relationship between opinion-congruency and profiles perception is mediated by perceived profile credibility.

Our results indicated that participants perceived the levels of Twitter profiles namely, human, social bot and ambiguous profiles differently (a significant main effect for the factor ambiguity) as can also be seen in Figure 3. The results of

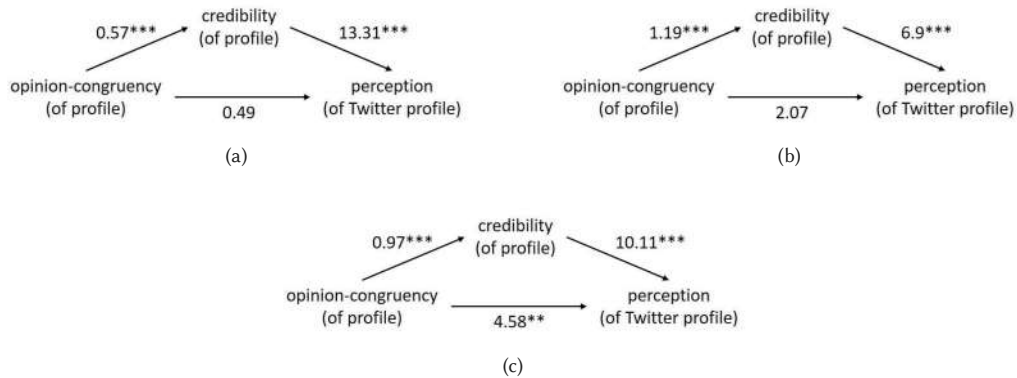


Fig. 3. Mediation analyses with relevant path coefficients for (a) bot profiles, (b) human profiles, and (c) ambiguous profiles. Significant path coefficients are marked as followed: *** $p < .001$, ** $p < .01$, * $p < .05$.

paired sample t-tests also supported this. Consequently, we observe that participants could generally differentiate between social bot and human accounts.

Moreover, we find that younger individuals differentiated less between the three levels of Twitter profiles as compared to older individuals, independent of the levels of opinion-congruency. We suggest that the increased exposure of younger individuals to social media (see significant correlations of age with time spend on social media and social media platforms used in Table 2) makes them more suspicious of accounts. This is why they are unable to ascertain the identity of any account (human or social bot) with greater certainty. However, to understand why participants perceived a particular account as social bot or human can only be answered through in-depths qualitative studies similar to [3].

5.1 The role of opinion-congruency when perceiving Twitter accounts

In relation to our central hypothesis, suggesting that profiles are perceived in a biased manner, we found only partial evidence to support this claim. When including control variables, we did not find a main effect for opinion-congruency. This indicates that opinion-congruent and opinion-incongruent profiles were perceived similarly. This result rejects our hypothesis that motivated reasoning guides profile perception. However, we found a motivated reasoning effect for the control variables *age*, *time spend on social media* and *social bot knowledge*. We found that, especially, participants with prior knowledge of social bots as well as participants that spend more time on social media showed a greater bias. In line with motivated reasoning, they perceived opinion-congruent profiles as more human and opinion-incongruent profiles as more bot-like.

We propose that these results stem from a different usage of the term social bot. We agree that participants with prior knowledge of social bots and participants who spend more time on social media might apply the term social bot as a pejorative term to indicate disagreement and discredit accounts in comparison to less well versed users. We believe that the effect of motivated reasoning is, in turn, amplified by wanting to show disagreement by labeling accounts as social bots (expressive disagreement). Popular media have discussed this labeling of others as social bots to discredit them before. In 2019, Sašo Ornik discussed in a Medium article how the political opposition in Slovenia, as well as the USA, have been demonized and degraded by associating them with social bots or an evil mastermind ².

²<https://medium.com/@saornik/everybody-i-dont-agree-with-is-a-russian-bot-or-how-it-is-easier-to-believe-an-evil-mastermind-ca02391055cb>

Moreover, when excluding the control variables from our analysis through paired-sampled t-test, we found that mean profile perceptions of opinion-congruent profiles were, as predicted, rated relatively more human (H1a) whereas opinion-incongruent profiles were rated as relatively more bot-like (H1b). This was a constant pattern overall profiles, independent of the levels of ambiguity. This means that a motivated reasoning effect could only be found when not controlling for individual differences such as age, time spend on social media or knowledge of social bots.

Our results also reveal the expected mediation of congruency and profile perceptions through profile credibility. This means that perceived profile credibility contributes significantly to the effect of opinion-congruency on the profile perception whereby opinion-congruent profiles were perceived as more credible than opinion-incongruent profiles.

5.2 Practical implications and future directions

Our results can be used to inform developers of assisting tools for social bot detection as well as media literacy scholars. For assisting tools like Botometer, which rely on users to self-identify accounts they want to check, our results suggest that users might enter simply profiles they disagree with, expecting it to be a social bot, due to motivated reasoning. In turn, the Botometer result might become less credible because it does not confirm one's expectation. Future studies should take into account the users' bias when investigating the usability and resulting credibility of assisting tools. For example, it could be asked whether accounts that are incongruent to one's own opinion are more likely to be tested of being a social bot? Moreover, are results of assisting tools less credible to the users, if they do not confirm the expected outcome?

Because motivated reasoning influences individuals to become overconfident as well as overcritical of what they perceive on social media, we suggest the inclusion of motivated reasoning and its effects in media literacy programs. This has been proposed before, for example, Lenker [30] argued that individuals need to be educated about the detrimental effects of motivated reasoning as well as to reflect upon one's own motivated reasoning when processing information. Future studies could investigate whether educating users about motivated reasoning results into less biased perception.

5.3 Limitations

Participants were only shown Twitter profiles. Limitations that come with this are twofold: First, our results are platform dependent. Although we asked whether participants actively used Twitter and controlled for this factor, we can neither make claims about specific other social media platforms, nor can we assess how generalizable our results are. Second, users were presented full Twitter profiles of only ten posts. Actual Twitter profile go beyond ten posts and offer more information which can help better identify an account. Third, users are most likely to come across social bots on their Twitter timeline, where they can only view one post. However, one post is less indicative about an accounts authenticity than an account profile.

Lastly, the social bot accounts which we used in this study were artificially created. Although similar to social bots on Twitter as well as pre-tested, they represent only a limited type of social bots. The cues that we used to build the profiles do not represent all possible cues and combination of cues. For example, we did not manipulate follower/followee ratio or profiles pictures, both of which are commonly used cues to assess whether an account is automated or not. Future studies would profit from assessing users' perception among a variety of different social media platforms, within different contexts (timeline versus profiles) and with different cues associated with social bots.

5.4 Conclusion

In this study, we investigated as to when users perceive an account to be a social bot or a human. In line with motivated reasoning, we hypothesized that opinion-congruent accounts are perceived as more human and opinion-incongruent accounts as more bot-like. Our main finding supports this hypothesis: users' ability to differentiate human and social bot accounts is affected by their prior-opinion. We conclude that participants who know about social bots and participants who spend more time on social media labeled profiles as social bots to express disagreement and discredit the profile. Moreover, we assert that opinion-congruent profiles were perceived as more credible than opinion-incongruent profiles.

REFERENCES

- [1] Jon-Patrick Allem, Emilio Ferrara, Sree Priyanka Uppu, Tess Boley Cruz, and Jennifer B Unger. 2017. E-Cigarette surveillance with social media data: Social bots, emerging topics, and trends. *JMIR Public Health and Surveillance* 3, 4 (2017), e98. <https://doi.org/10.2196/publichealth.8641>
- [2] Eiman Alothali, Nazar Zaki, Elfadil A. Mohamed, and Hany Alashwal. 2019. Detecting Social Bots on Twitter: A Literature Review. *Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018* (2019), 175–180. <https://doi.org/10.1109/INNOVATIONS.2018.8605995>
- [3] Anonymized Authors. Submitted to CHI2021. Spot the Bot: User Knowledge of Social Bots and Proficiency in their Detection. (Submitted to CHI2021).
- [4] Marco T. Bastos and Dan Mercea. 2019. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review* 37, 1 (2019), 38–54. <https://doi.org/10.1177/0894439317734157>
- [5] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21, 11 (2016), 1–17. <https://doi.org/10.5210/fm.v21i11.7090>
- [6] Martin Bisgaard. 2015. Bias will find a way: Economic perceptions, attributions of blame, and partisan-motivated reasoning during crisis. *The Journal of Politics* 77, 3 (2015), 849–860. <https://doi.org/10.1086/681591>
- [7] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.
- [8] David A. Broniatowski, Amelia M. Jamison, Si Hua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health* 108, 10 (2018), 1378–1384. <https://doi.org/10.2105/AJPH.2018.304567>
- [9] C Cai, L Li, and D Zengi. 2017. Behavior enhanced deep bot detection in social media. In *IEEE International Conference on Intelligence and Security Informatics*. 128–130.
- [10] Shelly Chaiken, Roger Giner-Sorolla, and Serena Chen. 1996. Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In *The Psychology of action: Linking cognition and motivation to behavior*, P. M. Gollwitzer and J. A. Bargh (Eds.). Guilford Press, New York, NY, 553–578.
- [11] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. Who is tweeting on twitter: Human, bot, or cyborg? *Proceedings - Annual Computer Security Applications Conference, ACSAC* (2010), 21–30. <https://doi.org/10.1145/1920261.1920265>
- [12] Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot: A System to Evaluate Social Bots. (2016), 4–5. <https://doi.org/10.1145/2872518.2889302> arXiv:1602.00975
- [13] James N. Druckman, Thomas J. Leeper, and Rune Slothuus. 2016. Motivated responses to political communications: Framing, party cues, and science information. *The Feeling, Thinking Citizens* (2016), 125–150. <https://doi.org/10.4324/9781351215947>
- [14] Ullrich K.H. Ecker and Li Chang Ang. 2019. Corrections, political attitudes and the processing of misinformation. *Political Psychology* 40, 2 (2019), 241–260. <https://doi.org/10.16309/j.cnki.issn.1007-1776.2003.03.004>
- [15] Chad Edwards, Autumn Edwards, Patric R. Spence, and Ashleigh K. Shelton. 2014. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior* 33 (2014), 372–376. <https://doi.org/10.1016/j.chb.2013.08.013>
- [16] Richard M. Everett, Jason R. C. Nurse, and Arnau Erola. 2016. The anatomy of online deception: What makes automated text convincing? *Proceedings of the 31st Annual ACM symposium on Applied Computing* (2016), 1115–1120. <https://doi.org/10.1145/2851613.2851813>
- [17] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday* 22, 8 (2017).
- [18] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. 2017. Social Bots: Human-Like by Means of Human Control? *Big Data* 5, 4 (2017), 279–293. <https://doi.org/10.1089/big.2017.0044> arXiv:1706.07624
- [19] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- [20] Hye-hyun Hong and Hyun Jee Oh. 2020. Utilizing bots for sustainable news business: Understanding users' perspectives of news bots in the age of social media. *Sustainability* 12, 16 (2020), 6515. <https://doi.org/10.3390/su12166515>
- [21] Philip N. Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum: Research note 2016.1. *Oxford: Computational Propagand Project, 2016* (2016).

- [22] Leonie Huddy, Lilliana Mason, and Lene Aarøe. 2015. Expressive partisanship: campaign involvement, political emotion, and partisan identity. *American Political Science Review* 109, 1 (2015), 1–17. <https://doi.org/10.1017/S0003055414000604>
- [23] Andreas Jungherr. 2015. Twitter as Political Communication Space: Publics, Prominent Users, and Politicians. In *Analyzing Political Communication with Digital Trace Data. Contributions to Political Science*. Springer, Cham. https://doi.org/10.1007/978-3-319-20319-5_4
- [24] Byungkyu Kang, Tobias Höllerer, and John O Donovan. 2015. Believe it or not? Analyzing information credibility in microblogs. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (2015), 611–616.
- [25] Minjeong Kang. 2010. Measuring Social Media Credibility: A Study on a Measure of Blog Credibility. *Institute for Public Relations* (2010), 59–68. <https://doi.org/10.1136/bmj.g5133>
- [26] Arzum Karataş and Serap Şahin. 2017. A Review on Social Bot Detection Techniques and Research Directions. *Proc. Int. Security and Cryptology Conferencme Turkey i* (2017), 156–161.
- [27] Franziska B Keller, David Schoch, Sebastian Stier, and Junghwan Yang. 2020. Political Astroturfing on Twitter : How to Coordinate a Disinformation Campaign Political Astroturfing on Twitter : How to Coordinate a Disinformation Campaign. *Political Communication* 37, 2 (2020), 256–280. <https://doi.org/10.1080/10584609.2019.1661888>
- [28] Mustafa Koc and Esra Barut. 2016. Computers in Human Behavior Development and validation of New Media Literacy Scale (NMLS) for university students. *Computers in Human Behavior* 63 (2016), 834–843. <https://doi.org/10.1016/j.chb.2016.06.035>
- [29] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological Bulletin* 108, 3 (1990), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- [30] Mark Lenker. 2016. Motivated reasoning, political information, and information literacy education. *portal: Libraries and the Academy* 16, 3 (2016), 511–528. <https://doi.org/10.1353/pla.2016.0030>
- [31] James C McCroskey and Jason J Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* 66, 1 (1999), 90–103. <https://doi.org/10.1080/03637759909376464>
- [32] Judith Meinert, Ahmet Aker, and Nicole C Krämer. 2019. The impact of Twitter features on credibility ratings - An explorative examination combining psychological measurements and feature based selection methods. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, Vol. 6. 2600–2609.
- [33] Hugo Mercier. 2016. Confirmation bias - myside bias. *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory* (2016), 99–114. <https://doi.org/10.4324/9781315696935>
- [34] Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of Communication* 60, 3 (2010), 413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- [35] Alessandro Murgia, Daan Janssens, Serge Demeyer, and Bogdan Vasilescu. 2016. Among the machines: Human-bot interaction on social Q&A websites. *Conference on Human Factors in Computing Systems - Proceedings 07-12-May-* (2016), 1272–1279. <https://doi.org/10.1145/2851581.2892311>
- [36] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.
- [37] Adrian Rauchfleisch and Jonas Kaiser. 2020. The False Positive Problem of Automatic Bot Detection in Social Science Research. *SSRN Electronic Journal* 7641 (2020). <https://doi.org/10.2139/ssrn.3565233>
- [38] David P. Redlawsk. 2002. Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making author(s). *The Journal of Politics* 64, 4 (2002), 1021–1044. <https://doi.org/10.1111/1468-2508.00161>
- [39] B. Reeves and C.I. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, New York.
- [40] Fabian Schäfer, Stefan Evert, and Philipp Heinrich. 2017. Japan’s 2014 general election: Political bots, right-wing internet activism, and Prime Minister Shinz o Abe’s hidden nationalist agenda. *Big Data* 5, 4 (2017), 294–309. <https://doi.org/10.1089/big.2017.0049>
- [41] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669. <https://doi.org/10.1073/pnas.1805871115>
- [42] Galen Stocking and Nami Sumida. 2018. Social media bots draw public’s attention and concern. *Pew Research Center* October (2018).
- [43] Charles S Taber and Milton Lodge. 2016. The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Political Psychology* 37 (2016), 61–85. <https://doi.org/10.1111/pops.12321>
- [44] J. J. Van Bavel and A. Pereira. 2018. The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences* 22, 3 (2018), 213–224.
- [45] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (2017), 280–289. arXiv:arXiv:1703.03107v2
- [46] Patrick Wang, Rafael Angarita, and Ilaria Renna. 2018. Is this the era of misinformation yet? Combining social bots and fake news to deceive the masses. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (2018), 1557–1561. <https://doi.org/10.1145/3184558.3191610>
- [47] Stephan Winter, Miriam J. Metzger, and Andrew J. Flanagin. 2016. Selective use of news cues: A multiple-motive perspective on information selection in social media environments. *Journal of Communication* 66, 4 (2016), 669–693. <https://doi.org/10.1111/jcom.12241>
- [48] Kai Cheng Yang, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emergent Technologies* 2018 (2019), 48–61. <https://doi.org/10.1002/hbe2.115>

ARTICLE 3

The following article is reused from:

Wischnewski, M., Ngo, T., Bernemann, R., Jansen, M., & Krämer, N. (2022). “I agree with you, bot!” How users (dis)engage with social bots on Twitter. *Manuscript under review for publication in the journal New Media & Society*, online first. <https://doi.org/10.1177/14614448211072307>



“I agree with you, bot!” How users (dis)engage with social bots on Twitter

new media & society
1–22

© The Author(s) 2022



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448211072307
journals.sagepub.com/home/nms



**Magdalena Wischnewski^{ID}, Thao Ngo,
Rebecca Bernemann, Martin Jansen
and Nicole Krämer**

University of Duisburg-Essen, Germany

Abstract

This article investigates under which conditions users on Twitter engage with or react to social bots. Based on insights from human–computer interaction and motivated reasoning, we hypothesize that (1) users are more likely to engage with human-like social bot accounts and (2) users are more likely to engage with social bots which promote content congruent to the user’s partisanship. In a preregistered 3×2 within-subject experiment, we asked $N=223$ US Americans to indicate whether they would engage with or react to different Twitter accounts. Accounts systematically varied in their displayed *humanness* (low humanness, medium humanness, and high humanness) and *partisanship* (congruent and incongruent). In line with our hypotheses, we found that the more human-like accounts are, the greater is the likelihood that users would engage with or react to them. However, this was only true for accounts that shared the same partisanship as the user.

Keywords

Human–computer interaction, motivated reasoning, social bot, social influence, social media, Twitter

Introduction

Communication on social media platforms is not always generated by human users. So-called *social bots*, automated accounts run by computer algorithms, increasingly

Corresponding author:

Magdalena Wischnewski, Research Training Group “User-Centred Social Media” University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany.

Email: magdalena.wischnewski@uni-due.de

populate social networking sites. While automated scripts have been used for some time, these social bots are design to interact with human users and to shape “the aggregate social behavior and patterns of relationships between groups of users online” (Hwang et al., 2012: 40). Being designed as such socio-technical entities, it has been argued that social bots go beyond previous functionalities (e.g. scripts that do undesirable labor) but are built to resemble a *social Self*, opening new frontiers of robo-sociality (Bakardjieva, 2015; Gehl and Bakardjieva, 2017).

Because previous research connects some of these social bots to a plethora of malicious activities such as the promotion of misinformation (Wang et al., 2018), political astroturfing (Keller et al., 2020), and the influence of election outcomes (Bessi and Ferrara, 2016; Ferrara, 2017; Schäfer et al., 2017), it is important to scrutinize the mechanisms of their effects. In this study, we are interested in such malicious social bots which try to disguise their automated nature by blending with human online activities (e.g. “liking” or “retweeting”).

While previous work has already investigated the effects of bots on social networks (Cheng et al., 2020; Keijzer and Mäs, 2021; Ross et al., 2019), in this article we offer a psychological perspective on how deceitful and manipulative social bots and social media users engage with each other through following, retweeting, quoted retweeting, and commenting on Twitter. We argue that investigating such engagement with social bots is crucial to better understand how social bot communication affects not only social media networks but also the user’s contribution to the amplification of malicious social bot communication.

To this end, we experimentally investigate two factors: the *humanness of an account* and the displayed *partisanship of an account*. We base our assumptions on previous findings which indicate that users rate human communication as more attractive than social bot communication (Edwards et al., 2014), as well as on insights from partisan-motivated reasoning (e.g. Bolsen et al., 2014) which suggests that users process information in a biased manner, favoring opinion-congruent information over opinion-incongruent information. Hence, the central research question of our study is the following:

RQ1. How does the humanness and partisanship of a social bot account influence users’ willingness to follow, comment, retweet, and share a quoted retweet of a social bot account?

Based on previous results on the interplay of humanness and partisanship (Wischnewski et al., 2021; Yan et al., 2020), we assume that the effects of humanness and partisanship influence each other. Opinion congruency might be more pronounced for highly human-like accounts, whereas opinion congruency might matter less when accounts are less human-like. Moreover, we want to know *why* participants chose for or against following, retweeting, quoted retweeting, and commenting. Hence, in this article, we are also interested in the motivations, that is, possible reasons for users to engage with and react to social bots. In two subsequent research questions, we ask,

RQ2. Does the humanness of an account interact with the congruency of the displayed account partisanship and users’ partisanship?

RQ3. Which engagement motivations drive users' engagement activities and are these engagement motivations rather dependent on the humanness or partisanship of an account?

Investigating how users engage with varying degrees of social bot humanness and partisanship, we contribute to a better understanding of the assumed effects of social bots on online communication networks. In particular, the results of our study help to gain a deeper understanding of the users' contribution to the amplification of malicious social bot content. Our results can also help to design countermeasures and inform policymakers and social media providers alike.

Theoretical background

To understand the effects of social bots on users and, ultimately, on society, different strategies have been pursued, for example, modeling approaches that simulate the (hypothetical) impact of social bots in social networks through (multi)agent-based modeling (Cheng et al., 2020; Keijzer and Mäs, 2021; Ross et al., 2019) or epidemiological models of contagion (Mønsted et al., 2017). For example, employing an agent-based model of the spiral of silence, Ross et al. (2019) found that, under certain circumstances, social bots can alter the opinion climate of a communication network. Another approach is to investigate how social bots spread information in online networks (Gorodnichenko et al., 2018; Salge et al., 2021). For example, Salge et al. (2021) observed information dissemination by social bots in the 2013 Brazilian Confederation Cup riots. Relying on conduit brokerage, the authors derive a theoretical model of information dissemination, incorporating an algorithmic process of social alertness and social transmission. In doing so, the authors thoroughly explain the complex processes that constitute information dissemination by social bots.

While these approaches provide meaningful conclusions on how deceitful and manipulative social bots can affect social networking platforms as a whole, they do not offer insights into how individual social media users might (unwillingly) promote social bot accounts. Moreover, knowing how individual users perceive and engage with social bots is crucial to inform modeling approaches. For example, Ross et al. (2019) employed two agent-based models with varying degrees of social bot influence on individual users. The authors found that, depending on the social bot influence on individual users, group-level effects varied: Bots with low influence on users were more effective in sparse networks, whereas bots with high influence on users were more effective in dense networks. Similarly, Salge et al. (2021: 4) were interested in the study of "bots taking action to disseminate information, regardless of whether the information they disseminate is received or not by the actors."

In this study, we want to know how much direct influence social bots exert on individual users to overcome this limitation. While the direct persuasive effect of social bots on individual users' opinions is difficult to assess, we argue to investigate, in a first step, how social bots and users engage with each other. Such user engagement can be investigated from different perspectives and varying depth. In this study, we limit engagement, however, to selected actions that individual users can take on Twitter which are

commenting, following, retweeting, and quoted retweeting. With engagement as a first proxy to assess a possible social bot influence, we follow the assumption “that the interaction with other individuals (or a group) [here: social bots] may affect or change subjects’ thoughts, feelings, or behavior” (Luceri et al., 2019:2). Empirical evidence supports this view. For example, investigating political persuasion on social media, Diehl et al. (2016) found that, besides news use, social interactions of users with each other positively affected attitude formations.

While not many studies have experimentally investigated under which circumstances users engage with social bots (see, for example, Edwards et al., 2014; Spence et al., 2018), previous observational evidence confirms that users, indeed, engage with social bots (Cardoso et al., 2019; Wagner et al., 2012). Trying to identify which users are more susceptible to engage with social bots, Wagner et al. (2012), for example, found that more interactive users were more likely to reply to and befriend social bots. Similarly, Wald et al. (2013) found that users’ Klout score¹ and the total number of followers predicted the likelihood of users replying to or following a social bot. Results by Cardoso et al. (2019) were especially alarming. The authors could show that “[o]ne in three posts reshared by humans is an original content created by bots” (Cardoso et al., 2019: 2).

Although these results indicate that users frequently engage with social bots, other findings suggest that this engagement is less likely, if the users suspect an account to be a social bot. For example, Edwards et al. (2014) found that, while users rated social bots as similarly credible and competent compared with human accounts, social bots were perceived as less social and task attractive. In another study by Edwards et al. (2015), the authors found that users display higher uncertainty, less liking, and less social presence when communicating with a social bot than a human user.

Considering these results, we suggest that users show an engagement preference for human-like social bot accounts over accounts perceived as social bots. Central to this claim is our hypothesis that the *perception* of an account determines how users would engage with an account. The perception, however, might deviate from the actual nature of an account. For example, highly technically sophisticated social bots might appear human, whereas accounts run by human users might appear to be social bots. It is also likely that some accounts are perceived as somewhat ambiguous, neither clearly automated nor clearly *not* automated. Hence, we argue that whether an account is truly a social bot or a human user is less relevant for our approach. Instead, our focus is on the effect of *users’ perceptions of humanness*. In doing so, we investigate the effects of different levels of humanness. Relaying this to the findings cited above, we assume that

H1. Users prefer to engage (commenting, following, retweeting, and quoted retweeting) with clearly human-like accounts over medium human-like accounts (ambiguous—neither clearly human nor social bot) over low human-like accounts (social bot-like).

Because engagement is not limited to users initiating engagement, we also investigate how users react when accounts initiate engagement. Similar to H1, we assume that

users are less likely to react to low human-like accounts and are more likely to react to high human-like accounts:

H2. Users prefer to react to high human-like accounts over medium human-like accounts (neither clearly human nor clearly social bot) over low human-like accounts (social bot-like).

Influence of users' and profiles' partisanship on engagement activities

Besides the humanness of accounts (high human-likeness vs medium human-likeness vs low human-likeness), we suggest that specific opinions expressed by accounts increase or decrease the likelihood of users engaging with the account. Especially findings in the context of political communication and political information processing could show that opinion-congruent information is more likely to lead to engagement on social media than opinion-incongruent information (Colleoni et al., 2014; Garz et al., 2020). Such effects have led researchers to assume the emergence of so-called echo chambers and filter bubbles within social media, suggesting that users become encapsulated only with like-minded views, reinforcing existing beliefs (Barberá et al., 2015; Colleoni et al., 2014). However, this notion has recently been challenged (Bruns, 2019) and refined Kitchens et al. (2020).

The psychological mechanisms driving this preferential behavior have been associated with motivated reasoning. Motivated reasoning generally proposes that information processing is driven by either accuracy or directional goals (Kunda, 1990). While the motive of accuracy goals is to arrive at an accurate conclusion, directional goals aim to arrive at predefined conclusions that support previous attitudes or identities. In an early study, Kunda (1987) found that female heavy coffee consumers were less convinced about the harmful effects of caffeine than female low coffee consumers.

Motivated reasoning has also been found for engagement activities of users on social media (Cinelli et al., 2020), indicating that users were more likely to engage with like-minded users, leading to homophilic interaction patterns. We assume that this preference for engagement with like-minded users, rooted in motivated reasoning, should also occur when engaging with social bots. However, we assume a differentiated pattern, depending on the engagement activities. For the case of Twitter, activities such as *retweeting* or *following*² signal endorsement, which should predominantly be used for like-minded users/content. Especially retweeting has been associated not only with agreement in a message but also with trust in the author of a tweet (Metaxas et al., 2015). In contrast, by sharing a *quoted retweet* or a *comment*, users can similarly express endorsement and disagreement and we assume that motivated reasoning would not affect these activities. We hypothesize the following:

H3a. Users *follow* and *retweet* accounts in a biased manner, favoring opinion-congruent accounts over opinion-incongruent accounts.

H3b. Users *comment* on and share *quoted retweets* of opinion-congruent and opinion-incongruent accounts equally likely.

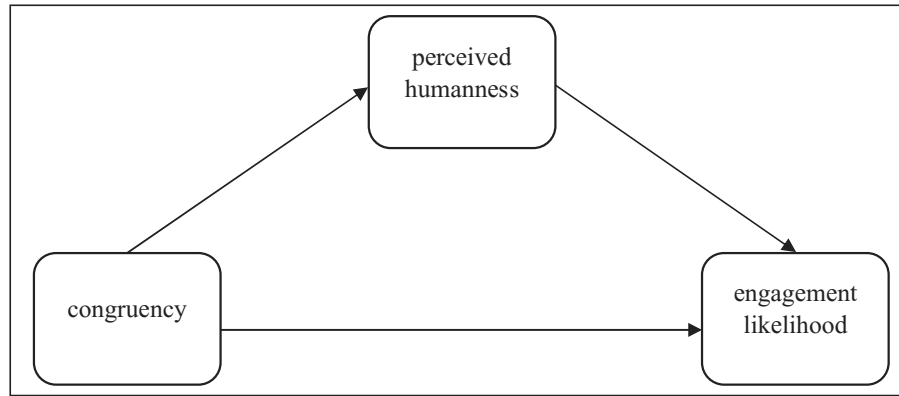


Figure 1. Mediation hypothesis (H5).

If a participant's partisanship and the displayed partisanship of a profile match (congruent condition), the participant is more likely to engage with the profile as compared with the non-matching condition. This effect can partly be explained by the effect of partisanship congruency on perceived humanness (see Wischniewski et al., 2021; Yan et al., 2020), where congruent profiles are perceived as more human-like than incongruent profiles.

Similar to H2, engagement can also be initiated by another account. In the same manner as H3a, we assume that motivated reasoning would affect this process so that users are less likely to react to opinion-incongruent accounts and more likely to react to opinion-congruent accounts:

H4. Users are more likely to react to accounts that are opinion-congruent than opinion-incongruent.

To answer RQ1, how does the perceived humanness and partisanship of a social media account influence users' willingness to engage with and react to it, we suggest so far that both perceived humanness of accounts (H1 & H2) and partisanship (H3 & H4) drive users' engagement with an account. However, previous studies have also found that motivated reasoning also affects how users perceive the humanness of profiles. Results indicate that users perceive opinion-congruent accounts as more human-like than opinion-incongruent accounts, which were perceived as less human-like and more bot-like (Wischniewski et al., 2021; Yan et al., 2020). Building on these results, we take our assumptions in H1–H4 one step further and suggest a mediating role of account perceptions. We hypothesize that

H5. The relationship between the users' partisanship and the likelihood to engage with an account is mediated by the perceived humanness of an account (see Figure 1).

In addition, humanness (H1/H2) and users' partisanship (H3/H4) might affect each other in a way that opinion-congruent preferences are more pronounced when accounts are perceived as more human-like, whereas opinion-congruency might matter less when accounts are perceived as less human-like. Independent of the accounts' displayed partisanship, users might generally be less likely to engage with accounts of low humanness. This would

suggest an important boundary condition of motivated reasoning, indicating that users do not blindly engage with any account on social media just because it shares the users' partisanship. Hence, we included RQ2 into our study: Does the perceived humanness of an account interact with the congruency of the displayed account partisanship and users' partisanship?

Finally, we want to know *why* participants chose for or against engagement with accounts. Hence, we are also interested in the motivations which drive the engagement process. We address this in a third research question (RQ3): Which engagement motivations drive users' engagement activities, and are these engagement motivations rather dependent on the perceived humanness or the partisanship of an account?

All hypotheses and research questions were preregistered prior to data collection and are publicly available via OSF (<https://osf.io/w42ca/>).

Method

The ethical committee of the University of Duisburg-Essen approved the study. The data set, stimulus material, analysis, and Supplementary Material are publicly available on OSF: <https://osf.io/w42ca/>.

Sample

To test our hypotheses, we recruited 223 US American Twitter users from the crowdsourcing platform Prolific. The sample size of 220 was determined through a prior power analysis (for details, see preregistration). Participants' age ranged from 18 to 75 ($M=30.75$, $SD=11$) years. A total of 115 identified as female, 96 as male, nine as non-binary, and three participants did not disclose their gender. Most participants held either a high school degree (75) or a university degree (BA=98; MA=28) and were White (139), Black or African American (25), Hispanic/Latino (25), or Asian American (24).

Experimental design and procedure

We conducted an online experiment, using a 3×2 within-subject design, with two independent factors, *humanness* and *congruency*. *Humanness* described the nature of a Twitter account and consisted of three levels: highly human-like accounts, medium human-like accounts (ambiguous), and low human-like accounts. The factor *congruency* consisted of two levels, opinion-congruent and opinion-incongruent, and referred to the agreement between the participants' partisanship and the partisanship displayed in the Twitter account. To determine the congruency, we measured the political partisanship of each participant (Democrat or Republican) and experimentally manipulated the expressed partisanship of the Twitter accounts in the stimulus material.

To ensure all participants had the same understanding of social bots, we provided a general definition of social bots³ before the experiment. After that, participants viewed 18 different Twitter profiles in a randomized order. For each profile, participants were asked (1) how likely they would engage (retweet, follow, quoted retweet, and comment) with the profile, (2) how they would react if the profile engaged with

them, (3) which motivations drove their engagement intentions, and (4) how automated they perceived the profile. Finally, the experiment asked about participants' basic demographic data, social media usage, time spent on social media, Twitter usage, political interest, and partisanship.

Stimulus material

Constituting the first factor, *humanness*, we manipulated Twitter profiles to appear either run by a human, ambiguous, or run by a social bot. We followed the procedure developed by Wischnewski et al. (2021), who manipulated profiles by varying different characteristics, including the timing of posting behavior (frequent and infrequent), content (retweet only, tweets only, shared links only, and mixed content), and the profile picture.

For the second factor, congruency, we manipulated the political partisanship expressed in each profile, representing either a Republican or a Democrat account. Each profiles' partisanship was matched with the participants' partisanship, resulting in the two factor levels: opinion-congruent (Democratic Twitter profile and self-identified Democrat/Republican Twitter profile and self-identified Republican) and opinion-incongruent (Democratic Twitter profile and self-identified Republican/Republican Twitter profile and self-identified Democrat).

For each of the overall six conditions, we created three Twitter profiles consisting of 10 posts per profile, resulting in 18 Twitter profiles. Each Democrat account had a matching Republican account, displaying similar features concerning their follower/followee ratio, posting timings, and posting behavior. An example of a low human-like Democratic profile and the corresponding low human-like Republican profile is shown in Figure 2.

Measures

After viewing each Twitter profile, we asked participants several questions concerning their intention to engage with the respective profile. First, we wanted to know how likely participants would engage with a profile. Hence, for each of the four engagement activities (retweet, follow, quoted retweet, and comment), we asked, "How likely is it that you would [activity] this account?" Answers were given on a 5-point-Likert-type item, ranging from 1 = *very unlikely* to 5 = *very likely*. Second, we wanted to know how participants would react if a profile engaged with them. Hence, for each of the four engagement activities (retweet, follow, quoted retweet, and comment), we asked participants to indicate whether they would (1) "engage in some way with the account (e.g. follow back, retweet, or comment)," (2) "block/report this account," or (3) "do nothing."

To assess how profile perceptions affect the relationship of opinion congruency and engagement intentions, we measured the *profile perception* on a continuous sliding scale from 0 to 100, with lower values indicating more bot-like perceptions and higher values a more human-like perception (see also, Wischnewski et al., 2021). Finally, we also wanted to know which motivations drove participants' engagement intentions. To this end, we collected previously found engagement motivations that have been shown to drive users' engagement online. Each motivation could then be answered on a 5-point-Likert-type scale, ranging from 1 = *completely disagree* to 5 = *completely agree*. A list of all engagement

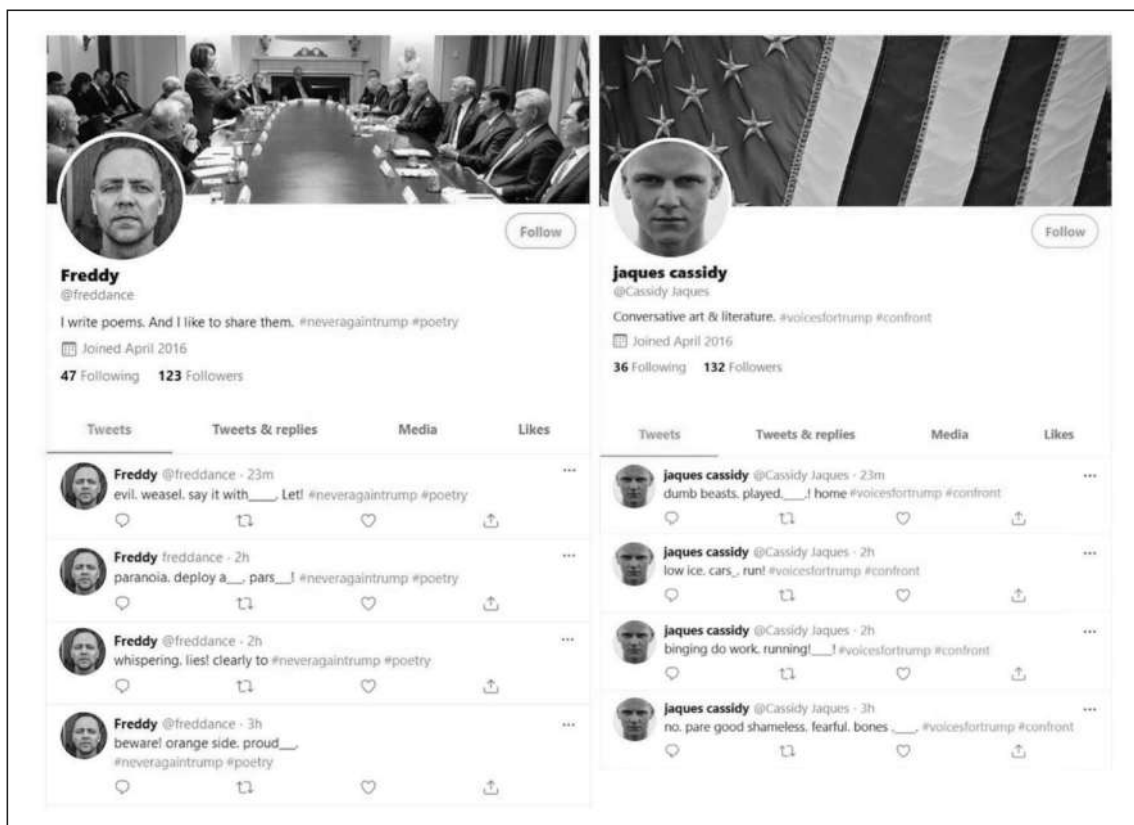


Figure 2. Examples of two low human-like profiles, depicting a Democrat profile (left) and a Republican profile (right).

motivations for each of the four engagement activities (retweet, follow, quoted retweet, and comment), including references, can be found in the Supplementary Material S1.

Control variables

Previous studies have shown that multiple variables, besides the variables of interest in our study, can affect how users perceive and engage with other users online. For example, Wischnewski et al. (2021) found that younger participants, participants who spent more time on social media and who know about social bots, showed a greater motivated reasoning bias. Hence, we included a one-item measure to assess how much time participants spent on social media (“less than 1 hour,” “1–3 hours,” “4–7 hours,” or “more than 8 hours”), and participant’s previous *social bot knowledge* (“How much do you think you know about ‘social bots’?” “a great deal/I am an expert on this topic,” “a lot/I have read quite a lot about them,” “some/I have some knowledge about them,” “a little/I have only heard of them,” or “none/I have no idea what they are”), following Yan et al. (2020).

In addition, we assessed participant’s *everyday engagement on social media* (“How often do you engage in the following activities [retweet, follow, quoted retweet, comment] on Twitter?” with answers on a 5-point-Likert-type scale, 1 = *never*, 5 = *always*; Cronbach’s $\alpha = .82$) and Twitter-related items from the *Twitter and Facebook Usage*

Scale by Hughes et al. (2012), resulting in two measures: Twitter for information sharing and searching (Cronbach's $\alpha = .8$) and Twitter for socializing (Cronbach's $\alpha = .86$). This allowed us to account for natural differences in engagement habits and accommodate findings by Wagner et al. (2012), who found that more engaged users were more susceptible to engage with social bots. Previous research could also identify that, for political context, political interest is an important moderator (e.g. Carrus et al., 2018), with higher levels of interest increasing partisan effects. Consequently, we included Shani's (2012) three-item measure of *political interest* (Cronbach's $\alpha = .94$).

Results

User-initiated engagement with social bots

The main interest of this study was the effects of humanness and partisanship on different engagement activities. Hence, we preregistered four repeated measures analyses of variance (ANOVAs), including predefined control variables, for all four different engagement intentions (following, retweeting, commenting, and quoted retweeting). As hypothesized, we found a significant main effect of humanness (H1) for following, $F(2, 428) = 61.84, p < .001, \eta_p^2 = .22$; retweeting, $F(2, 428) = 75.99, p < .001, \eta_p^2 = .26$; commenting, $F(2, 428) = 61.39, p < .001, \eta_p^2 = .22$; and quoted retweeting, $F(2, 428) = 76.63, p < .001, \eta_p^2 = .26$. All means and standard deviations are given in Table 1. Planned contrasts supported the results of the omnibus test, indicating that, for each engagement activity, participants were most willing to engage with highly human-like accounts over medium human-like accounts over low human-like accounts. Results of the planned contrasts can be found in Table S2 in the Supplementary Material.

Moreover, in H3a, we hypothesized that participants would show a motivated reasoning bias for the endorsement activities, following, and retweeting (main effect of congruency). Results of the repeated measures ANOVAs supported our hypothesis. We found a significant main effect of opinion-congruency for following, $F(1, 214) = 183.93, p < .001, \eta_p^2 = .46$, and retweeting, $F(1, 214) = 185.79, p < .001, \eta_p^2 = .47$. For the ambiguous engagement, commenting, and quoted retweeting, we did not expect an effect of opinion congruency (H3b). However, for both commenting, $F(1, 214) = 131.09, p < .001, \eta_p^2 = .38$, and quoted retweeting, $F(1, 214) = 174.53, p < .001, \eta_p^2 = .45$, the main effect of congruency was significant. This indicates that, for all engagement activities, participants preferred to engage with congruent profiles, instead of incongruent profiles, even for engagements that do not indicate endorsement, such as commenting and quoted retweeting.

In addition to the main effects of humanness and partisanship congruency, we wanted to know whether the effect of partisanship congruency was different for different levels of humanness (RQ2). Over all four engagement activities, we found that the effect of congruency was dependent on the level of humanness, following: $F(2, 428) = 72.2, p < .001, \eta_p^2 = .25$; retweeting: $F(2, 428) = 174.53, p < .001, \eta_p^2 = .45$; commenting: $F(2, 428) = 64.93, p < .001, \eta_p^2 = .23$; and quoted retweeting: $F(2, 428) = 66.75, p < .001, \eta_p^2 = .24$. When visually inspecting (see Figure 3) the mean engagement likelihoods, we found that the effect of congruency was much more pronounced for human-like accounts

Table 1. Mean engagement likelihoods and standard deviations for all four engagement activities.

Engagement	Human-likeness	Congruency	M	SD
Following	Low	Congruent	1.86	0.81
		Incongruent	1.39	0.67
	Medium	Congruent	2.05	0.95
		Incongruent	1.54	0.82
	High	Congruent	2.58	1.15
		Incongruent	1.44	0.84
Retweeting	Low	Congruent	1.90	0.80
		Incongruent	1.39	0.66
	Medium	Congruent	2.11	0.97
		Incongruent	1.44	0.76
	High	Congruent	2.62	1.21
		Incongruent	1.43	0.85
Commenting	Low	Congruent	1.82	0.85
		Incongruent	1.45	0.71
	Medium	Congruent	2.00	0.99
		Incongruent	1.57	0.81
	High	Congruent	2.46	1.18
		Incongruent	1.55	0.91
Quoted retweeting	Low	Congruent	1.84	0.81
		Incongruent	1.40	0.70
	Medium	Congruent	2.04	0.97
		Incongruent	1.46	0.76
	High	Congruent	2.55	1.19
		Incongruent	1.44	0.82

SD: standard deviation.

as compared with medium human-like, ambiguous accounts and low human-like accounts, indicating a boundary effect of motivated reasoning.

Reactions to social bot-initiated engagement

Similar to users initiating engagement with social bot accounts, social bot accounts can initiate engagement with users by following user accounts and commenting on, retweeting, or quoted retweeting user posts. Visually inspecting the descriptive outcomes of the four reactive engagement decisions (Figure 4), we observed several overall trends. Confirming the user-initiated engagement results (see the previous section), incongruent Twitter profiles received similar (dis-)engagement reactions, independent of the level of perceived humanness.

In contrast, reactions to congruent profiles were dependent on the level of humanness. Fewer participants indicated to report/block highly human-like accounts and more participants reported reacting to such accounts. Notably, participants likely just ignored

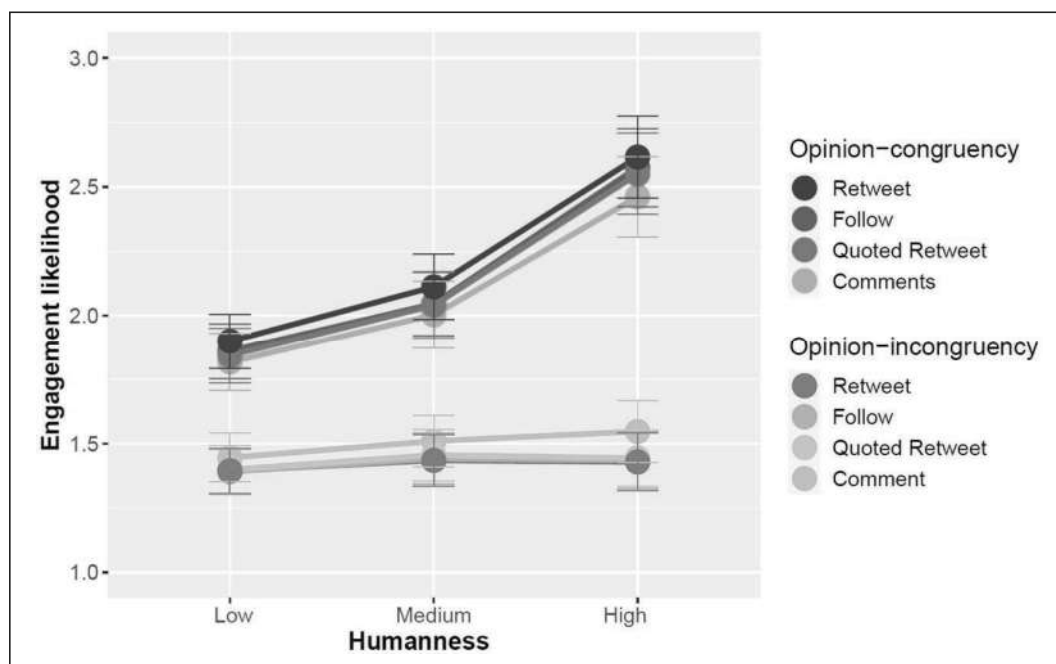


Figure 3. Mean engagement likelihoods for retweeting, following, quoted retweeting, and commenting.

Error bars depict 95% confidence intervals.

congruent accounts of low and medium humanness. For highly human-like accounts, the reaction depended on the initiating behavior. While following, retweeting, and quoted retweeting were still likely to be ignored, commenting human-like accounts are most likely to engage participants.

To confirm the visual analysis, we conducted mixed multinomial regressions. Because the visual analysis conveyed that participants were most inclined to ignore any engagement behavior, for the dependent variable, we used “nothing” as the baseline category, which we compared with the decision to “react” and “block/report.” For the factor congruency, we classified “congruent” and for the factor sophistication “ambiguous” at baseline. Control variables were included in each model. Coefficients and standard errors of all models can be found in Table 2.

Results of the mixed multinomial regression models are similar to all initiating activities and support the visual analysis. Compared with less human-like (ambiguous) Twitter accounts, less human-like accounts were more likely to be blocked/reported but equally (un)likely to be reacted to. Compared with less human-like (ambiguous) Twitter accounts, human-like accounts were more likely to be reacted to and less likely to be blocked/reported. Moreover, congruent profiles were more likely to be reacted to and less likely to be blocked than incongruent profiles.

Profile perception as the driver of engagement decisions

As suggested by motivated reasoning theory, we assumed that matching the participants’ partisanship and the displayed partisanship of the account (i.e. congruent) would drive

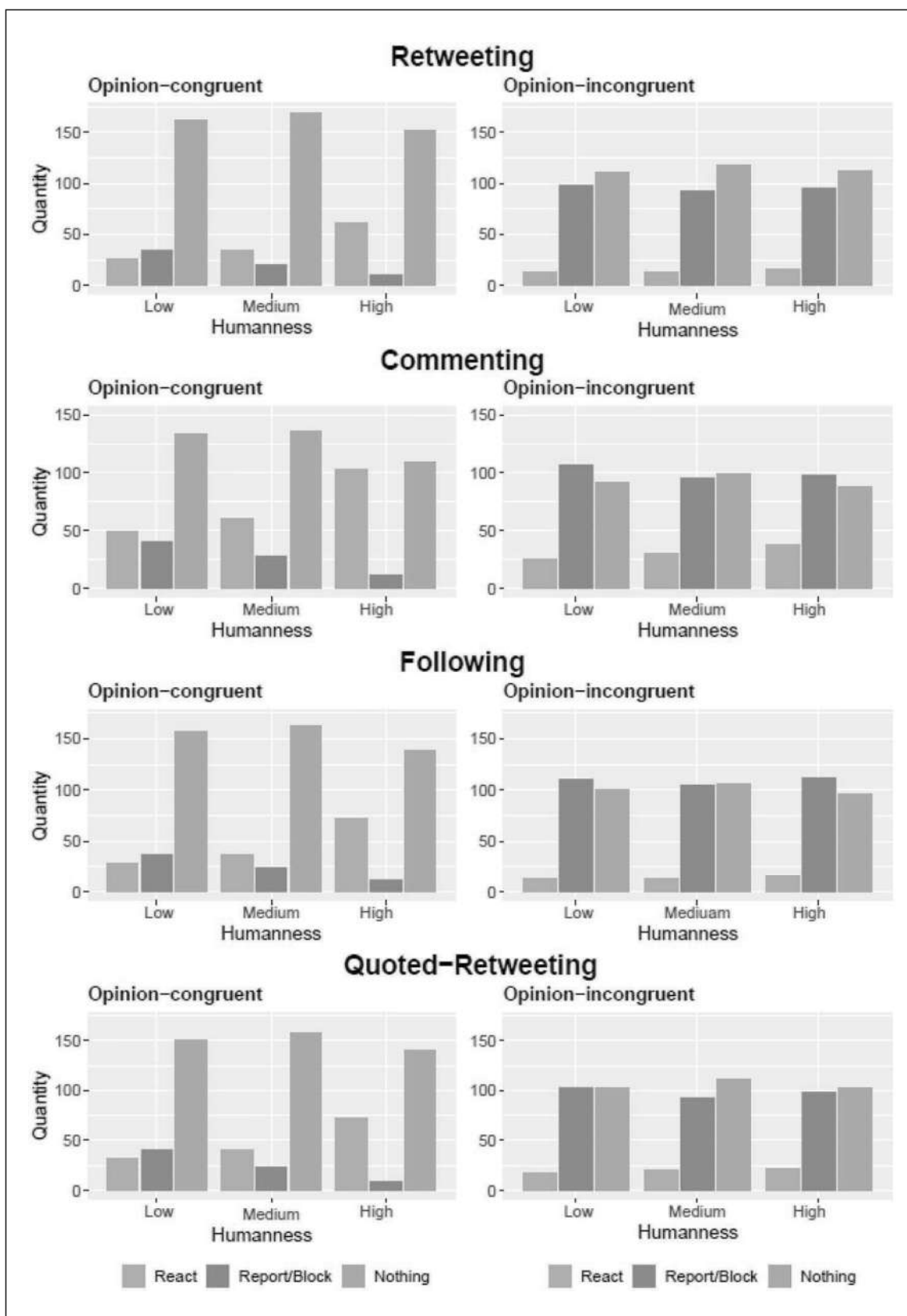


Figure 4. Count data of reaction decisions for each of the four behaviors (retweeting, commenting, following, and quoted retweeting).

Table 2. Results of the mixed multinomial regression analysis.

Engagement	Term	Low humanness		High humanness		Incongruent	
		Estimate	SE	Estimate	SE	Estimate	SE
Following	React	-0.23	0.16	0.88**	0.14	-0.66**	0.15
	Block/ report	0.45**	0.16	-0.56**	0.21	1.89**	0.21
Retweeting	React	-0.11	0.08	0.35**	0.07	-0.33**	0.10
	Block/ report	0.29**	0.08	-0.32**	0.12	0.94**	0.08
Commenting	React	-0.09	0.07	0.4**	0.06	-0.19*	0.07
	Block/ report	0.20*	0.08	-0.31**	0.11	0.79	0.07
Quoted retweeting	React	-0.11	0.08	0.36**	0.07	-0.19*	0.07
	Block/ report	0.29**	0.08	-0.41**	0.12	0.87**	0.09

SE: standard error.

* $p < .05$; ** $p < .01$; *** $p < .001$.

the decision to engage with this account. We found this hypothesis supported (see “User-initiated engagement with social bots” section). In addition, we wanted to know whether we can explain this effect through the participants’ perception of the account. Hence, we hypothesized that partisan congruency indirectly affected the engagement likelihood through a biased perception of the account (H5).

To test H5, we employed mediation analyses, using the PROCESS macro for SPSS by Hayes (2017). We initially preregistered mediation analyses only for the endorsement activities (following and retweeting) because we expected these activities to be affected by motivated reasoning. However, the analysis in section “User-initiated engagement with social bots” indicated that commenting and quoted retweeting were also affected by partisanship congruency. Hence, we also conducted mediation analyses for commenting and quoted retweeting, assuming the same mediating role of profile perception.

Overall, for engagement activities, the mediation analyses supported H5 only for low human-like and highly human-like accounts but not for less human-like (ambiguous) accounts. For both low and highly human-like accounts, incongruent profiles were less likely to be engaged with (significant negative c-paths). They were also perceived as less likely to be human (significant negative a-paths). In turn, being perceived as less human decreased the likelihood of engagement with an account (significant negative b-path). For both low human-like accounts and highly human-like accounts, including the perception of an account partially explained the effect of congruency on engagements (significant indirect effects). All path coefficients and confidence intervals are reported in Table S5 in the Supplementary Material.

Similar to low and highly human-like accounts, we found for less human-like (ambiguous) that congruency increased engagement likelihoods accounts and a more human-like profile perception (significant negative c’- and a-paths). However, we did not find

that the profile perception affected the likelihood of engagement (non-significant b-paths). Trying to understand this null effect, we first inspected the means and standard deviations of profile perceptions for ambiguous profiles: $M_{\text{congruent}} = 59.84$, $SD = 20.62$ and $M_{\text{incongruent}} = 57.83$, $SD = 21.33$. A paired-samples t test indicated that both means did not differ significantly, $t(222) = 1.43$, $p = .154$. Hence, we concluded that partisanship congruency did not affect how users perceived profiles with less human-likeness (ambiguous accounts).

Engagement motivations

As a follow-up to the findings above, we also wanted to know *why* users decided for or against engaging with an account (RQ3). Through four multiple regression analyses, we found which motivations drove users' decisions. Standardized regression coefficients and significance tests are summarized in Table S3 in the Supplementary Material.

To understand whether engagement motivations depended on the level of humanness of an account and/or the congruency of an account, we ran repeated measures ANOVAs with the two within-factors, humanness and partisanship congruency, and the previously mentioned control variables. Across all motivations, we consistently found a main effect for humanness and partisanship congruency and an interaction effect of both. Planned contrasts revealed that engagement results were similar to the engagement and reaction results. For incongruent accounts, levels of humanness did not matter. Most engagement motivations were equally low independent of an account's humanness. In contrast, for congruent accounts, most motivations increased with increased human-likeness. For a detailed report of the F statistics, p values, and planned contrasts, see Table S4 in the Supplementary Material. We found one exception for this pattern for the motivation to share a quoted retweet ("I want to argue by adding my own opinion to a retweet"). Here, we did not find a significant interaction of perceived humanness and partisanship congruency, $F(1.88, 428) = 0.89$, $p = .41$, $\eta_p^2 = .004$. While this motivation was less relevant for incongruent profiles, the motivation became more relevant with increased human-likeness.

Discussion

Drawing on insights from previous research on human–social bot interaction and motivated reasoning theory, we hypothesized that the likelihood of social bot accounts to engage with users depends on two factors: the humanness of an account and the partisanship displayed by the account. In doing so, we differentiated between initiating engagement of users with social bots and reactive engagement of users with social bots. The behaviors of interest were the engagement activities: following, retweeting, commenting, and quoted retweeting. In addition to the direct effect of partisanship displayed by the account, we also hypothesized partisanship to indirectly affect engagement likelihoods by altering how profiles are perceived (mediation hypothesis). Finally, to better understand why users chose to engage or not, we also explored users' engagement motivations.

Through repeated measures ANOVAs, we found for all four engagement activities of interest (following, retweeting, commenting, and quoted retweeting) the expected effects

of humanness and partisanship congruency as well as an interaction of both. These results indicated that (1) highly human-like accounts were more likely than medium and low human-like accounts to receive engagement from and also that (2) this was only true for congruent accounts. In contrast, accounts that did not share participants' partisanship were highly unlikely to receive engagement from participants. Hence, our study highlights that Twitter users are more willing to engage with human-like accounts, especially when they share the same political partisanship.

Similarly, when investigating how likely participants would react to accounts, only human-like, congruent accounts were likely to receive any engagement. However, it was most likely that participants would not react at all when an account initiated engagement. Independent of the level of humanness, incongruent accounts were most likely to be either blocked/reported or ignored. This implies that only very sophisticated social bots, which can successfully disguise their automated nature, are likely to engage with or receive engagement from users.

Moreover, we found that the impact of partisanship congruency was dependent on the level of humanness. The effect of partisanship congruency was largest for highly human-like accounts, smaller for medium human-like accounts, and smallest for low human-like accounts. This implies that users do not "blindly" engage with any account which shares their political partisanship but incorporate their perception of the humanness of the account into their engagement decision.

Results of the mediation hypothesis support this. Here, we revealed that the effect of congruency was partly due to biased humanness perceptions, indicating that congruent profiles were perceived as more human-like which, in turn, lead to an increased likelihood of user engagement. Similar to the engagement results, this suggests that users are unlikely to react to clear social bot accounts and most likely ignore or block/report these accounts. However, we did not find this effect for profiles that fell neither into the clearly social bot category (low human-like accounts) nor into the clearly human category (highly human-like accounts). We assume that due to the ambiguous nature of these accounts, participants needed to engage in more deliberative processing, which has previously been found to reduce the effect of motivated reasoning (Pennycook and Rand, 2019).

Especially the results concerning partisanship congruency confirm previous findings of motivated reasoning and homophilic patterns in social networks (Colleoni et al., 2014; Garz et al., 2020; Mosleh et al., 2021). Our results add to this that, in the context of social bot accounts, this pattern is partly due to biased perceptions of profiles with partisan-congruent profiles being perceived as more human-like (see also Wischnewski et al., 2021; Yan et al., 2020). However, our results also show limitations of this effect. Partisanship congruency mattered the least for bot-like accounts. With previous studies indicating that human-like social bot accounts are likely to be rare (Assenmacher et al., 2020), we conclude that the influential impact of social bots is likely to be overestimated. In fact, our results suggest that most accounts that show low to medium levels of humanness are likely to be ignored if they are congruent or blocked/reported if they are incongruent. However, this also implies that, as soon as social bots are well enough developed to successfully disguise their automated nature,

users become increasingly more susceptible, especially if accounts are tailored to support specific partisan views.

These results for humanness extend previous findings on user social bot engagement which conventionally do not differentiate between different levels of social bot humanness (Cardoso et al., 2019; Wagner et al., 2012). For example, Cardoso et al. (2019) found that an increasing number of users interact with social bots and share content from social bot accounts. However, the authors do not differentiate between different levels of humanness of these accounts. Our findings indicate that, besides the strong impact of partisan congruency, user engagement with social bots is most likely driven by highly human-like bots.

Finally, by investigating different engagement motivations, we could also show that the effects of humanness and partisanship congruency are reflected by users' motivations to engage. If accounts are incongruent, participants were generally not motivated to engage. If accounts were congruent, participants were most motivated to engage when the accounts were also human-like. In addition, we found that the different levels of humanness and partisanship (in)congruency of a profile affect all engagement motivations equally, except for the motivation to share a quoted retweet. Here, we found that the motivation "I want to argue by adding my own opinion to a retweet" was not affected by the displayed partisanship of the profile. This supports our initial hypothesis that different engagement activities are affected differently by partisanship congruency. Activities that do not imply endorsement, such as commenting or sharing a quoted retweet, should be less affected by the displayed partisanship.

Limitations and future work

The discussed results include methodological and theoretical limitations. By choosing Twitter as a social media platform, assumptions are restricted to it. Similarly, we can only make assumptions about political partisanship in the US context. However, different dynamics might occur when transferring our experimental setup to different cultures but also different polarizing contexts. For example, while partisans in the US are less likely to engage with each other (Finkel et al., 2020), other contexts might show different engagement patterns.

Moreover, participants in our study were immediately confronted with Twitter profiles. Consequently, suspicious behavior such as repetitive retweeting was immediately evident. However, in their everyday social media browsing experience, participants are more likely to come across single posts of accounts. In addition, we measured participants' engagement intentions but not actual behavior. While previous research suggests that intentions are generally a good indicator for behavior, research on the intention-behavior gap suggests that, under certain circumstances, behavior deviates from the intention (see Sheeran and Webb, 2016, for a review). Furthermore, the introductory definition of social bots might have primed participants to more aware of a possible bot presence which would not occur in a real scenario. To overcome these limitations and increase ecological validity, field experiments similar to Mosleh et al. (2021) are necessary to strengthen our findings.

Theoretically, our argumentation relies on the assumption that social bots exert influence through communication with users. In doing so, we imply that users actively engage with or react to social bots. This assumption bears at least two limitations. First, active engagement is not a necessary but only a sufficient requirement for influence. Especially findings on *mere exposure* and *prior exposure* show that one or multiple exposures to a stimulus can change perceptions (Pennycook et al., 2018; Zajonc, 1968). This implies that passive engagement without any reactions such as following, retweeting, commenting, or quoted retweeting can already influence users. To account for such passive engagement, eye-tracking studies could complement the findings of our study to detect concentration and attention to social bot accounts (e.g. Counts and Fisher, 2011). Second, our findings cannot account for the influence of social bots on a social network. Previous studies could show that social bots can amplify specific content, leading to “megaphone effects” (Woolley and Guilbeault, 2019: 193), as well as initiating political astroturfing campaigns (Keller et al., 2020). Because our results suggest that the *social influence* of social bots on individual users is low, we suspect that the actual impact of social bots originates in their ability to affect network structures, thus *network influence*.

Besides these methodological and theoretical limitations, our work also holds important social implications. In particular, our results suggest that with increased sophistication of social bots, in other words, an increased robotization of social media users, the line between “real” human users and “automated” users becomes increasingly blurred which can lead to feelings of alienation and dehumanization of human users (Fortunati et al., 2019).

Conclusion

In this article, we wanted to know under which conditions Twitter users engage with and react to social bots. We found that highly human-like social bots were most likely to receive user engagement and were also more likely to initiate engagement with users. We also found that this was only true for accounts that shared the same partisanship as the user. Thus, users prefer to engage with and react to highly human-like accounts that share the same political opinion. Moreover, this effect of partisanship congruency decreased for accounts displaying medium or low levels of humanness, indicating that users do not blindly engage with any account that shared their political partisanship. Thus, we conclude that the impact of social bots on individual users is nuanced and most likely overestimated. Social bot engagement is only effective if they achieve to disguise their automated nature.

Acknowledgements

We would like to thank Carolina Alves de Lima Salge and Björn Ross as well as both anonymous reviewers for their helpful comments.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the German Research Foundation (DFG) under Grant No. GRK 2167, Research Training Group “User-Centred Social Media.”

ORCID iD

Magdalena Wischnewski  <https://orcid.org/0000-0001-6377-0940>

Supplemental material

Supplemental material for this article is available online (<https://osf.io/w42ca/>).

Notes

1. A single indicator of an account's influence on social media networks developed by Klout.com.
2. We acknowledge that both following and retweeting are not always clear indications of endorsement. Some users on Twitter explicitly state, for example, in their bios that "RTs [retweets] ≠ endorsement." However, the disclaimer "RT ≠ endorsement" is likely a (meaningless) phrase dating back to the early use of Twitter by journalists (<https://www.buzzfeednews.com/article/charliewarzel/meet-the-man-behind-twitters-most-infamous-phrase>). Moreover, while some users follow others, for example, to surveil an account, most following motivations relate to endorsement (Ouwerkerk and Johnson, 2016).
3. Social bots are automated online accounts that communicate more or less autonomously and typically operate on social media sites, such as Twitter. They serve different functions like forwarding, liking, or commenting on specific topics like the weather, sport results, but also political issues. Social bots can also be programmed by companies to disseminate advertisement within social networking sites. They can openly disclose that they are automated accounts.

References

- Assenmacher D, Clever L, Frischlich L, et al. (2020) Demystifying social bots: on the intelligence of automated social media actors. *Social Media and Society* 6(3): 1–14.
- Bakardjieva M (2015) Rationalizing sociality: an unfinished script for socialbots. *The Information Society* 31(3): 244–256.
- Barberá P, Jost JT, Nagler J, et al. (2015) Tweeting from left to right: is online political communication more than an echo chamber? *Psychological Science* 26(10): 1531–1542.
- Bessi A and Ferrara E (2016) Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* 21(11): 1–17.
- Bolsen T, Druckman JN and Cook FL (2014) The influence of partisan motivated reasoning on public opinion. *Political Behavior* 36(2): 235–262.
- Bruns A (2019) *Are Filter Bubbles Real?* Cambridge: Polity Press.
- Cardoso F, Luceri L and Giordano S (2019) Digital weapons in social media manipulation campaigns 2018–2020. http://workshop-proceedings.icwsm.org/pdf/2020_32.pdf
- Carrus G, Panno A and Leone L (2018) The moderating role of interest in politics on the relations between conservative political orientation and denial of climate change. *Society and Natural Resources* 31(10): 1103–1117.
- Cheng C, Luo Y and Yu C (2020) Dynamic mechanism of social bots interfering with public opinion in network. *Physica A: Statistical Mechanics and Its Applications* 551: 124163.
- Cinelli M, De Francisci Morales G, Galeazzi A, et al. (2020) Echo chambers on social media: a comparative analysis. *Arxiv* 1–15. Available at: <https://arxiv.org/abs/2004.09603>
- Colleoni E, Rozza A and Arvidsson A (2014) Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64(2): 317–332.

- Counts S and Fisher K (2011) Taking it all in? Visual attention in microblog consumption. In: *Proceedings of the fifth international AAAI conference on weblogs and social media*, pp. 97–104. Available at: <https://ojs.aaai.org/index.php/ICWSM/article/view/14103>
- Diehl T, Weeks BE and Gil de Zúñiga H (2016) Political persuasion on social media: tracing direct and indirect effects of news use and social interaction. *New Media and Society* 18(9): 1875–1895.
- Edwards C, Edwards A, Spence PR, et al. (2014) Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior* 33: 372–376.
- Edwards C, Edwards A, Spence PR et al. (2015) Initial interaction expectations with robots: testing the human-to-human interaction script. *Communication Studies* 67(2): 1–12. <https://doi.org/10.1080/10510974.2015.1121899>
- Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22(8): 8005.
- Finkel EJ, Bail CA, Cikara M, et al. (2020) Political sectarianism in America. *Science* 370(6516): 533–536.
- Fortunati L, Manganelli AM, Cavallo F, et al. (2019) You need to show that you are not a robot. *New Media and Society* 21(8): 1859–1876.
- Garz M, Sörensen J and Stone DF (2020) Partisan selective engagement: evidence from Facebook. *Journal of Economic Behavior and Organization* 177: 91–108.
- Gehl RW and Bakardjieva M (2017) *Social Bots and Their Friends. Digital Media and the Automation of Sociality* (eds RW Gehl and M Bakardjieva). New York: Routledge.
- Gorodnichenko Y, Pham T and Talavera O (2018) Social media, sentiment and public opinions: evidence from #Brexit and #USElection. *European Economic Review* 136: 103772.
- Hayes AF (2017) *Introduction to Mediation, Moderation, and Conditional Process Analysis*. New York: Guilford Press.
- Hughes DJ, Rowe M, Batey M, et al. (2012) A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior* 28(2): 561–569.
- Hwang T, Pearce I and Nanis M (2012) Socialbots: voices from the fronts. *Interactions* 19(2): 38–45.
- Keijzer MA and Mäs M (2021) The strength of weak bots. *Online Social Networks and Media* 21: 100106.
- Keller FB, Schoch D, Stier S, et al. (2020) Political astroturfing on Twitter: how to coordinate a disinformation campaign. *Political Communication* 37(2): 256–280.
- Kitchens B, Johnson SL and Gray P (2020) Understanding echo chambers and filter bubbles: the impact of social media on diversification and partisan shifts in news consumption. *MIS Quarterly: Management Information Systems* 44(4): 1987–2011.
- Kunda Z (1987) Motivated inference: self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53(4): 636–647.
- Kunda Z (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3): 480–498.
- Luceri L, Braun T and Giordano S (2019) Analyzing and inferring human real-life behavior through online social networks with social influence deep learning. *Applied Network Science* 4(1). <https://doi.org/10.1007/s41109-019-0134-3>
- Mønsted B, Sapieżyński P, Ferrara E, et al. (2017) Evidence of complex contagion of information in social media: an experiment using Twitter bots. *PLoS ONE* 12(9): 1–12.
- Metaxas PT, Mustafaraj E, Wong K, et al. (2015) What do retweets indicate? Results from user survey and meta-review of research. *Proceedings of the AAAI International Conference on Web and Social Media* 9: 658–661.

- Mosleh M, Martel C, Eckles D, et al. (2021) Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences of the United States of America* 118(7): 9–11.
- Ouwerkerk JW and Johnson BK (2016) Motives for online friending and following: the dark side of social network site connections. *Social Media and Society* 2(3). <https://doi.org/10.1177/2056305116664219>
- Pennycook G, Cannon TD and Rand DG (2018) Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* 147(12): 1865–1880.
- Pennycook G and Rand DG (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>
- Ross B, Pilz L, Cabrera B, et al. (2019) Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems* 28(4): 394–412.
- Salge CA, de L, Karahanna E, et al. (2021) Algorithmic processes of social alertness and social transmission: how bots disseminate information on Twitter. *MIS Quarterly*. <https://doi.org/10.25300/MISQ/2021/15598>
- Schäfer F, Evert S and Heinrich P (2017) Japan’s 2014 general election: political bots, right-wing internet activism, and Prime Minister Shinzō Abe’s hidden nationalist agenda. *Big Data* 5(4): 294–309.
- Shani D (2012) Measuring political interest. In: Aldrich JH and McGraw KM (eds) *Improving Public Opinion Surveys: Interdisciplinary Innovation and the American National Election Studies*, pp. 139–157. Princeton, NJ: Princeton University Press.
- Sheeran P and Webb TL (2016) The intention–behavior gap. *Social and Personality Psychology Compass* 10(9): 503–518.
- Spence PR, Edwards A, Edwards C, et al. (2018) ‘The bot predicted rain, grab an umbrella’: few perceived differences in communication quality of a weather Twitterbot versus professional and amateur meteorologists. *Behaviour and Information Technology* 38(1): 101–109.
- Wagner C, Mitter S, Körner C, et al. (2012) When social bots attack: modeling susceptibility of users in online social networks. *Proceedings of the 2nd Workshop on Making Sense of Microposts* 838: 41–48.
- Wald R, Khoshgoftaar TM, Napolitano A, et al. (2013) Which users reply to and interact with Twitter social bots. In: *Proceedings of the international conference on tools with artificial intelligence (ICTAI)*, Herndon, VA, 4–6 November, pp. 135–144. New York: IEEE.
- Wang P, Angarita R and Renna I (2018) Is this the era of misinformation yet? Combining social bots and fake news to deceive the masses. In: *Proceedings of the companion of the Web conference*, Lyon, 23–27 April, pp. 1557–1561. New York: ACM.
- Wischnewski M, Bernemann R, Ngo T, et al. (2021) Disagree? You must be a bot! How beliefs shape Twitter profile perceptions. In: *Proceedings of the CHI conference on human factors in computing systems (CHI’21)*, 8–13 May, Yokohama, Japan. New York: ACM.
- Woolley SC and Guilbeault DR (2019) United States: manufacturing consensus online. In: Woolley SC and Howard PN (eds) *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press, pp. 185–211.
- Yan HY, Yang KC, Menczer F, et al. (2020) Asymmetrical perceptions of partisan political bots. *New Media and Society* 23(10): 3016–3037.
- Zajonc RB (1968) Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology* 9(2): 1–27.

Author biographies

Magdalena Wischnewski is a researcher in social and media psychology. Her interests are in the underlying processes of motivated reasoning and how motivated reasoning plays out on social media.

Thao Ngo is a researcher with a background in psychology and human factors with an interest in human–computer interaction, and explainable artificial intelligence (AI).

Rebecca Bernemann is a researcher in theoretical computer science. She is interested in graph analysis, especially in investigating uncertainty in social networks with the help of such concepts as Bayesian Networks and Petri Nets.

Martin Jansen is a researcher with a background in communication science and commerce and with an interest in micro-targeting, transparency, and consumer behavior.

Nicole Krämer is a researcher in social and media psychology with an interest in human–computer interaction and computer-mediated communication.

ARTICLE 4

The following article is reused from:

Wischnewski, M. & Krämer, N. (2020). I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption. In Proceedings of the 11th International Conference on Social Media and Society, 148–155.
<https://doi.org/10.1145/3400806.3400824>

I Reason Who I am? Identity Salience Manipulation to Reduce Motivated Reasoning in News Consumption

Magdalena Wischniewski *
Department of Social Psychology, University of
Duisburg-Essen, Germany
magdalena.wischniewski@uni-due.de

Nicole Krämer
Department of Social Psychology, University of
Duisburg-Essen, Germany
nicole.kraemer@uni-due.de

ABSTRACT

Past research has drawn on motivated reasoning theories in order to explain why some people fall for fake news while others do not. One such motivated reasoning paradigm proposes an elicitation of identity threat when incoming information is inconsistent with prior attitudes and beliefs. This experienced identity threat leads to biased information processing in order to defend those prior attitudes and beliefs. Building on this, we conducted two studies to test the overarching hypothesis that shifting identity salience changes information processing outcomes. In two experimental studies with $N = 353$, we tried to (1) increase factual information acceptance and (2) decrease misinformation acceptance. Our data support the previously found results that identity-threatening information decreases the evaluation of information compared to a control group. Findings also suggested that identity-supporting information was evaluated better, respectively. However, in both studies, identity salience manipulation did not change the evaluation of the information. Still, we found that those participants for whom another identity was made more salient indicated reduced feelings of anger compared to participants who were threatened and received no identity salience manipulation. We interpret these results as a promising first step to counter motivated reasoning processes.

CCS CONCEPTS

• **Social and Professional Topics**; • **User Characteristics**; • **Cultural Characteristics**; • **Collaborative and social computing**; • **Collaborative and social computing theory, concepts and paradigms**;

KEYWORDS

Motivated Reasoning, Identity Protection Cognition, Identity Salience, Misinformation, Anger

ACM Reference Format:

Magdalena Wischniewski and Nicole Krämer. 2020. I Reason Who I am? Identity Salience Manipulation to Reduce Motivated Reasoning in News

*corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SMSociety '20, July 22–24, 2020, Toronto, ON, Canada

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7688-4/20/07...\$15.00

<https://doi.org/10.1145/3400806.3400824>

Consumption. In *International Conference on Social Media and Society (SM-Society '20)*, July 22–24, 2020, Toronto, ON, Canada. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3400806.3400824>

1 INTRODUCTION

Today, misinformation spreads online with rapid velocity and seems to become ever more pervasive [42]. Repercussions of this vast increase of shared misinformation have not only become tangible in the health sector (see outbreaks of measles) [30] but more generally relate to increased distrust of public institutions [28]. Notably, social media sites such as Facebook, Twitter, and Reddit, that facilitate the creation and sharing of user-generated content are often associated with dissemination and consumption of misinformation [29]. In addition to that, messaging services like WhatsApp contribute increasingly to the circulation of misinformation [31]. These developments coincide with two facilitating processes, namely, (1) increasing numbers of people receive their news through social media platforms [32] and (2) content is created and spread by automated accounts (e.g., social bots) [43] and so-called trolls (e.g., Russian Intelligence Research Agency) [3] who target individual users. In all likelihood, more social media users will be exposed to misinformative content in the future.

In light of these trends, the question of why people fall for fake news has not only been raised within the research community [26] but also on popular news sites [27]. We can find answers in prior research on motivated reasoning. Motivated reasoning theories propose that individuals sometimes process incoming information with a bias in favor of existing beliefs and attitudes, “to arrive at a desired conclusion” [[2], p. 485]. According to this, motivated reasoning affects the perception of both misinformation and accurate information: misinformation that supports existing beliefs and attitudes is likely to be accepted (false positive acceptance), just as misinformation that contradicts those beliefs and attitudes is easily detected (true positive). In return, information rejection of factual information contradicting an individual’s prior beliefs and attitudes are higher (false negative) and vice-a-versa, in cases where information acceptance coincides with the individual’s prior belief’s and attitudes (true negative).

Analogous to the discussion of why people fall for fake news, studies examining the possibilities to protect individuals from misinformation, as well as to correct misinformation at its source, have increased in number. Many of these investigate misinformation corrections [8] and the acceptance of these corrections [25], but also inoculation strategies [6] and media literacy interventions [34]. This paper contributes to the existing literature by introducing another approach to deal with the increase of misinformation circulation and propose ways of protecting users. We employ a social

identity approach to address this. We propose manipulating identity salience of users' to either a) make a threatened identity less salient or b) to reduce the salience of a misinformation affirming identity.

2 THEORETICAL BACKGROUND

2.1 Information processing and motivated reasoning

The central claim of motivated reasoning theories is that the individual assessment of information is driven by prior attitudes [36] and beliefs as well as group belonging [5] and identity [16]. Motivated reasoning theories claim that if incoming information is congruent with an individual's prior beliefs and attitudes or if it supports one's world view, it is more likely to be believed, whereas information contradicting individuals' views is more likely to be rejected.

To explain the underlying psychological mechanisms of motivated reasoning, different theoretical approaches have been proposed. With origins in cognitive dissonance theory [7], earlier theories suggest that incoming information elicits different motives like the need for accuracy, self-defense and impression management [4] which are then associated with biased memory search and selection of reasoning strategies [20]. In particular, defense motivation as a result of perceived social identity threat has been related to directionally motivated information processing [13]. More recently, emotional processes have been introduced to add to the existing literature on cognitive and motivational approaches. For example, in coining the term *affective contagion*, Lodge and Taber [22], acknowledge these affective processes within motivated reasoning. They argue that early affective associations towards the incoming information determine memory retrieval and set the direction for subsequent processing via associative pathways [37]. By contrast, instead of focusing on early affective reactions, Suhay and Erisen [35] investigate the discrete emotions of anger, anxiety and enthusiasm, which all occur later, and their effect on motivated reasoning. Through mediation models, they confirmed the role of anger for attitude inconsistent information and enthusiasm for attitude consistent information. In their analysis, however, motivated reasoning was mostly fueled by anger. In addition to this research, a third theoretical approach centers on the role of "Identity Protection Cognition" (IPC) within the psychological processes of motivated reasoning.

2.2 Identity Protection Cognition and Identity Salience

A prerequisite for Identity Protection Cognition posits that specific ideas, ideologies or world views become "a badge of membership with identity-defining affinity groups" [[9], p. 2]. The formation of such groups is evident, especially when issues become politicized. In other words, group identities and alliances become tightly bound to specific belief systems which in turn affect behavior. According to Identity Protection Cognition, it is considered *individually rational* for individuals to reject information that is contrary to group beliefs, because acceptance would threaten their status within their affinity group. "[I]n fact, identity-protective cognition is a mode of engaging information rationally suited to the ends of the agents

who display it" [[9], p.1]. It is then the primary goal of an individual to protect her or his status within the respective group. In this, Identity Protection Cognition draws on both evolutionary psychology and utility maximation theory. While the former proposes that social groups fulfil basic needs of belonging, protection and safety, the latter supplements utility maximation theory in proposing that the benefits of conforming to group beliefs outweigh the costs of rejecting them. Furthermore, empirical studies in the context of politics [2] and technology acceptance [11] support Identity Protection Cognition. Van Bavel and Pereira [2] suggest in their identity-based model of beliefs that "accuracy goals compete with [partisan] identity goals to determine the value of beliefs" (p. 215). For our part, instead of focusing on partisan related identities and goals, we want to focus on the underlying processes of identity threat and identity affirmation.

As stated above, identity-based beliefs may either support or reject factual information or misinformation. In our two studies we focussed on false negatives and false positives (see *c* and *d* in Table 1). The overall aims are to increase acceptance and decrease rejection of factual information (thereby addressing false negatives), and decrease acceptance and increase rejection of misinformation (with an eye to false positives), through a manipulation of identity salience. We hypothesize that changing the salience of an individual's identity from a threatened identity to an unthreatened identity will, in turn, increase the likelihood of accepting factual information and rejecting misinformation.

In general, identity salience manipulations have been successfully implemented to change attitudes and behavior as observed in studies on stereotype susceptibility [33], stereotype threat [24] and performance [18] but also within the realm of policy support [41]. These studies draw mostly on social identity theory and self-categorization theory. According to social identity theory, people can define themselves both in terms of who they are as an individual as well as their membership in various groups [38]. Turner, Oakes, Haslam and McGarty [40] extend this view by what they called self-categorization theory. Self-categorization theory picks up the concept of personal and social identity and describes these as "different levels of self-categorization" (p.1). It asserts an experience of self through varying identities, and that the prevalent identity shifts in response to contextual and social cues. Consequently, once another identity becomes more salient, the respective norms of the salient group guide downstream cognition and emotion [15].

Moreover, we hypothesize that a shift in identity salience would not only affect how individuals think about incoming information but also how they feel about it. Because experiences of identity threat have previously been shown to induce either feelings of anxiety or anger [16], we hypothesized that a manipulation of identity salience should also be reflected by an identifiable emotional reaction. Depending on the emotion induced by the identity-threatening information (either anger or anxiety), this emotion should decrease upon the manipulation of identity salience. Building on this, we conducted two studies, (1) which tried to increase the acceptance of factual information and (2) which tried to increase the rejection of misinformation.

Table 1: Processing (mis-)information.

	Information is true (factual information)	Information is false (misinformation)
Belief supports information	(a) True positive	(b) False positive
Belief contradicts information	(c) False negative	(d) True negative

2.3 Hypotheses Study 1

In study 1, we hypothesized the following: (H1.1) If incoming information (news) threatens the identity of individuals, the information is evaluated more poorly than compared to a group whose identity is not threatened by the information. (H1.2) This identity threat is accompanied by an emotional reaction. Therefore, identity-threatened individuals experience higher levels of anger or anxieties. (H2.1) When the salient identity of an individual is changed to an unthreatened identity, the individual will evaluate the factual information better. (H2.2) An identity salience manipulation for an unthreatened identity will result in no changes in evaluation. (H3.1) The change from a threatened to an unthreatened identity will also be reflected in a change to the emotion experienced (i.e., individuals will experience less anger or anxiety). (H3.2) The identity salience manipulation for unthreatened identities will result in no changes in experienced emotion. See Appendix A.1 for an overview of all hypotheses.

2.4 Hypotheses Study 2

For study 2, instead of increasing the acceptance of factual information, we intended to increase the rejection of misinformation (i.e., (b) false positives in Table 1). Building on IPC, we hypothesized the following. (H4.1) If incoming information (misinformation) supports the identity of individuals, the information evaluation is better than compared to a group whose identity is not supported by the information. (H4.2) This identity support is accompanied by an emotional reaction towards the identity-supporting stimulus. Therefore, individuals experience higher levels of enthusiasm. (H5.1) By changing the salient identity of individuals to a non-supporting identity, the affected individuals will have a poorer evaluation of the misinformation. (H5.2) An identity salience manipulation from one non-supporting identity to another will result in no changes in evaluation. (H6.1) The change from a supporting to a non-supporting identity will also be reflected in a change of experienced emotion. Individuals should experience less enthusiasm. (H6.2) The identity salience manipulation from non-supported to another non-supported identity will result in no changes in experienced emotion. See Appendix A.2 for an overview of all hypotheses.

3 METHODS

Both studies received ethical approval by the ethics committee of the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen.

3.1 Study 1 - Design and Procedure

To test our hypotheses, we conducted a 2 (manipulation versus no manipulation) x 2 (women versus men) between-subject study with

247 University students (175 female) with an age range of 18 to 62 years ($M = 23.36$, $SD = 3.65$). Participants of both groups (control and experimental group) were asked to read and evaluate a news article. In addition, we asked for participants' affective reactions after reading the article. After being randomly assigned to one of the two groups, participants in the experimental group received an identity salience manipulation before reading the news article. All participants were debriefed upon completion.

In order to reduce identity-threatening potential, per our hypotheses, of the news article, we implemented an identity salience manipulation. Since all participants were university students, we intended to shift identity salience to that identity. For this, we applied a manipulation used by Shi, Pittinsky, and Ambady [33], which has previously been found to induce identity salience change. Participants of the experimental group were asked the following four questions: 1) if they were enrolled at a university, 2) what subject they were studying, 3) which semester they were in, and 4) if they were living in a student home.

3.2 Stimulus material

The article presented to all participants discussed factual information about domestic violence. Specifically, it focused on domestic violence committed by female perpetrators against their male partners, and, hence, was written in a way that threatened women's identity. The article was relatively short (approximately 350 words) and was presented in the design of a well-known newspaper.

3.3 Evaluation of the articles

To measure how the participants evaluated the article, we used the Trust in News Media (TiNM) scale by Kohring and Matthes [19]. The scale is a standardized and validated multidimensional measurement to depict trust and credibility of news media. It consists of four lower-order factors that are assessed by four items: (1) trust in the selectivity of topics, (2) trust in the selectivity of facts, (3) trust in the accuracy of depictions and (4) trust in journalistic assessment. Answers were given on a five-point Likert scale (1 = "do not agree at all", 5 = "fully agree") and items reached acceptable reliability of Cronbach's $\alpha = .86$. The complete list of items can be assessed in Table 2

3.4 Emotions

To assess emotional reactions towards the news article, we asked participants to self-report their experience of anger and anxiety. To do so, we created six items based on Affective Intelligence Theory (AIT) [23]. Anger and anxiety were measured through three items respectively: hateful, angry, outraged (Cronbach's $\alpha = .65$) and afraid, worried, anxious (Cronbach's $\alpha = .71$). Participants indicated on a five-point Likert scale (1 = "do not agree at all", 5 = "fully

Table 2: Items of the Trust in News Media Scale by Kohring and Matthes.

Items	
Selectivity of topics	The topic of domestic violence by female perpetrators received the necessary attention.
	The topic of domestic violence by female perpetrator is assigned an adequate status
	The frequency with which domestic violence by female perpetrators is covered is adequate.
Selectivity of facts	The topic is covered on the necessary regular basis.
	The essential points are included.
	The focus is on important facts.
Accuracy of depiction	All important information regarding the topic of domestic violence by female perpetrators is provided.
	Reporting includes different points of view.
	The information in a report would be verifiable if examined.
	The reported information is true.
Journalistic assessment	The report recounts the facts truthfully.
	The facts that I received regarding domestic violence by female perpetrator are correct
	Criticism is expressed in an adequate manner.
	The journalist's opinions are well-founded.
Journalistic assessment	The commentary regarding domestic violence by female perpetrators consists of well-considered conclusions.
	I feel that the journalistic assessment regarding the topic of domestic violence by female perpetrators is useful.

Table 3: Mean scores of TiNM, self-reported anger and anxiety by group.

Condition		TiNM		Anger		Anxiety	
		M	SD	M	SD	M	SD
Salience manipulated	female	40.40	9.46	6.57	2.54	7.40	2.80
	male	42.50	9.24	6.30	2.05	6.65	2.56
Salience not manipulated	female	38.62	9.83	7.46	2.87	7.53	2.71
	male	43.13	10.91	6.41	3.01	6.84	2.55

agree”) how they felt when reading the article. The use of AIT to assess emotions via self-reporting is sound and has previously been used in several studies.

3.5 Results study 1

Mean scores over all groups for the Trust in News Media scale, self-reported anger and self-reported anxiety are presented in Table 3. To test the influence of identity threat on article evaluation (H1.1) and the proposed effect of an identity salience manipulation to decrease that effect (H2.1), we conducted an analysis of variance (ANOVA) with planned contrasts, with a standard $p < .05$ criterium of significance. The groups (manipulation/no manipulation) were entered as independent factors whereas the Trust in News Media score was entered as a dependent variable.

The overall model did not reach significance, $F(3, 242) = 2.6, p = .053, \eta_p^2 = .03$. Yet, to answer H1.1, we compared the evaluation of female readers with male readers through planned contrast. H1.1 hypothesized that female readers would evaluate the article worse than male readers. This was supported in our data: male readers assigned a higher quality evaluation to the article than did female

readers ($F(1, 242) = 6.22, p = .01, \eta_p^2 = .03$). However, this was neither reflected by the reported experience of anger ($F(1, 242) = 3.1, p = .08, \eta_p^2 = .01$) nor of anxiety ($F(1, 242) = 3.4, p = .06, \eta_p^2 = .01$), as we had hypothesized in H1.2.

Concerning H2.1, which hypothesized that an identity salience manipulation would lead to a decreased identity threat, we conducted another planned contrast. However, the identity salience manipulation in one group did not change how females evaluated the information compared to females of the group without a manipulation ($F(1, 242) = 1.37, p = .24$). As hypothesized in H2.2, we did not find a change in the evaluation between male readers ($F(1, 242) = 0.1, p = .76$). Although our data did not support that the identity salience manipulation changed the evaluation of the article, we found that female readers of the salience manipulation reported lower levels of anger than females of the control group ($F(1, 242) = 4.82, p = .03$). These data support H3.1. This was, however, only true for reported levels of anger but not for anxiety ($F(1, 242) = 0.12, p = .74$). As hypothesized (H3.2), male readers of the salience manipulation group did not differ from male readers of the control

Table 4: Items of the Group Identification Scale. (R) indicates a reversed-scored item.

Subscale	Item
Affective	1. I would like to be in a different group (R).
	2. Members of this group like one another.
	3. I enjoy interacting with the members of this group.
	4. I don't like many of the people in this group (R).
Behavioral	5. In this group, members don't have to rely on one another (R).
	6. All members need to contribute to achieve the group's goals
	7. This group accomplishes things that no single member could achieve alone.
	8. In this group, members do not need to cooperate to complete group tasks (R).
Cognitive	9. I think of this group as a part of who I am.
	10. I see myself as quite different from other members of the group (R).
	11. I don't think of this group as a part of who I am (R).
	12. I see myself as quite similar to other members of the group.

group concerning anger ($F(1, 242) = 0.03, p = .87$) and anxiety ($F(1, 242) = 0.1, p = .75$).

The results from study 1 showed that Identity Protection Cognition was supported. The threatened group gave a worse evaluation of the factual news article than did the non-threatened group, as was also reflected in self-reported levels of anger. The manipulation of identity salience to close this gap was unsuccessful. After receiving the manipulation, mean evaluations did not change significantly, although we did observe an increase in mean evaluation in the hypothesized direction. Self-reported levels of anger changed significantly after the manipulation, but not self-reported levels of anxiety.

3.6 Study 2 - Design and Procedure

The basic construction of study 2 was similar to study 1. To test our hypotheses, we conducted a 2 (manipulation versus no manipulation) x 2 (vegetarian versus meat-eater) between-subject study with 106 University students (75 female) with an age range of 18 to 56 years ($M = 23.44, SD = 4.01$). Participants of both groups (control and experimental group) were asked to read and evaluate a news article. In contrast to study 1, the news article was written in the style of misinformation. In doing so, we oriented ourselves to findings by Horne and Adali [14], including elements like simple language and sensationalizing headlines. After evaluating the quality of the article, participants were asked to indicate their experienced enthusiasm. As in study 1, one group's participants received an identity salience manipulation before reading the news article in order to reduce our hypothesized identity-supporting potential of the misinformation article. Here we aimed to make the identity of *student* more salient. After the experience in study 1 of employing a manipulation consisting of only four questions, we decided to use a more elaborate manipulation in study 2. We applied the group identification scale developed by Henry, Arrow, and Carini [10]. The scale consists of 12 items which can be divided into three subscales: cognitive (social categorization), affective (interpersonal attraction) and behavioral (interdependence). For a full list of all questions see Table 4. After completion, all participants were debriefed.

3.7 Stimulus Material

The article presented to all participants discussed misinformation about the apparent positive effect of a vegetarian diet on job success. It was written in such a way as to venerate individuals who are vegetarian or vegan. The article was similar in length to study 1 (~ 330 words) and was presented in the design of a well-known newspaper.

3.8 Evaluation of the article

As in study 1, we used the Trust in News Media (TiNM) scale to assess how participants evaluated the article.

3.9 Emotions

Because we expected that participants following a vegetarian diet would feel positively about the identity-supporting misinformation, instead of assessing anger and anxiety, participants were asked to self-report feelings of enthusiasm. As with the evaluation of anger and anxiety in study 1, the measure of enthusiasm was based on *Affective Intelligence Theory* (AIT) by [23], and was assessed through three categories: proud, enthusiastic and hopeful (Cronbach's $\alpha = .86$). Participants indicated on a five-point Likert scale (1 = "do not agree at all", 5 = "fully agree") how they felt when reading the article.

3.10 Results study 2

For an overview of mean responses concerning TiNM scores and self-reported enthusiasm, see Table 5. To determine whether individuals whose identity was supported by the misinformation evaluated the article better than those whose identity was not supported (H4.1), we conducted an analysis of variance (ANOVA) with planned contrasts and controlled for age and gender, with a standard $p < .05$ criterion of significance. The overall model was not significant concerning the grouping variable ($F(3, 99) = 2.63, p = .054, \eta_p^2 = .07$). However, the planned contrast revealed that vegetarian readers evaluated the misinformation significantly better than meat-eaters ($F(1, 99) = 4.66, p = .03$), as was predicted

Table 5: Mean scores of TiNM score and self-reported enthusiasm by group.

Condition		TiSM		Enthusiasm	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Salience manipulated	vegetarian	46.62	8.22	10.00	2.51
	meat eater	45.57	9.26	6.03	2.88
Salience not manipulated	vegetarian	48.75	7.78	9.88	3.27
	meat eater	42.61	7.8	6.52	2.58

by identity-support hypothesis (H4.1). This was also reflected in higher self-reported levels of enthusiasm ($F(1, 99) = 39.07, p < .001, \eta_p^2 = .28$) (H4.2)

To test whether shifting identity salience from an identity-support to an uninvolved identity, we conducted another planned contrast. However, results revealed no significant change between vegetarian readers of the manipulation group and vegetarian readers of the control group ($F(1, 99) = 1.24, p = .27, \eta_p^2 = .01$) (H5.1). Nevertheless, inspecting the mean scores (see Table 5), we noted a change in the hypothesized direction ($M_{\text{vege_manipulated}} < M_{\text{vege_control}}$). For the non-supported identity, we found that the manipulation did not change the evaluation as was expected ($F(1, 99) = 2.14, p = .15, \eta_p^2 = .02$) (H5.2). This overall picture was confirmed when running an ANOVA with planned contrast for the dependent variable enthusiasm as well. Manipulation did not affect the responses for enthusiasm of identity-supported individuals ($F(1, 99) = 0.04, p = .85, \eta_p^2 < .001$) (H6.1) and non-supported individuals ($F(1, 99) = 0.63, p = .43, \eta_p^2 = .02$) (H6.2).

4 DISCUSSION

In this paper, we wanted to assess if an individual's evaluation of a news article changes when their identity salience is manipulated. In doing so, our overall aim was to increase the acceptance of factual information on the one hand and to increase rejections of misinformation on the other. We build on Identity Protection Cognition which argues both that information is rejected when it threatens an individual's identity and that a piece of information is accepted when it supports an individual's identity. Our hypothesis was, then, that by changing an individual's identity from a threatened to an unthreatened identity, they would be more likely to accept incoming information. Further, by our changing an individual's salient identity from a supporting identity to a non-supporting, they would be more likely to reject incoming information. We applied this to increase factual information acceptance (study 1) and increase misinformation rejection (study 2).

The results of study 1 confirmed Identity Protection Cognition insofar as individuals whose identity was threatened by the information of the news article did indeed evaluate the article to be worse than individuals whose identity was not threatened. They also self-reported significantly higher instances of experiencing anger. However, even after changing identity salience to an unthreatened identity, we did not find that individuals indicated significantly increased acceptance of the news article compared to the control group who did not receive the salience manipulation. However, we did find that, although evaluation did not change after the salience manipulation, levels of self-reported anger were significantly lower

for individuals of the manipulation condition. We did not observe this finding for levels of self-reported anxiety.

The aim of study 2 was to increase misinformation rejection for individuals whose identity was supported by the misinformation. Our results confirmed Identity Protection Cognition, as individuals whose identity was supported evaluated the misinformation better than those whose identity was not supported. We found this to be reflected by the self-reported levels of enthusiasm, as individuals whose identity was supported indicated significantly higher levels of enthusiasm. Contrary to our hypotheses, changing identity salience to a non-supporting identity did not result in lower levels of misinformation acceptance. Although mean acceptance decreased, this change was not significant. Likewise, we found no support for our hypothesis in self-reported experiences of enthusiasm. The identity salience manipulation did not change how individuals felt about the misinformation.

4.1 Limitations

Before we dive deeper into implications, we want to elaborate on the limitations of our findings. In study 1, we found evidence in our data that manipulating identity salience led to lower levels of anger. This suggests that the information was less identity-threatening. However, we did not find this reflected in the mean evaluation (TiNM scores). Although the mean evaluation of participants from the salience manipulation condition was higher as hypothesized ($M_{\text{manipulation}} = 40.40; M_{\text{no manipulation}} = 38.62$), the difference was not significant. Results from study 2 confirmed this. Similarly, while the mean evaluation decreased in the hypothesized direction, the change was not significant. We identify four possibly explanations for these results.

(1) The manipulation had no real effect across the two studies. However, we would like to call attention to the significant change in experienced anger that was recorded. Given this, we assume that at least in study 1 the manipulation worked. Future studies would need to further investigate the shift in identity salience via manipulation.

(2) Identity salience was manipulated, but the effect of such a manipulation was not lasting. Again, we cannot refuse this either. We know that the duration of framing effects depend heavily on individual characteristics like a person's political knowledge [21]. Nevertheless, little is known about the effectiveness of identity salience manipulations.

(3) The manipulation worked but its effect was too weak to be statistically significant.

(4) Similarly to the previous possible explanation, both studies may have been underpowered. To this and the previous explanation,

we see reason to believe that our results support our original hypotheses. Our data demonstrated the expected direction of change for both studies when scrutinizing the mean differences, and we saw the same pattern for the experienced emotion of anger.

4.2 Implications and Conclusion

Despite its limitations, our research offers some valuable conclusions. First, both studies support claims made by Identity Protection Cognition. Threatened individuals evaluated factual information as being of poorer quality, and experienced higher levels of anger. At the same time, information, in our case misinformation, that supports an identity is evaluated as being of better quality, while also reflecting a higher level of experienced enthusiasm on the part of the reader. This has important implications for how we approach misinformation and factual information acceptance, namely, going further than confirmation and partisan bias. While it is true that issues arise when prior attitudes are confirmed or rejected by incoming information, incorporating attitudes in a model of identity might explain why misinformation corrections [8], fact-checking [1], and other attempts to rectify false information fail. If an attitude or previously held opinion becomes closely associated with an affinity group, as Identity Protection Cognition suggests, corrections to information are likely to be ignored, counterargued, or outright dismissed. Although studies have shown that individuals are willing to adapt their views [39]—a phenomenon that has previously been related to “Cognitive Reflection” [26]—we find those cases to be generally less polarized. Studies have shown that beliefs can become more polarised after increasing the saliency of the reader’s political identity [41], while others have indicated that intergroup comparisons can maximise perception of between-group differences [12]. In light of this, we argue that whether it is delivering news or offering misinformation corrections, practitioners need to consider these effects that are so associated with the readers’ identities. This implies refraining from framing issues in a politicized light but instead focusing on what is said rather than by whom.

In closing, let us turn to the by far more obvious problem: social media. Here, admittedly, little can be done concerning the visibility of political identities. We fear that other issues like content moderation [9] and the monetization of clicks and views may be far more decisive than any identity salience manipulation we might devise. Nevertheless, we do not intend to discourage anyone. In fact, building on our results, we encourage our colleagues to investigate not only the role of identity in information processing and motivated reasoning, but also emotional processes and contextual cues.

ACKNOWLEDGMENTS

This research was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group “User-Centred Social Media”. Special thanks to Louisa Nosch, Agnes Dyszlewski, Ida Schaffeld, and Britta Tröger for their help in the conceptualization and data collection.

REFERENCES

- [1] Michael J. Aird, Ullrich K.H. Ecker, Briony Swire, Adam J. Berinsky, and Stephan Lewandowsky. 2018. Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *R. Soc. Open Sci.* 5, 12 (2018). DOI:https://doi.org/10.1098/rsos.180593
- [2] J. J. Van Bavel and A. Pereira. 2018. The partisan brain: An identity-based model of political belief. *Trends Cogn. Sci.* 22, 3 (2018), 213–224.
- [3] Brandon Boatwright, Darren L. Linvill, and Patrick L. Warren. 2018. Troll factories: The IRA and state-sponsored agenda building. *Resour. Cent. Media Free. Eur.* (2018).
- [4] Shelly Chaiken, Roger Giner-Sorolla, and Serena Chen. 1996. Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In *The Psychology of action: Linking cognition and motivation to behavior*, P. M. Gollwitzer and J. A. Bargh (eds.). Guilford Press, New York, NY, 553–578.
- [5] G. L. Cohen. 2003. Party over policy: The dominating impact of group influence on political beliefs. *J. Pers. Soc. Psychol.* 85, 5 (2003), 808–822. DOI:https://doi.org/10.1037/0022-3514.85.5.808
- [6] John Cook, Stephan Lewandowsky, and Ullrich K.H. Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One* 12, 5 (2017), 1–21. DOI:https://doi.org/10.1371/journal.pone.0175799
- [7] L. Festinger. 1976. *A theory of cognitive dissonance*. Stanford University Press, Stanford, CA.
- [8] D. J. Flynn, Brendan Nyhan, and Jason Reifler. 2017. The Nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Polit. Psychol.* 38, 1 (2017), 127–150. DOI:https://doi.org/10.1111/pops.12394
- [9] Giovanni De Gregorio. 2019. Democratising online content moderation: A constitutional framework. *Comput. Law Secur. Rev.* xxx (2019), 105374. DOI:https://doi.org/10.1016/j.clsr.2019.105374
- [10] Kelly Bouas Henry, Holly Arrow, and Barbara Carini. 1999. A tripartite model of group identification: Theory and measurement. *Small Gr. Res.* 30, 5 (1999), 558–581. DOI:https://doi.org/10.1177/104649649903000504
- [11] Shirley S. Ho, Dietram A. Scheufele, and Elizabeth A. Corley. 2010. Making sense of policy choices: Understanding the roles of value predispositions, mass media, and cognitive processing in public attitudes toward nanotechnology. *J. Nanoparticle Res.* 12, 8 (2010), 2703–2715. DOI:https://doi.org/10.1007/s11051-010-0038-8
- [12] Michael A. Hogg, J. Turner, and Barbara Davidson. 1990. Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic Appl. Soc. Psychol.* 11, 1 (1990).
- [13] Natascha de Hoog. 2013. Processing of social identity threats: A defense motivation perspective. *Soc. Psychol. (Gott)* 44, (2013), 361–372. DOI:https://doi.org/10.1027/1864-9335/a000133
- [14] Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. (2017), 759–766. Retrieved from <http://arxiv.org/abs/1703.09398>
- [15] Matthew J. Hornsey. 2008. Social identity theory and self-categorization theory: A historical review. *Soc. Personal. Psychol. Compass* 2, 1 (2008), 204–222. DOI:https://doi.org/10.1111/j.1751-9004.2007.00066.x
- [16] Leonie Huddy, Lilliana Mason, and Lene Aarøe. 2015. Expressive partisanship: campaign involvement, political emotion, and partisan identity. *Am. Polit. Sci. Rev.* 109, 1 (2015), 1–17. DOI:https://doi.org/10.1017/S0003055414000604
- [17] Dan M. Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition.
- [18] D. Van Knippenberg and N. Ellemers. 2003. Social identity and group performance. Identification as the key to group-oriented effort. In *Social identity at work: Developing theory for organizational practice*, S. Alexander Haslam, Daan van Knippenberg, Michael J. Platow and Naomi Ellemers (eds.). Psychology Press.
- [19] Matthias Kohring and Jörg Matthes. 2007. Trust in news media: Development and validation of a multidimensional scale. *Communic. Res.* 34, 2 (2007), 231–252. DOI:https://doi.org/10.1177/0093650206298071
- [20] Ziva Kunda. 1990. The case for motivated reasoning. *Psychol. Bull.* 108, 3 (1990), 480–498. DOI:https://doi.org/10.1037/0033-2909.108.3.480
- [21] Sophie Lecheler, Andreas R.T. Schuck, and Claes H. De Vreese. 2013. Dealing with feelings: Positive and negative discrete emotions as mediators of news framing effects. *Communications* 38, 2 (2013), 189–209. DOI:https://doi.org/10.1515/commun-2013-0011
- [22] Milton Lodge and Charles S. Taber. 2013. Affective contagion and political thinking. In *The Rationalizing voter*.
- [23] George E. Marcus, W. Russell Neuman, and Michael B. MacKuen. 2000. *Affective intelligence and political judgement*. University of Chicago Press.
- [24] Matthew S. McGlone and Joshua Aronson. 2006. Stereotype threat, identity salience, and spatial reasoning. *J. Appl. Dev. Psychol.* 27, 5 (2006), 486–493. DOI:https://doi.org/10.1016/j.appdev.2006.06.003
- [25] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Polit. Behav.* 32, 2 (2010), 303–330.
- [26] Gordon Pennycook and David G. Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018), 1–12. DOI:https://doi.org/10.1016/j.cognition.2018.06.011
- [27] Gordon Pennycook and David G. Rand. 2019. Why do people fall for fake news? *The New York Times*.

[28] Nathaniel Persily. 2017. The 2016 U.S. Election: Can democracy survive the internet? *J. Democr.*28, 2 (2017), 63–76.

[29] Pew Research Center. 2016. Many americans believe fake news is sowing confusion. *Encycl. Fam. Stud.* (2016). DOI:<https://doi.org/10.1002/9781119085621.wbefs533>

[30] Jacek Radzikowski, A. Stefanidis, Kathryn H Jacobsen, Arie Croitoru, A. Crooks, and Paul L. Delamater. 2016. The measles vaccination narrative in twitter: A quantitative analysis. *JMIR Public Heal. Surveill.*2, 1 (2016), 1–15. DOI:<https://doi.org/10.2196/publichealth.5059>

[31] Gustavo Resende, Philippe Melo, Hugo Sousa, Johnatan Messias, Marisa Vasconcelos, Jussara M. Almeida, and Fabricia Benevenuto. 2019. (Mis) Information dissemination in WhatsApp: Gathering, analyzing and countermeasures. *Proc. World Wide Conf.* (2019), 818–828.

[32] Elisa Shearer and Elizabeth Grieco. 2019. Americans are wary of the role social media sites play in delivering the news. *Pew Res. Cent.* (2019).

[33] Margaret Shih, Todd L. Pittinsky, and Nalini Ambady. 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychol. Sci.*10, 1 (1999), 80–83. DOI:<https://doi.org/10.1111/1467-9280.00111>

[34] Susan Currie Sivek. 2018. Both facts and feelings: Emotion and news literacy. *J. Media Lit. Educ.*10, 2 (2018), 123–138. DOI:<https://doi.org/10.23860/jmle-2018-10-2-7>

[35] Elizabeth Suhay and Cengiz Erisen. 2018. The role of anger in the biased assimilation of political information. *Polit. Psychol.*39, 4 (2018), 793–810. DOI:<https://doi.org/10.1111/pops.12463>

[36] Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *Am. J. Pol. Sci.*50, 3 (2006), 755–769. DOI:<https://doi.org/10.1111/j.1540-5907.2006.00214.x>

[37] Charles S Taber and Milton Lodge. 2016. The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Polit. Psychol.*37, (2016), 61–85. DOI:<https://doi.org/10.1111/pops.12321>

[38] H. Tajfel and J. Turner. 1986. The social identity theory of intergroup behavior. In *Key Readings in Social Psychology: Political Psychology*, J.T. Jost and J. Sidanius (eds.). Psychology Press, 276–293.

[39] Ben M. Tappin, Gordon Pennycook, and David G Rand. 2018. Rethinking the link between cognitive sophistication and identity-protective bias in political belief formation. *Preprint*. Retrieved from <https://doi.org/10.31234/osf.io/yuzfj>

[40] J. Turner, Penelope J. Oakes, S. Alexander Haslam, and Craig McGarty. 1994. Self and collective: Cognition and social context. *Soc. Personal. Soc. Psychol.*20, 5 (1994), 454–469.

[41] Kerrie L. Unsworth and Kelly S. Fielding. 2014. It’s political: How the salience of one’s political identity changes climate change beliefs and policy support. *Glob. Environ. Chang.*27, 1 (2014), 131–137. DOI:<https://doi.org/10.1016/j.gloenvcha.2014.05.002>

[42] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science (80-.)*.359, 6380 (2018), 1146–1151. DOI:<https://doi.org/DOI:10.1126/science.aap9559>

[43] Samuel C. Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday* 21, 4 (2016), 1–24.

A APPENDIX

A.1 Hypothesis for study 1 and their respective results.

	Hypotheses	Confirmed?
H1.1	Incoming information threatens the identity of individuals. As a result, the information is evaluated more poorly than compared to a group whose identity is not threatened by the information.	Yes
H1.2	Identity threat is accompanied by the emotional reactions of anger or anxieties.	Yes
H2.1	By changing the salient identity of individuals to an unthreatened identity, the individuals will evaluate the factual information better.	No
H2.2	An identity salience manipulation for an unthreatened identity will result in no changes in evaluation.	Yes
H3.1	The change from a threatened to an unthreatened identity will also be reflected in a change of experienced emotion. Individuals will experience less anger or anxiety.	Yes
H3.2	The identity salience manipulation for unthreatened identities will result in no changes in emotion experienced.	Yes

A.2 Hypothesis for study 2 and their respective results.

	Hypotheses	Confirmed?
H4.1	If incoming information (misinformation) supports the identity of individuals, the information evaluation is better than compared to a group whose identity is not supported by the information.	Yes
H4.2	This identity support is accompanied by higher levels of enthusiasm.	Yes
H5.1	Changing the salient identity of individuals to a non-supporting identity, the individuals will evaluate the misinformation worse.	No
H5.2	An identity salience manipulation from one non-supporting identity to another will result in no changes in evaluation	Yes
H6.1	The change from a supporting to a non-supporting identity will also be reflected in a change of experienced emotion. Individuals will experience less enthusiasm.	No
H6.2	The identity salience manipulation from nonsupported to another non-supported identity will result in no changes in experienced emotion	Yes

ARTICLE 5

The following article is reused from:

Wischnewski, M. & Krämer, N. (2021). The role of emotions and identity-protection cognition when processing (mis) information, *Technology, Mind, and Behavior*, 2(1).
<https://doi.org/10.1037/tmb0000029>

The Role of Emotions and Identity-Protection Cognition When Processing (Mis)Information

Magdalena Wischniewski and Nicole Krämer

Social Psychology: Media and Communication, University of Duisburg-Essen



In this study, we investigate the role of emotions in identity-protection cognition to understand how people draw inferences from politicized (mis-)information. In doing so, we combine the identity-protection cognition theory with insights about the effects of emotions on information processing. Central to our study, we assume that the relationship between an individual's political identity and inference-conclusions of politicized information is mediated by the experienced emotions anger, anxiety, and enthusiasm. In an online study, 463 German adults were asked to interpret numerical information in two politically polarizing contexts (refugee intake and driving ban for Diesel cars) and one nonpolarizing context (treatment of skin rash). Results showed that, although emotions were mostly unrelated to political identity, they predicted performance more consistently than political identity and cognitive sophistication.

Keywords: misinformation, identity-protection cognition, emotions, mediation, cognitive sophistication

Supplemental materials: <https://doi.org/10.1037/tmb0000029.supp>

The increased circulation of misinformation and its direct linkage to damaging consequences, such as information avoidance about the recent Coronavirus disease 2019 (COVID-19) outbreak (Kim et al., 2020), decreased trust in the media (Turcotte et al., 2015), and environmental harm (Farrell, 2019), has resulted in the proliferation of academic studies investigating misinformation. Different approaches and fields investigated, for example, the


spread of misinformation (Brady et al., 2017), susceptibility for it (Druckman, 2012; Pennycook & Rand, 2019), and the difficulty to correct it (De keersmaecker & Roets, 2017; Flynn et al., 2017).

Previous studies about misinformation acceptance have placed their investigation within the broader concept of motivated reasoning to understand its underlying psychological mechanisms. Motivated reasoning, also known as biased assimilation (Lord et al., 1979), generally assumes that information processing and assimilation are sometimes biased in favor of one's prior-beliefs and attitudes.

In this article, we contribute to this growing knowledge on misinformation by investigating the role of identity protection cognition and protection-related emotions. To do so, in our theory-driven approach, we combine insights from the theory of identity-protection cognition (IPC), an identity-centered motivated reasoning conceptualization, with theories of identity threat, affirmation, and emotional reaction to identity threat. We choose IPC, which was originally developed to explain public disagreement about risk (Kahan et al., 2007) and scientific consensus (Kahan et al., 2011), as it was recently introduced to explain why individuals believe misinformation (Kahan, 2017). Moreover, although it has the individual at the center, it also incorporates cognition related to group belonging and social identity which the classical motivated reasoning literature has generally disregarded. However, especially in highly politically polarizing contexts, group belongings and social identity play a crucial role (Cohen, 2003).


By applying IPC, we asked participants to draw numerical inferences in two politically polarized contexts as well as how they experienced the inference-tasks concerning the emotions of anger, anxiety, and enthusiasm, emotions which have previously been closely related to political reasoning (Marcus et al., 2000). We

Action Editor: Danielle S. McNamara was the action editor for this article.

ORCID iD: Magdalena Wischniewski  <https://orcid.org/0000-0001-6377-0940>

Disclosure and Acknowledgments: This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media." Authors declare no conflicts of interest. All data are publicly available under <https://osf.io/8WT59/>.

Open Science Disclosures:

 The data are available at <https://osf.io/8WT59/>

Open Access License: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC-BY-NC-ND). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Disclaimer: Interactive content is included in the online version of this article.

Contact Information: Correspondence concerning this article should be addressed to Magdalena Wischniewski, Social Psychology: Media and Communication Department, University of Duisburg-Essen, Forsthausweg 2, Room LE 243, Duisburg 47057, Germany. Email: magdalena.wischniewski@uni-due.de

argue that if information threatens one's status within an affinity group, as IPC suggests, the identity threat elicits anger or anxiety. In turn, if incoming information affirms one's identity, enthusiasm will be elicited. As emotions have been shown to influence downstream information processing, we hypothesize that both the experience of anger and anxiety, as well as enthusiasm, mediate the relationship of IPC and the reasoning outcome. Our overarching objective is to contribute to a better understanding of cognitive and emotional processes in combating misinformation.

Theoretical Background

An Identity Approach to (Mis)Information Processing

With the increasing emergence of misinformation, Yeo et al. (2015) found that, especially in social media, the question of why people believe misinformative content has experienced a renewal of interest. Findings from previous studies reveal that belief in misinformation is either rooted in motivational processes (Chaiken et al., 1996; Kunda, 1990), cognitive biases such as confirmation bias (Nickerson, 1998) and myside bias (Mercier, 2016), biased memory retrieval (Taber & Lodge, 2016), or cognitive sophistication (Pennycook & Rand, 2019). Contrary to this, we focus on a different approach that places information processing in an identity-driven framework of motivated reasoning.

In *identity-protection cognition* (IPC), Kahan (2017) argues that beliefs and political views become “a badge of membership with identity-defining affinity groups” (p. 2). IPC considers it *individually rational* to reject information that opposes group beliefs, not solely individuals' beliefs, engaging in information processing that is “rationally suited to the ends of the agents who display it” (Kahan, 2017, p. 1). The main goal of an individual becomes then to protect one's status within the affinity group. Empirical results of recent studies support this claim by showing that the rejection of scientists as well as anthropogenic climate change is driven by social identity threats (Nauroth et al., 2017; Postmes, 2015).

Relating IPC to misinformation, Kahan (2017) argued that IPC's contribution is twofold. On the one hand, individuals accomplish their goal of expressing group belonging in selectively dismissing factual information while on the other hand crediting misinformation that confirms affinity-group identities. This argumentation is supported by studies that experimentally modified identity salience to increase misinformation rejection and increase genuine information acceptance, respectively (Wischnewski & Krämer, 2020).

Testing IPC empirically against deliberate and reflective information processing, Kahan et al. (2017) gave participants numerical information about the effect of a gun-ban and asked them to derive from the numbers the correct conclusion of the ban: either an increase or a decrease of crime. Unlike deliberative and reflective information processing would suggest, namely, that individuals with higher numeracy scores performed better in solving the task, they found that individuals were more likely to answer correctly if the answer confirmed their political identity.

Although Kahan et al. (2017) do not directly compare the acceptance of real and fake news like others have done (e.g., Pennycook et al., 2018) but rather investigate inferential conclusions about facts, we considered their approach fruitful: The actual validity of information put aside, it explains not only why people believe misinformation but also why some real information is rejected.

Moreover, in Kahan et al.'s (2017) study, the authors used numeric information and mathematical evaluations. One would expect that even strong identifications cannot change numerical inferences—after all, the numbers in their study allowed for only *one* correct interpretation. However, that was not the case. Although numbers unequivocally implied one answer, strong identifications had an impact on the answering behavior, even for individuals who were highly numerate. We argue that even supposedly undebatable arguments such as numbers can be perceived in a biased manner.

The theoretical groundwork of IPC draws on evolutionary psychology and utility maximation theory. According to the former, social groups fulfill humans' inherent need of belonging, protection, and safety, whereas the latter suggests that the benefits of conforming to group beliefs outweigh the costs of accepting group-inconsistent information. Empirical studies support IPC in the context of politics (Kahan et al., 2017) and risk perception (Kahan et al., 2007).

Attitudes or values that are transformed into a badge of group membership leading to IPC are manifold. However, in line with the original works on IPC (Kahan et al., 2017), our goal is to investigate politicized attitudes. We argue that this is appropriate given that political partisanship has previously been associated with group-belongings and social identity theory (Greene, 2004). Hence, we combine IPC with partisanship and social identity to suggest that political identity-incongruent information becomes identity-threatening, whereas political identity-congruent information becomes identity-affirming.

While the resulting bias of IPC and other theories of motivated reasoning within reasoning about politics is widely accepted, the moderating variable of cognitive sophistication yielded contradicting results. Findings by Kahan et al. (2017) indicated that cognitive sophistication increased identity-protection cognition and, hence, resulted in a stronger bias. However, recent investigations of this relationship found the opposite effect. Individuals with higher cognitive abilities showed less bias (Lind et al., 2018; Pennycook & Rand, 2019; Tappin et al., 2020). We position our investigation within the latter findings. In addition to this, IPC has been a source of criticism concerning its earlier label *cultural cognition* (used in, e.g., Kahan et al., 2011). Specifically, the notion of “cultural” bias has led to conceptual criticism (van der Linden, 2016). For our study, we do not want to argue for or against a notion of culture but rather follow IPC's argumentation of an identity-based approach to understand how people reason about (mis)information. In doing so, we follow Van Bavel and Pereira (2018) who adopt the idea of IPC in their identity-based model of beliefs. Their main argumentation proposes that “accuracy goals compete with identity goals to determine the value of beliefs” (p. 215). Examples of these identity goals are belonging goals, epistemic goals, status goals, or system goals which Van Bavel and Pereira connect to partisan identities. Similar to Kahan (2017), they argue that “maintaining beliefs and judgements that are aligned with one's political identity [. . .] is a higher priority than achieving accuracy” (p. 217).

Concludingly, we hypothesized, based on the above-reviewed literature on identity-protection cognition:

Hypothesis 1: If participants are asked to draw inferences from neutral stimuli, accurate inferences are predicted by participants' cognitive sophistication, not their political identity (see Figure 1 Hypothesis 1).

Hypothesis 2: If participants are asked to draw inferences from politicized stimuli, accurate inferences are predicted by participants’ political identity, not by cognitive sophistication (see Figure 1 Hypothesis 2).

Hypothesis 3: If participants are asked to draw inferences from politicized stimuli, the relationship between participants’ political identity and accurate inferences is moderated by participants’ cognitive sophistication (see Figure 1 Hypothesis 3).

(heuristic). In a later version of the HSM, Chaiken et al. (1996) revised this understanding, saying that accuracy goals are not always the primary driver of cognition but instead goals that preserve beliefs and self-concept (defense motivation) or impression management goals (impression motivation). This defense motivation has later been associated not only to the protection of the self-concept but also to favor in-groups (De Dreu et al., 2008). Concludingly, the defense mechanism becomes a central part of our theoretical understanding of IPC.

In the next section, we combine this defense motivation with emotional experiences related to identity threat.

The Central Role of Identity-Defense

The psychological defense mechanism against threats to self-integrity, that IPC uses, has long been discussed (e.g., Sherman & Cohen, 2002). For example, Chaiken (1987) argued in the heuristic-systematic model (HSM) that individuals’ primary motivation is to arrive at an accurate conclusion. Given sufficient cognitive capacities and motivation, individuals process information thoroughly (systematic), whereas they otherwise rely on mental shortcuts

Emotions and Identity-Protection Cognition

For the longest time, emotions have been regarded as corroding rational thought. In the last two decades, however, they found their way into many theories on human reasoning (Blanchette & Caparos, 2013; Jung et al., 2014; Ray & Huntsinger, 2017). Emotions have been introduced to political science (Brader et al., 2008;

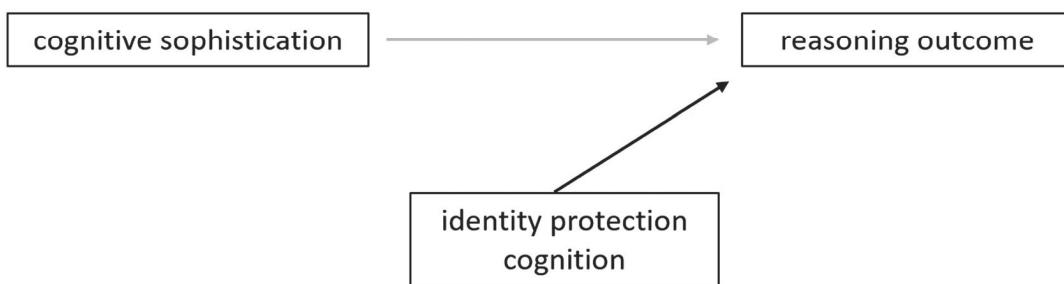
Figure 1

Visualization of Hypotheses 1–3. Hypothesis 1 Describes the Inferential Process for Neutral Stimuli, Whereas Hypotheses 2 and 3 Describe the Inferential Process for Politicized Stimuli in Accordance to IPC

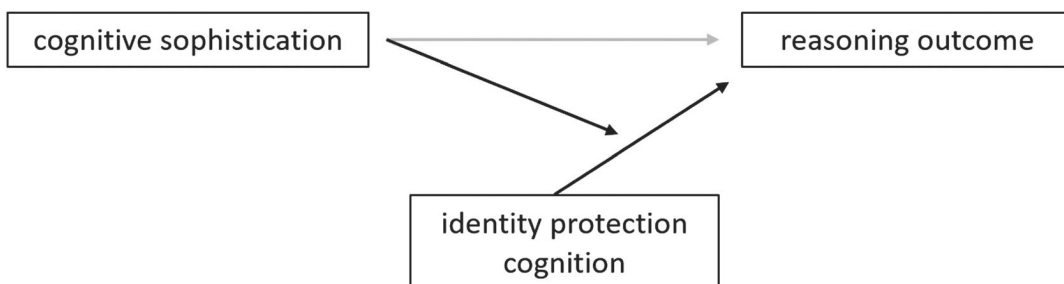
H1: Inferential process for *neutral* stimuli – the reasoning outcome is predicted by cognitive sophistication.



H2: Inferential process for *politicized* stimuli – the reasoning outcome is predicted by identity protection cognition, not cognitive sophistication.



H3: Inferential process for *politicized* stimuli – the reasoning outcome is predicted by identity protection cognition and moderated by cognitive sophistication.



MacKuen et al., 2010; Marcus et al., 2000, 2011), as well as studies on misinformation (Bakir & McStay, 2018; Brady et al., 2017; Van Damme & Smets, 2014; Weeks, 2015) and motivated reasoning (Lodge & Taber, 2013; Lord et al., 1979; Martel et al., 2019; Suhay & Erisen, 2018; Taber & Lodge, 2016).

Earlier research on the effects of emotion on cognition differentiated primarily between positive and negative emotions. Concerning social cognition, for example, it was found that positive moods induce top-down processing strategies, whereas negative moods induce a more systematic, stimulus-driven, bottom-up processing (Fiedler, 2001). In line with this, Schwarz (2002) suggested that emotions signal the level of required vigilance and effort, where negative states signal potential threat and positive states signal a safe environment which he called *cognitive tuning*. Later findings differentiate, however, between negative emotions, such as anger and anxiety. Contrary to previous findings, it was found that anger, as well as enthusiasm, led to a general overreliance on prior beliefs and superficial reasoning strategies (Huddy et al., 2007)—a top-down processing strategy. In the same lines, Weeks (2015) found that angry people were also “more likely to be motivated to defend their attitudes or partisanship” (p. 126). In turn, it was found that the discrete emotion of anxiety facilitates attention to available information and prompts thorough information seeking and processing (Brader et al., 2008). People who felt anxious were more likely to put their prior attitudes aside and to consider evidence in a balanced manner. These findings are consistent with the argumentation of affective intelligence theory (AIT) by Marcus et al. (2000). In AIT, Marcus and colleagues introduce two affective systems, the surveillance and the dispositional systems. While the former is alerted when an individual encounters new and unknown situations or information, the latter monitors habitual behavior. As part of the surveillance system, Marcus et al. propose that anger increases reliance on heuristics as well as enthusiasm. In contrast, anxiety facilitates the use of careful considerations.

Another strand of research regards emotions as additional information for the evaluation of information. The *feelings-as-information* hypothesis suggests that feelings are used to infer conclusions in a “how-do-I-feel-about-it” manner (Schwarz & Clore, 1983). Positive emotions signal positive evaluations, whereas negative emotions signal negative evaluations, respectively. Slovic et al. (2007) have argued that this reliance on affective cues gives individuals an

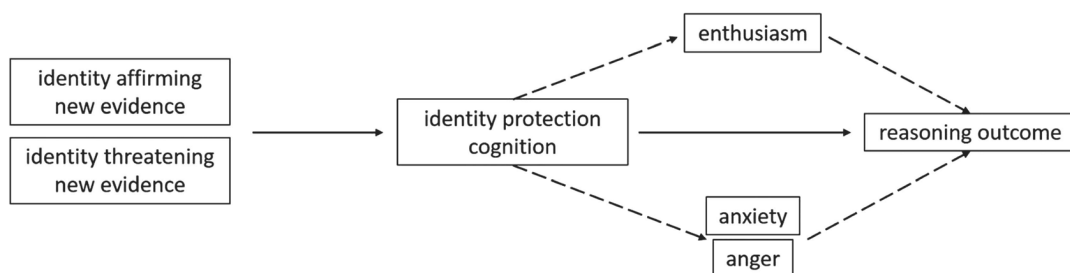
advantage over the more effortful and time-consuming reasoning processes of weighing pros and cons, independent of whether the affective cue is consciously perceived or not. They propose that the use of a mental shortcut, an affect heuristic, is especially likely “when the required judgement or decision is complex or mental resources are limited” (Slovic et al., 2007, p. 1336). In a complex inference task, it is, hence, more likely that feelings guide individuals’ processing. If a task holds identity-threatening potential, negative emotions such as anger or anxiety amplify this threat, leading to either avoidance or rejection of the given information, as IPC suggests. Positive emotions such as enthusiasm signal safety and approval of incoming information which is, in turn, used as additional information in the decision-making process.

This research aligns with Lazarus’ (1991) appraisal theory which proposes that any emotional state is preceded by a subjective evaluation of the event. Especially going beyond a dimensional understanding of emotions, discrete emotions like anger, anxiety, or enthusiasm require some degree of cognitive appraisal (Lodge & Taber, 2013). For the case of motivated reasoning which IPC relates to, Suhay and Erisen (2018) found, for example, that the discrete emotion anger successfully mediated the relationship between prior attitude and quality ratings and counter arguments of political campaigns.

Concerning IPC, as stated above, it is assumed that group-inconsistent information threatens an individual’s identity. Consequently, the individual reacts to the threat with identity-protective cognition which leads to the rejection of information, independent of its veracity. Previous studies on identity threat have shown, however, that identity threat elicited feelings of either anger or anxiety (Huddy et al., 2005; Huddy et al., 2015; Wischnewski & Krämer, 2020). We can expect that anger is elicited as a result of a perceived violation of one’s standards (Carver & Harmon-Jones, 2009), whereas anxiety is associated with a lack of personal control and uncertainty (Eysenck et al., 2007). Hence, if IPC is a reaction to identity threat, we can expect that this threat is accompanied by emotions integral to the situation like anger or anxiety (see dashed arrows in Figure 2). However, if incoming information is identity affirming, we can expect that enthusiasm is elicited due to identity affirmation (Marcus et al., 2000). In turn, we can expect that an individual, who identifies less with a certain issue, experiences less identity threat or identity affirmation. Hence, the individual should also be less biased in her/his reasoning.

Figure 2

Original Identity-Protection Cognition Path in Unbroken Lines. We Hypothesize That Dependent on an Individual’s Attitudes, Anger, Anxiety (for Identity Threat) or Enthusiasm (for Identity Affirmation) are Elicited (dashed Lines). In Turn, the Effect of Identity-Protection Cognition on the Reasoning Outcomes is Mediated by the Respective Elicited Emotions



In accordance with the discussed literature and empirical results, we theorize that emotional reactions are an integral part of IPC. We, therefore, hypothesize a mediating role of emotions:

Hypothesis 4: If participants are asked to draw inferences from politicized stimuli, the relationship of participants' political identity and accurate inferences is mediated by experienced anger, anxiety, and enthusiasm.

The Current Study

In this study, we investigated the role of emotions in identity-protection cognition to understand how people draw inferences from politicized information. Central to our study, we assume that the relationship between an individual's political identity and inference conclusions of politicized information is mediated by the experienced emotions anger, anxiety, and enthusiasm (H4). To assess this, we first hypothesize that inferring conclusions from nonpoliticized information are driven by an individual's cognitive sophistication skills (Hypothesis 1), whereas inferences about politicized information are driven by individual's political identity (Hypothesis 2). In addition, we hypothesized that the relationship between an individual's political identity and the inference conclusion for politicized information is moderated by cognitive sophistication (Hypothesis 3).

To test our hypotheses, we gave participants information about two politicized scenarios and one unpoliticized scenario and asked them to interpret the given information. Information was presented in a way to either confirm or reject one side of the selected politicized issue. Through self-report measures, we assessed participants' emotional responses as well as their cognitive sophistication.

Method

This study received ethical approval from the ethics committee of the Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen.

Sample

In an online experiment using convenience sampling, 463 (304 female) German citizens were recruited via different online platforms¹ (to arrive at our final sample size, we used the software G * Power version 3.1.9.4). The mean age of participants was 27.82 years ($SD = 8.64$), with most participants having a University degree (54%). Subjects who indicated so could participate in a raffle to win one out of overall 18 gift cards, worth between 10€ and 100€, as compensation. The data collection took place throughout late April to early June 2019.

General Procedure

After assessing individual cognitive sophistication skills, participants encountered three different fictitious scenarios in a randomized order. Within each scenario, they were confronted with a math task, asking them to draw inferential conclusions from numerical data. Immediately after each scenario, we asked participants to self-report their emotional reaction. The study closed with measures of political identity, basic demographic data, and a debriefing statement.

Pilot Study

To select potential polarized issues, we tested the threatening and affirming potential of seven controversies in a prestudy ($N = 64$). Controversies were selected through purposive sampling. A detailed report of the pilot study can be found here: <https://osf.io/8wt59/>. We selected the two most threat- and affirmation-inducing topics of the pretest, which were refugee intake and a driving ban for diesel cars. The former has been a highly discussed issue in multiple European countries and has induced IPC before (Lind et al., 2018). The latter relates to environmental concerns which have shown to induce biased reasoning as well (e.g., Hart & Nisbet, 2012; Kahan, 2013). Furthermore, the pretest results confirmed that issue positions (pro-refugee intake versus against, and pro-driving ban versus against) were closely related to specific political identities. However, positions for refugee intake were more pronounced and divergent than for a driving ban.

Experimental Design

In a 3 (neutral/polarized I/polarized II) \times 2 (increase/decrease) mixed experimental design, participants encountered three different fictitious scenarios in a randomized order. Constituting the first factor, in each of the three scenarios participants were asked to draw inferential conclusions: (a) Use of a skin cr me and the occurrence of a skin rash (neutral), (b) refugee intake and crime rates (polarized I), and (c) driving ban for Diesel cars and air quality (polarized II). Constituting the second factor, each of these three scenarios was randomly presented in a way to indicate either an increase or a decrease (between factors), resulting in overall six different scenarios. For example, in the neutral condition, the skin cr me could either increase or decrease the rash. In turn, for the polarized conditions, it meant that a refugee intake either increased or decreased crime rates as well as a Diesel car driving ban resulting to better (increase) air quality or worse (decrease).

Dependent Variable

General scenario structure, instructions, and data presented in each condition (see Figure 3 for an example) were adapted from Kahan et al. (2017) and translated to German. We asked participants to read the text as well as the numbers to indicate which conclusion they thought to be accurate, according to the data presented. The correct answer could be derived through inferential reasoning about data presented in a 2 \times 2 contingency table (see Figure 3). Hence, the dependent measure *task performance* was the answer given by the participants, which resulted in a binary measure (correct vs. incorrect).

Independent Variables

Political Identity

Similar to Kahan et al. (2017), we assessed participants' political identity through two different measures: a left-right scale and a progressive-conservative scale. Participants were asked to indicate on two 5-point Likert scales to self-identify as either political left or right and progressive versus conservative. High scores indicated far

¹ Different Facebook groups, Twitter, Survey Circle, eBay.

Figure 3

One of Six Experimental Conditions, Depicting the Inferential Task. Participants Were Asked to Read the Text as Well as the Numbers in the Table and Then to Indicate Which of the Statements Was Correct (Here Decrease). Statements Either Aligned With or Opposed Political Identities. to Create the Increase Condition, the Column Heads “Crime Rate Decrease” and “Crime Rate Increased” Were Swapped

Limited entry and crimes rates

Recently, a lot of controversy has risen from refugee intake and its effects on crime rates in German cities. Because opinions about this differ immensely, a large-scale study was conducted to find answers to this hypothesized correlation by the German Ministry for Migration. Refugee intake and crime rates for multiple communes within Germany have been measured. Among the communes, some have taken in a lot of refugees, whereas other have only accepted very few. The results of the study are stated in the table below. Although the two groups were not the same size, relative differences can still be detected.

	Crime rate decrease	Crime rate increase
Communes that took in many refugees	181	61
Communes that took in few refugees	87	17

Which indicate which of the following statements is correct:

- The crime rate decreased in communes that took in many refugees.
- The crime rate increased in communes that took in many refugees.

right and very conservative, respectively. Both measures were skewed toward the left and progressive side, which was found in German samples before (Bauer et al., 2017; $M = 2.43$, $SD = 0.75$). We joined both scales to one continuous political identity score, which proofed to show acceptable reliability (Cohen’s $\alpha = .71$).

Independent and Moderator Variable: Cognitive Sophistication

To assess cognitive sophistication, we followed operationalizations of previous studies (e.g., Lind et al., 2018). We assessed individual abilities through two different measures: the Computer Adaptive Berlin Numeracy Test (BNT) developed by Cokely et al. (2012) and the cognitive reflection task (Frederick, 2005). Both measures were added to one overall cognitive sophistication measure (Cohen’s $\alpha = .66$; see in Appendix Table A1 for all questions) with values ranging from zero to seven. Participants’ mean cognitive sophistication was $M = 3.21$ ($SD = 1.98$).

Mediator Variables: Emotions

In addition to the independent variable *political identity*, we implemented three self-report measure of affective reactions, analog to affective intelligence theory (Marcus et al., 2000), which have previously been used (Weeks, 2015): anger (angry, outraged, disgusted; Cohen’s $\alpha = .84-.93$), anxiety (afraid, anxious, nervous; Cohen’s $\alpha = .84-.86$), and enthusiasm (enthusiastic, hopeful, proud; Cohen’s $\alpha = .74-.89$). Affect judgments were given in percentages on a sliding scale from 0 (*not at all*) to 100 (*very much*), in whole integers.

Statistical Analysis

To account for the binary nature of the dependent variable, we conducted logistic regression analyses. In general, logistic regression estimates the probability of an outcome, in our case if participants answered correctly, via estimation of the log odds as a linear combination of the independent variables. Concerning levels of statistical significance, we followed the conventional alpha level of .05.

In the results section, we first entered cognitive sophistication, political identity as well as the control variables age, gender, and education into the regression model (Hypotheses 1 and 2). To account for the hypothesized moderation of cognitive sophistication, we included in a next step an interaction term of cognitive sophistication with political identity (Hypothesis 3). If the interaction coefficient is positive (negative) in the crime increase condition (crime decrease condition) and in the air quality decrease condition (air quality increase condition), Hypothesis 3 is supported

Finally, to assess the hypothesized mediating effects of anger, anxiety, and enthusiasm (H4), we used mediation analysis for binary-dependent variables according to Feingold et al. (2019) for all four politicized conditions. We used the PROCESS macro Version 3 (Hayes, 2017) for SPSS for all mediation models.

Results

An overview of the descriptive statistics of all variables can be found in Table 1 (dependent variables per scenario) and Table 2 (emotions per scenario). As described in the section *Statistical Analysis*, we performed binary logistic regressions for all six

Table 1
Descriptive Statistics of the Dependent Variable Task Performance Per Experimental Scenario

Dependent variable	<i>N</i>	<i>M</i>	<i>SD</i>
Rash increase	228	-0.05	1
Rash decrease	235	-0.03	1
Crime increase	225	-0.25	0.97
Crime decrease	238	-0.01	1
Air quality increase	242	0.52	0.85
Air quality decrease	221	0.18	0.98

scenarios (see Table 3 for numerical results and Figure A1 in the Appendix for a visual analysis).

Binary Logistic Regression for Nonpolarized Scenarios (Hypothesis 1)

As hypothesized in Hypothesis 1, cognitive sophistication was a significant predictor for the nonpolarized rash increase tasks, whereas the measure for political identity remained insignificant. However, in the second neutral tasks, the rash decrease condition, cognitive sophistication was not significant ($p = .057$). The positive regression coefficients for cognitive sophistication support the hypothesis that more numerate people were more likely to answer correctly (Hypothesis 1).

Binary Logistic Regression for Polarized Scenarios (Hypothesis 2)

When examining the politicized scenarios, the effects of cognitive sophistication were, as hypothesized (Hypothesis 2), mostly not significant. For the crime increase scenario, none of the suggested predictors reached significance. Instead, we noted that one of the control variables, gender, reached significance, indicating that men

Table 2
Descriptive Statistics of All Emotion Scores Per Experimental Scenario

Scenario	Emotion	<i>N</i>	<i>M</i>	<i>SD</i>
Rash increase	Anger	226	17.71	22.41
	Anxiety	225	15.55	20.09
	Enthusiasm	223	18.16	21.48
Rash decrease	Anger	234	20.53	23.95
	Anxiety	234	16.29	20.93
	Enthusiasm	234	17.15	22.46
Crime increase	Anger	222	14.55	19.92
	Anxiety	220	16.89	20.85
	Enthusiasm	223	29.02	25.49
Crime decrease	Anger	234	20.53	23.95
	Anxiety	238	19.98	23.02
	Enthusiasm	237	20.57	22.37
Air quality increase	Anger	240	22.49	25.41
	Anxiety	242	14.85	18.99
	Enthusiasm	242	22.6	21.15
Air quality decrease	Anger	218	22.49	25.41
	Anxiety	216	20.53	22.28
	Enthusiasm	217	18.21	22.54

were more likely to answer correctly than women. Nonetheless, in the crime decrease scenario, political identity reached significance which was hypothesized (Hypothesis 2). The negative regression coefficient indicated that people who self-identified as left and progressive were more likely to answer correctly, supporting the identity-protection hypothesis. Interestingly, for one of the driving ban scenarios, cognitive sophistication was a significant predictor. Again, the positive regression coefficient indicates that with increased cognitive sophistication, individuals became more likely to answer correctly, as was expected for the neutral rash conditions. Nagelkerke’s R^2 for all six models was between .044 and .098, which implied that the models could explain 4.5%–10% of the variance of our dependent variable, the response.

Moderation Analysis for Cognitive Sophistication (Hypothesis 3)

In the next step, we included an interaction term of cognitive sophistication and political identity in each model. Our results partly supported the hypothesis (Hypothesis 3). The interaction became significant only in the crime decrease condition ($b = -.24$, $p = .015$) but not in the crime increase and both Diesel ban conditions (see in Appendix Table A2 for regression coefficients). The nonsignificant results indicated that cognitive sophistication did not increase or decrease identity-protection cognition. The significant interaction of cognitive sophistication and political identity for the crime decrease condition supported Hypothesis 3, suggesting that cognitive sophistication decreases identity-protection cognition. Entering cognitive sophistication as a moderator revoked, however, the previously found effect of political identity in the crime decrease condition.

Mediation Effects of Anger, Anxiety and Enthusiasm

We hypothesized that the relationship between political identity and task performance is mediated by experienced anger, anxiety, and enthusiasm. As the results of the previous binary logistic regression analyses conveyed, we already knew that political identity and task performance are only associated in the case of crime decrease. We, therefore, focused on possible indirect effects of emotional experiences.

Results of the binary logistic regression analysis were generally confirmed. Political identity was only associated with the task performance in the crime decrease condition: participants with relatively left attitudes were more likely to correctly respond which supports the identity protection hypothesis. The respective relevant path coefficients per condition are displayed in Figure 4, while the respective results by emotion are reported in the next sections.

Anger

For the conditions of crime increase and air quality increase, political identity was a significant predictor for anger, whereas political identity and anger were not associated with the crime decrease condition and air quality decrease. Anger was a significant predictor for task performance only in the crime increase condition (see Figure 4a). We tested the significance of this effect using bootstrapping procedures, computing 5,000 bootstrapped samples with a confidence interval of 95%. The unstandardized indirect

Table 3
Regression Coefficients of the Logistic Regressions (Controlled for Age, Gender, and Education)

Predictor Variables	Rash condition		Refugee intake		Diesel ban	
	Increase	Decrease	Increase	Decrease	Increase	Decrease
Cognitive sophistication	0.21**	0.13 ⁺	0.05	-0.02	0.17*	0.12
Political identity	0.07	-0.08	0.21	-0.57**	-0.07	-0.26
Gender	0.65*	-0.19	0.9**	0.1	-0.06	-0.07
Nagelkerke's R^2	0.1	0.05	0.08	0.06	0.05	0.04

⁺ $p = .057$. * $p < .05$. ** $p < .01$. *** $p < .001$.

effect coefficient of anger was .19 with a 95% confidence interval ranging from $-.01$ to $.52$ (note, since the dependent variable response is binary, the metric of all effects is log-odds). Thus, the indirect effect of anger on the response was not significant, and the mediation hypothesis for anger was in no condition confirmed.

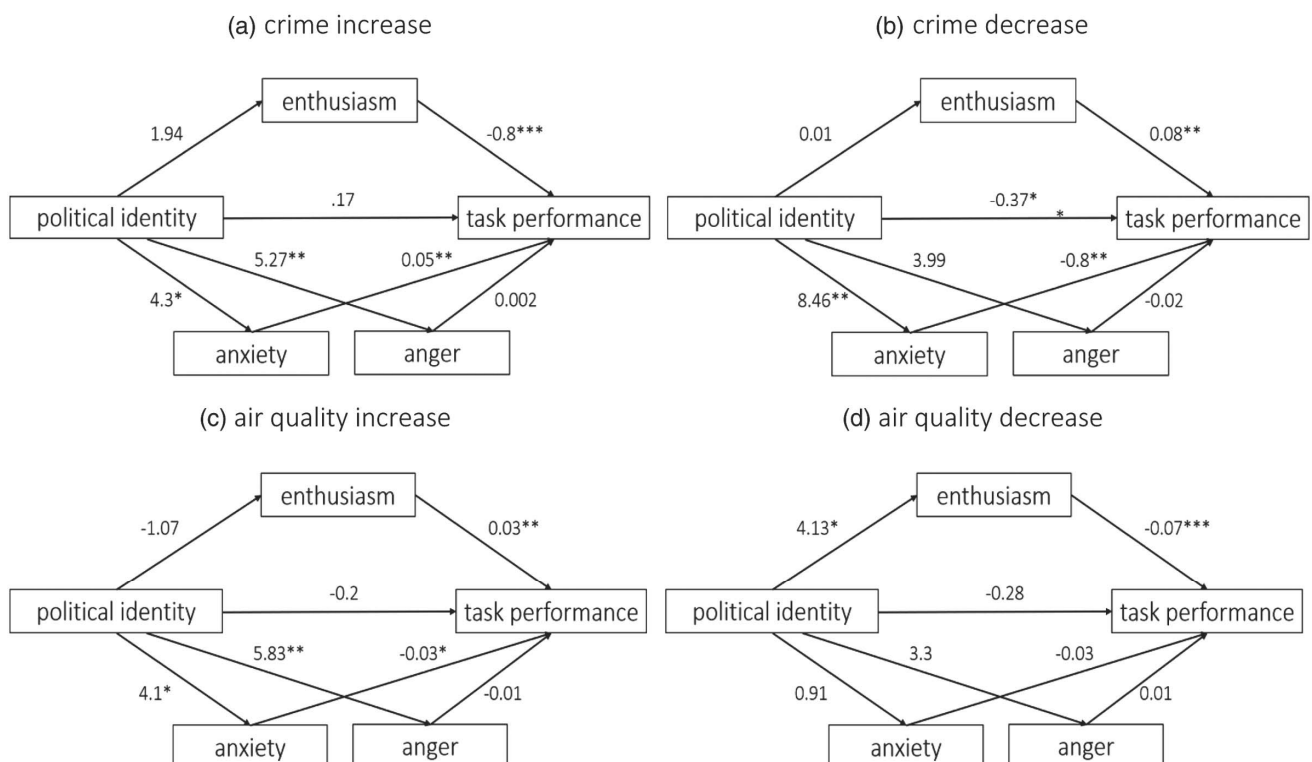
Anxiety

The picture for the experienced emotion anxiety was slightly different. Except for the air quality decrease condition, political identity was always a significant predictor for anxiety (see Figure 4d). In all conditions, anxiety was also a significant predictor for task performance, mainly indicating a negative relationship which implied that feelings of anxiety associated with political identity

were detrimental to the correct responding behavior. In order to gauge the significance of the indirect effects of anxiety, we used bootstrapping procedures for the respective cases. The unstandardized indirect effect coefficient of anxiety in both the crime increase and air quality increase condition was, however, not significant (crime increase = $.21$, CI $[-.04, .6]$; air quality increase = $-.13$, CI $[-.38, .02]$). In contrast to that, we found that for the crime decrease condition, the indirect effect of anxiety was significant ($-.69$, CI $[-2.13, -.17]$; see Figure 4b). The mediation hypothesis of anxiety must broadly be declined, apart from one condition: In the crime decrease condition, we found an indirect-only mediation of anxiety. This conveys that individuals, identifying with the political right, experienced higher levels of anxiety which, in turn, deteriorated task performance.

Figure 4

Mediation Analysis With Relevant Path Coefficients for the (a) Crime Increase Condition, (b) Crime Decrease Condition, (c) Air Quality Increase Condition, and (d) Air Quality Decrease Condition



Note. Significant path coefficients are marked as followed: * $p < .05$. ** $p < .01$. *** $p < .001$.

Enthusiasm

Political identity was generally not a predictor for feelings of enthusiasm except for the condition air quality decrease (see Figure 4d), whereas enthusiasm significantly predicted task performance throughout all conditions. Considering a possible mediation, we, therefore, looked only into the condition of air quality decrease. We found that the unstandardized, indirect effect of enthusiasm was significant, $-.28$, CI $[-.72, -.01]$, supporting an indirect-only mediation. Individuals from the political right experienced higher levels of enthusiasm which, in turn, deteriorated responses.

Discussion

In this study, we investigated the role of emotions in identity-protection cognition to understand how people draw inferences from politicized information. To do so, we relied on insights from motivated reasoning theories in social psychology and political science as well as advances in emotion research. Our central hypothesis assumed that the relationship between an individual's political identity and inference conclusions of politicized information is mediated by the experienced emotions anger, anxiety, and enthusiasm. To assess this, we first hypothesized that inferring conclusions from nonpoliticized information are driven by an individual's cognitive sophistication skills, whereas inferences about politicized information are driven by individual's political identity (identity-protection cognition hypothesis). In addition, we hypothesized that the relationship between an individual's political identity and the inference conclusion for politicized information is moderated by cognitive sophistication. We operationalized our hypotheses thematically in the field of refugee intake and driving bans for Diesel cars, both of which were highly politicized topics in Germany at the time of data collection.

Upon data analysis, we found only partial support for our hypotheses. Logistic regression analyses revealed that political identity did not predict task performance in three politicized conditions (Hypothesis 1), except one (crime decrease). Our data support IPC, therefore, only in the case of crime decrease. Overall, our results indicate that there is no bias related to opposing political identities. However, adding cognitive sophistication in the crime decrease condition as a moderator revoked this effect. The biasing effect of political identity was successfully moderated by individuals' cognitive ability, as we hypothesized (Hypothesis 3) and aligns with previous findings (Pennycook & Rand, 2019; Tappin et al., 2020). Concludingly, neither the political identity nor cognitive sophistication in isolation seemed to fully explain our data but rather the interaction of both.

Concerning our main research question, to investigate the role of emotions in identity-protection cognition, results were clearer, although not as predicted. Mediation analyses revealed that, whereas political identity was mostly neither associated with anger nor enthusiasm, we found a significant association of political identity and anxiety (see Figure 4b) which indirectly mediated the association of political identity and task performance. Mediation analyses also revealed that, unlike political identity and cognitive sophistication (as discussed above), emotional responses were related to task performance. In our data, we found that, other than expected, political identity did not predict identity-protection cognition but that, instead, emotional reactions determined responses. These

findings are in line with previous studies. Slovic et al. (2007) argued that people use their emotional responses as heuristics when tasks are complex, as can be found, for example, in evaluative priming (Hofmann et al., 2010). Results by Lind et al. (2018) indicated that individuals with higher numeric abilities (which we refer to as cognitive sophistication) showed less identity-protection cognition, arguing, in turn, that identity-protection cognition is more likely to be driven by emotions. Recent findings on fake and genuine news differentiation further support this argumentation. It was found that participants, when encouraged to rely on their intuition and gut feelings, were less likely to differentiate real from fake news (Martel et al., 2019). Because we did not find that emotions were consistently related to political identities, we assume that elicitation was somehow the result of political identities as well as the task content. This could have only been avoided if emotions were exogenously induced, as has been done before (Weeks, 2015), to a loss of external validity.

Limitations

There are several potential limitations to our study. First, the inconsistent relation of political identity and emotional reactions could be explained by a weaker association of self-identity and political identity. Although we tested how strongly political identities related to the selected topics (refugee intake and ban of diesel cars), the mere self-reported identification on a political spectrum might have reflected actual identification less well, especially considering the faceted nature of a multiparty system such as Germany. This might also explain why we found no relationship between political identities in the Diesel Ban scenario. While people may have held strong beliefs, these were not necessarily bound to a specific political identity but were rather associated with contextual factors that influenced an individuals' attitude (e.g., living in the countryside and being dependent on a car). Future studies could incorporate an identity scale like the four-item self-report measure, developed and tested by Bankert et al. (2017) or assess actual ego-involvement (see also Carpenter, 2019), to not only assess identification through a more direct measure but also accommodate for identity intensity.

Second, we want to remind the reader that our analyses rely on convenience sampling. We found that our sample was younger, more female, and more educated than a representative German sample, making it less clear how these findings generalize. However, the theoretical considerations and implications remain valid, despite missing representativeness. From a methodological view, we also want to point out that we followed the conventional standard to not correct for multiple testing. However, we are aware that calculating six regression analyses increases the chance for alpha error cumulation.

Third, it is theoretically unclear when identity threat evokes anger and when anxiety. It was previously found that a perceived violation of an individual's standards elicited anger (Carver & Harmon-Jones, 2009), whereas the latter is generally associated with a lack of personal control and uncertainty (Eysenck et al., 2007). Empirical findings are, however, mixed. While our results indicate a stronger association between political identity and anxiety, others have found anger to be strongly associated with prior opinions (Suhay & Erisen, 2018). Future studies could investigate if these differences are

merely measurement artifacts (identity measure versus opinion measure) or represent actual underlying psychological differences.

Fourth, we suggest that future studies could accommodate a cognitive psychological perspective on processing conflicting information. If we argue that identity threat is evoked, we assume that information inconsistent with one's own beliefs has been processed. This cognitive conflict should result into slower reaction times for stimuli that are inconsistent with prior-opinion which has been, for example, found in the Stroop task or the Simon task (Simon & Berbaum, 1990).

Implications for Misinformation Research in Social Media

Differentiating true from fake material online, and especially on social media, has become one of the greatest challenges in today's information society. Flagging and correcting misinformation (Flynn et al., 2017) for users is one way to reach accurate beliefs but has also shown to create new challenges such as *implied truth effects* where unchecked misinformation is considered validated (Pennycook et al., 2019). The ability to draw correct inferences from presented information is, hence, central to the constitution of accurate beliefs. Previous studies on misinformation have also shown that emotions are central to the language of misinformation (Bakir & McStay, 2018), its acceptance (Zollo et al., 2015), its spread (Vosoughi, Roy, & Aral, 2018), and its likelihood to be shared (Brady et al., 2017). Our study adds to this literature that emotional reactions indeed guide inferences, showing that not only identity-threatening information per se (as has been shown before: Kahan et al., 2017) but also identity-induced emotional reactions contribute to inaccurate inferences. To include these emotional reactions has already been addressed within the media literacy literature. For example, Sivek (2018) suggests mindfulness techniques to answer the palpable influence of emotion by, first, raising awareness of news exposure and, second, raising awareness for emotional responses. This relates to some degree to research on mood misattribution where individuals erroneously incorporated unrelated, incidental emotions into their judgments (Schwarz & Clore, 1983).

Considering the role of social media platforms in the proliferation of misinformation, applied research could investigate, for example, which information processing style might be more dominant when reading news on social media. Do contextual cues of social media feeds induce heuristic-driven information processing compared to traditional online news media? This would connect to phenomena like incidental news exposure (Kaiser et al., 2018) which describes the unintentional exposure to news via a user's social media feed and is driven by heuristics decision processes concerning news selection (Marewski et al., 2009).

General Conclusion

In this study, we investigated the role of emotions in motivated reasoning to understand how people draw inferences from politicized information. With increased misinformation dissemination, the skill to draw accurate inferences about politicized information has become more and more critical. From a psychological perspective, it is critical to understand the underlying cognitive and emotional processes of misinformation acceptance to advise policy

making in the best possible way. Our results align with previous findings on the role of cognitive sophistication, which can be considered a protection factor to decrease the likelihood of falling for misinformation. We did not find, however, consistent support for identity-protection cognition. More strikingly, our findings show that emotional processes, only partially related to political identity, drive the inferential processes.

References

- 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>
- Bankert, A., Huddy, L., & Rosema, M. (2017). Measuring partisanship as a social identity in multi-party systems. *Political Behavior*, 39(1), 103–132. <https://doi.org/10.1007/s11109-016-9349-5>
- Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology? Individual-level variation in associations with “left” and “right” and left-right self-placement. *Political Behavior*, 39(3), 553–583. <https://doi.org/10.1007/s11109-016-9368-2>
- Blanchette, I., & Caparos, S. (2013). When emotions improve reasoning: The possible roles of relevance and utility. *Thinking and Reasoning*, 19(3–4), 399–413. <https://doi.org/10.1080/13546783.2013.791642>
- Brader, T., Valentino, N. A., & Suhay, E. (2008). What triggers public opposition to immigration? Anxiety, group cues, and immigration threat. *American Journal of Political Science*, 52(4), 959–978. <https://doi.org/10.1111/j.1540-5907.2008.00353.x>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- Carpenter, C. J. (2019). Cognitive dissonance, ego-involvement, and motivated reasoning. *Annals of the International Communication Association*, 43(1), 1–23. <https://doi.org/10.1080/23808985.2018.1564881>
- Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin*, 135(2), 183–204. <https://doi.org/10.1037/a0013965>
- Chaiken, S., Giner-Sorolla, R., & Chen, S. (1996). Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 553–578). Guilford Press.
- Chaiken, S. (1987). The heuristic model of persuasion. In M. Zanna, J. Olson, & C. Herman (Eds.), *Social influence: The Ontario symposium* (pp. 3–39). Lawrence Erlbaum.
- Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, 85(5), 808–822. <https://doi.org/10.1037/0022-3514.85.5.808>
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgement and Decision Making*, 7(1), 25–47. <https://doi.org/10.1146/annurev.psych.49.1.447>
- De Dreu, C. K. W., Nijstad, B. A., & van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, 12(1), 22–49. <https://doi.org/10.1177/1088868307304092>
- De keersmaecker, J., & Roets, A. (2017). ‘Fake news’: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, 65, 107–110. <https://doi.org/10.1016/j.intell.2017.10.005>
- Druckman, J. N. (2012). The politics of motivation. *Critical Review*, 24(2), 199–216. <https://doi.org/10.1080/08913811.2012.711022>

- Schwarz, N., & Clore, G. L. (1983). How do I feel about it? The information function of affective states. In K. Fiedler & J. P. Forgas (Eds.), *Affect, cognition and social behavior: New evidence and integrative attempts* (pp. 44–63). C.J. Hogrefe.
- Sherman, D. K., & Cohen, G. L. (2002). Accepting threatening information: Self-affirmation and the reduction of defensive biases. *Current Directions in Psychological Science, 11*(4), 119–123. <https://doi.org/10.1111/1467-8721.00182>
- Simon, J. R., & Berbaum, K. (1990). Effect of conflicting cues on information processing: The 'Stroop effect' vs. the 'Simon effect'. *Acta Psychologica, 73*(2), 159–170. [https://doi.org/10.1016/0001-6918\(90\)90077-S](https://doi.org/10.1016/0001-6918(90)90077-S)
- Sivek, S. C. (2018). Both facts and feelings: Emotion and news literacy. *The Journal of Media Literacy Education, 10*(2), 123–138. <https://doi.org/10.23860/JMLE-2018-10-2-7>
- Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research, 177*, 1333–1352. <https://doi.org/10.1016/j.ejor.2005.04.006>
- Suhay, E., & Erisen, C. (2018). The role of anger in the biased assimilation of political information. *Political Psychology, 39*(4), 793–810. <https://doi.org/10.1111/pops.12463>
- Taber, C. S., & Lodge, M. (2016). The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Political Psychology, 37*, 61–85. <https://doi.org/10.1111/pops.12321>
- Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Psychology, 34*, 81–87. <https://doi.org/10.1016/j.cobeha.2020.01.003>
- Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication, 20*(5), 520–535. <https://doi.org/10.1111/jcc4.12127>
- 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 Damme, I., & Smets, K. (2014). The power of emotion versus the power of suggestion: Memory for emotional events in the misinformation paradigm. *Emotion, 14*(2), 310–320. <https://doi.org/10.1037/a0034629>
- van der Linden, S. (2016). A conceptual critique of the cultural cognition thesis. *Science Communication, 38*(1), 128–138. <https://doi.org/10.1177/1075547015614970>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- 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>
- Wischnewski, M., & Krämer, N. (2020). *I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption* [Conference session]. Proceedings of the 11th International Conference on Social Media and Society, Toronto, Ontario, Canada. <https://doi.org/10.1145/3400806.3400824>
- Yeo, S. K., Cacciatore, M. A., & Scheufele, D. A. (2015). News selectivity and beyond: Motivated reasoning in a changing media environment. *Publizistik Und Gesellschaftliche Verantwortung, 83*–104. <https://doi.org/10.1007/978-3-658-04704-7>
- Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozeti, I., Scala, A., & Caldarelli, G. Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLOS ONE, 10*(9), Article e0138740. <https://doi.org/10.1371/journal.pone.0138740>

Appendix

Table A1

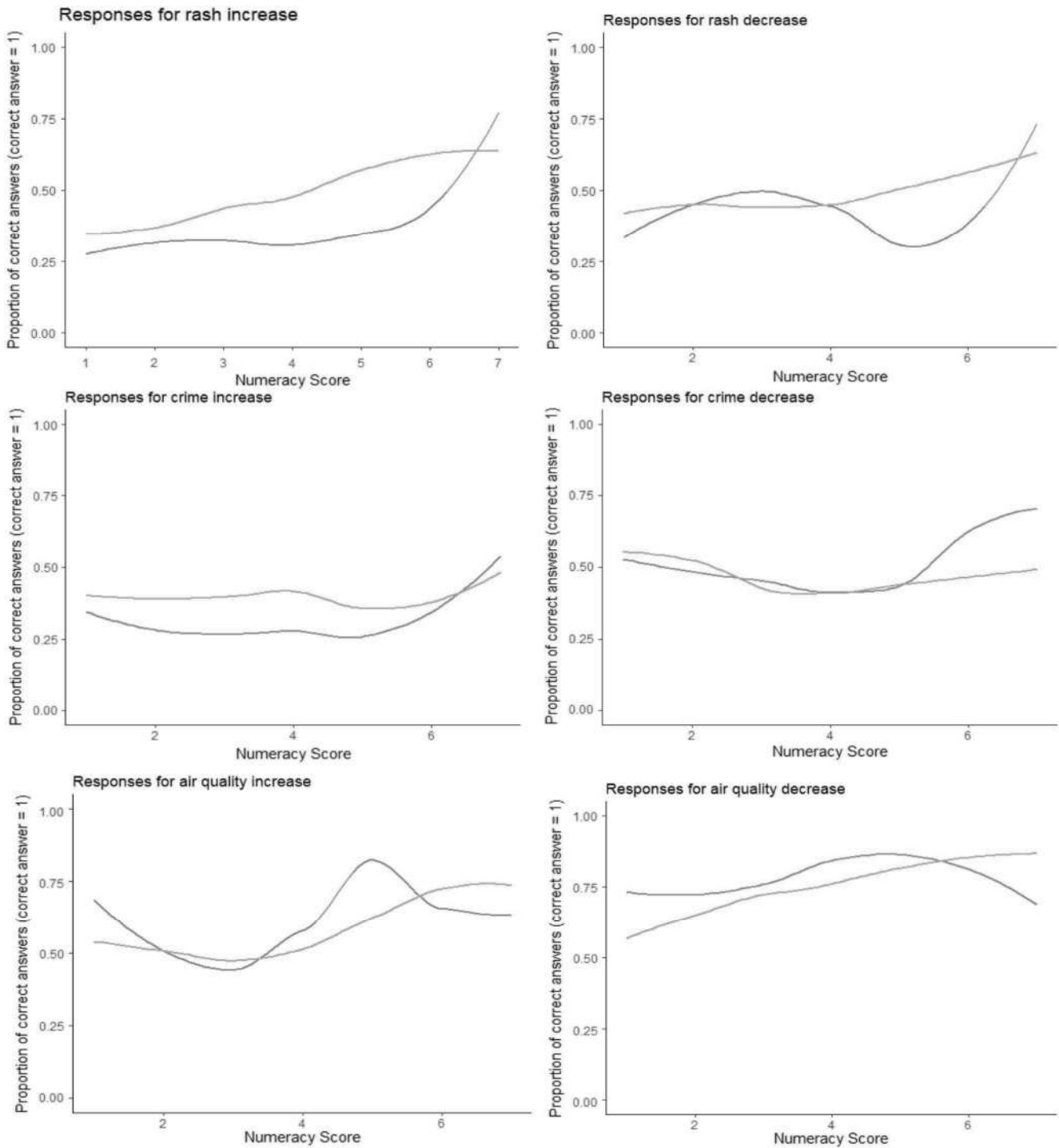
Questions of the Computer Adaptive Berlin Numeracy Test and Cognitive Reflection Test, Used to Assess an Overall Cognitive Sophistication Score

Questionnaire	Question	Correct answer
Computer adaptive berlin numeracy test	Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent.	25%
	Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3, or 5)?	30 out of 50
	Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?	20 out of 70 throws
	In a forest 20% of mushrooms are red, 50% brown, and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red?	50%
Cognitive reflection test	A bat and a ball cost 1,10€ in total. The bat costs 1,00€ more than the ball. How much does the ball cost?	0,05 €
	If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?	5 min
	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?	47 days

(Appendix continues)

Figure A1

Graphic Analysis of Each Task, Illustrating Task Performance by Cognitive Sophistication (Numeracy Score)



Note. Red Lines Indicate People From the Political Left, Whereas Blue Lines Indicate People From the Political Right.

(Appendix continues)

Table A2*Binary Logistic Regression Results for All Six Conditions, Including the Interaction Term Cognitive Sophistication*ideology*

Predictor Variables	Rash condition		Crime condition		Driving ban condition	
	Increase	Decrease	Increase	Decrease	Increase	Decrease
Cognitive sophistication	0.16	0.1	-0.02	-0.13	0.31**	0.13
Political identity	0.31	0.12	0.56	0.16	-0.83	-0.29
Cognitive sophistication*political identity	0.47	-0.07	-0.1	-0.24*	0.248	0.01
Gender	0.66*	0.2	0.94**	0.12	-0.01	-0.07
Nagelkerke's R^2	0.10	0.05	0.08	0.09	0.08	0.04

Received June 15, 2020

Revision received December 14, 2020

Accepted December 24, 2020 ■