Why Do People Spread Disinformation on Social Media? The Role of Social Identity and Perceived Morality

Joyner, Laura

A PhD thesis awarded by the University of Westminster.

https://doi.org/10.34737/w702w

Why Do People Spread Disinformation on Social Media?

The Role of Social Identity and Perceived Morality

LAURA CAMPBELL JOYNER

A thesis submitted in partial fulfilment of the

requirements of the University of Westminster

for the degree of Doctor of Philosophy

November 2023

Abstract

"Disinformation" is false or misleading information that is deliberately created or spread. In recent years, social media platforms have been used to rapidly disseminate disinformation for personal, political, and financial gain, or by those wishing to cause harm. Yet how far disinformation will digitally spread is also dependent on whether other users interact with it, regardless of whether they know it is inaccurate or not (i.e. "misinformation").

A series of five studies within the present thesis has therefore sought to understand if social media users are more likely to amplify the spread of misinformation and disinformation that allows them to express their identity and beliefs. It also investigated whether misinformation and disinformation are morally evaluated in the context of social identity, and whether any identity-related adjustments can help explain intentions to spread the content further.

Study 1 used a correlational design to explore whether degree of belief consistency influenced intentions to digitally interact (like, share privately, share publicly) with misinformation. Participants (N = 218) were presented with a series of 12 misinformation posts about the UK Government's handling of the COVID-19 pandemic (framed either "favourably" or "unfavourably") and misinformation about the risks of the COVID-19 virus (framed to either "minimise" or "maximise" risk). Related beliefs were also measured (i.e. trust in the UK Government's handling of the pandemic and perceived risk of COVID-19). Greater belief consistency predicted increased intentions to interact with misinformation. After informing participants the content was inaccurate, the degree of belief consistency also predicted the moral acceptability of spreading disinformation. The findings suggest that users may be more lenient towards and more likely to amplify the spread of inaccurate content when it is consistent with their beliefs about an issue.

Study 2 examined whether users would be more morally lenient towards misinformation or disinformation that may allow them to make favourable comparisons of their ingroup. London-based Conservative and Labour voters (N = 206) were recruited in the run up to the London mayoral elections in 2021. An experimental 2x2 between-groups design was employed, where participants were shown a social media post featuring fabricated information which either supported or undermined their own or the opposition party. Participants were more morally lenient towards spreading misinformation and disinformation that could help their ingroup (i.e. supported their ingroup or undermined an outgroup). However, exploratory analysis suggests that biased moral judgements of disinformation may have been driven by Conservative voters only.

A new scale was then tested and developed within **study 3** which incorporated digital actions that can potentially help reduce the wider spread of a post, as well as those which may amplify it further. This study replicated study one with the new scale (N = 251) and

showed that degree of belief consistency predicted the likelihood of contributing to the onwards spread of misinformation. It was also found that users may be more morally lenient towards spreading belief-consistent misinformation, and that such leniency can help explain (but does not entirely mediate) the relationship between belief consistency and spread.

Study 4 used a 2x2 between-groups design to test the effect of message framing (i.e. positive or negative) and fact-check tags on moral evaluations of misinformation and understand how any moral leniency may influence intentions to spread. Supporters of 5 English Premier League teams (N = 262) were recruited and shown inaccurate posts about their own team. Moral judgments were again biased in favour of the ingroup, even when participants were aware the information was untrue, and helped to explain increased intentions to spread the content further. Participants also provided written explanations to support their responses which were analysed against the Extended Moral Foundations Dictionary. The computational text analysis indicated that engagement with "fairness" related values differed across the four conditions. Specifically, participants were least likely to consider fairness when presented with positively framed ingroup misinformation, and this reduced consideration of fairness was related to increased moral acceptance of posts generally. Moreover, despite the content being unrelated to politics, political asymmetry was again observed in moral judgements of ingroup supporting disinformation. The findings indicated that politically left-leaning participants may have been more likely than others to consider fairness when making evaluations of identity-affirming disinformation.

Finally, two moral reframing interventions were developed in **study 5** which aimed to help reduce intentions to spread identity-affirming misinformation. These appeals framed the sharing of unverified content as violations of individualising moral values (fairness, harm) or violations of binding moral values (loyalty, authority, sanctity) and were tested alongside a pre-existing accuracy nudge intervention in a 3x2x2 between-groups design. Democrat and Republican voters (N = 508) were recruited in the run up to the 2022 US mid-term elections and shown political misinformation that positively compared their own party to the opposition. Both moral appeals were more effective at reducing moral acceptability and intentions to spread misinformation than pre-existing accuracy nudge interventions, but only in Democrat voters. The findings indicate that accuracy nudges may dissuade strong identifiers from amplifying misinformation further but have no influence on moral evaluations. In contrast, any reduced intentions to spread misinformation after viewing a moral appeal may be explained by adjustments to the perceived moral acceptability of spreading the content further.

Together, this research demonstrates that moral evaluations of misinformation and disinformation are situational and change in relation to the viewer's social identity. The present thesis also provides insight into the role of moral cognition in influencing decisions to spread misinformation and disinformation. It also may help explain why certain users may appear more susceptible to spreading inaccurate content generally.

Table of Contents

List of Tables

List of Figures

List of Appendices

Publications & Conferences

Parts of the research reported in this thesis have appeared in the following forms to date:

Journal Articles

Joyner, L. C., Buchanan, T., & Yetkili, O. (2023). Moral leniency towards belief-consistent disinformation may help explain its spread on social media. *PLOS ONE*, *18*(3), e0281777. https://doi.org/10.1371/journal.pone.0281777

Selected Presentations

Joyner, L. C., Buchanan, T., & Yetkili, O. (2023, July 24-25). *Political orientation may influence moral judgements of disinformation, but only when people know it is untrue* [Paper presentation]. British Psychological Society - Cyberpsychology Section Conference, Northumbria, United Kingdom.

Joyner, L. C., Buchanan, T., & Yetkili, O. (2022, October 26). *Political ideology may help explain group-differences in moral evaluations of disinformation* [Paper presentation]. Political Psychology Conference, University of Westminster, London, United Kingdom.

Joyner, L. C., Buchanan, T., & Yetkili, O. (2022, September 22-23). *Flexible moral judgements and the spread of belief-consistent disinformation on social media* [Paper presentation]. British Psychological Society - Cyberpsychology Section Conference, Brighton, United Kingdom.

Joyner, L. C., Buchanan, T., & Yetkili, O. (2022, September 5-7). *Making exceptions for belief consistent content: Moral psychology & spreading disinformation online media* [Paper presentation]. British Psychological Society - Social Psychology Section Conference. London, United Kingdom.

Joyner, L. C., Buchanan, T., & Yetkili, O. (2021, August 29-30). *Group identity and event-related beliefs influence potential sharing and moral judgements of disinformation* [Paper presentation]. Psychology Postgraduate Affairs Group Conference, Online.

Joyner, L. C., Buchanan, T., & Yetkili, O. (2021, April 30). *Issue-related beliefs influence moral judgements and potential sharing of disinformation on social media* [Paper presentation]. Disinformation, Language, and Identity Workshop, Cardiff University, Online.

Acknowledgements

Firstly, I would first like to express my sincerest gratitude to my supervisors: Professor Tom Buchanan and Dr Orkun Yetkili. I feel so fortunate to have gone through this journey with such a brilliant team. This would not have been possible without your invaluable guidance and support. Thank you both for your kindness, enthusiasm, and above all your belief in me.

I would also like to thank the University of Westminster for awarding me the Research Studentship in Psychology and for supporting other financial aspects of this PhD through the Globally Engaged Researcher and Psychology PhD Funds.

Beginning this venture during the midst of a pandemic presented a host of unique challenges. However, I luckily had the privilege to cross paths with many wonderful people during this time; both at Westminster and beyond. I therefore want to say a huge thank you to my PhD community. Special thanks go to Charlie, Giota & Kathryn for your ongoing support and friendship over the past three years (both in the early Zoom days and the fish tank). Also to Amy, Pippa & Tash for their wonderful advice and check in texts (especially in recent months). This journey would not have been the same without you all and I feel very fortunate to be able to call you my friends.

Thanks also to the many staff members at Westminster who have helped me to feel welcome over the past three years. Whether you took the time to attend a presentation run through, offer sage advice or simply asked how I was; it was appreciated. Notable mentions go to Rotem (for your encouragement both in terms of work as well as getting me to take much needed coffee breaks), Dave & Karen, Kathryn W (your Friday drop-in calls made the challenge of starting a PhD remotely feel possible), and Haulah & Lejla.

I am also thankful for all of my very patient friends from “the outside world” who have been so supportive over the past few years. I appreciate everyone who has checked in, asked how things are going (and rarely asked when my PhD will be done) and sent motivational messages along the way. Special thanks go to Catriona & Steph for their constant encouragement and to Celia for taking me under your wing at UEL and showing me this was even possible.

Last but by no means least, I would like to thank my family. To Mum, Dad & Sandy, thank you for always being there when I need you, for being so supportive of my career change, and for the help with proofreading in recent months! Finally, thank you to Andrew for always reminding me to back up my work, looking after me through the highs and the lows, and for constantly being my cheerleader. I could not have done this without you by my side.

Author's Declaration

I declare that all the material contained in this thesis is my own work and has not been submitted to any other University.

Laura Campbell Joyner, August 2023

Chapter 1. Introduction

1.1. Background and Rationale for the Research

Disinformation is false or misleading information that is created or spread with the intention to mislead others, for instance to cause harm, or for personal, political, or financial gain (Digital Culture Media and Sport Committee, 2019). Although disinformation is in no way a new phenomenon, its spread has historically been aided by advances in communication technology (Burkhardt, 2017), including the creation of the internet. In recent years, the spread of false and misleading information on social media platforms (SMPs) has been a growing concern. Estimates suggest around 60% of the global population are active social media users (DataReportal, 2023). Therefore SMPs provide environments where large proportions of the population may be reached on their personal devices from remote locations for relatively little cost.

Indeed, while technically anyone can create disinformation, it is also spread as part of professionally orchestrated campaigns that have the potential to recruit significant attention from users. For instance, 61,500 Facebook disinformation posts created by Russia's Internet Research Agency (IRA) to target Americans in the years surrounding the 2016 US presidential election collectively received 77 million interactions, and were reported to have been seen by 126 million people (DiResta et al., 2019). For context, this is equal to reaching more than a third of the US population. Notably, this campaign employed complex networks of faux accounts (spread within and across SMPs), each designed to target specific groups across the US population with carefully curated messages and content. While some posts received little to no attention, other content posted by their most popular accounts received tens-to-hundreds of thousands of interactions suggesting that many users were willing to interact with their posts. Moreover, analysis of disinformation disseminated surrounding four terrorist attacks in the UK in 2017 discovered 475 Tweets posted by 47 IRA run accounts (Innes, 2020). These tweets

(which targeted users at either end of the ideological spectrum) were subsequently reposted by other users more than 153,000 times. Social media users therefore arguably play a key role in amplifying the spread of disinformation.

Notably, the real-world impact of disinformation is itself difficult to measure (e.g. Colley et al., 2020) and therefore goes beyond the scope of the present thesis. However, previous disinformation campaigns have been linked to election interference (Digital Culture Media and Sport Committee, 2019), market disruption (S. C. Johnson, 2013), attempts to amplify and sow societal division (DiResta et al., 2019), and even genocide (Amnesty International, 2022). To help prevent disinformation from having a serious impact, it is therefore vital to reduce its spread. This thesis therefore focuses on how users contribute to the spread of disinformation content once it has been posted within a social media platform. The remainder of the present chapter will begin by briefly outlining how the technological and commercial features of social media platforms help create an environment where disinformation can be spread with relative ease. An overview of some of the key disinformation research within psychology is then provided followed by an introduction to Social Identity Theory. The research objectives and questions are stated, before an outline of the thesis as a whole.

1.1.1. Gaming Algorithms to Amplify Disinformation Spread

"Sharing" is not the only means by which users can spread disinformation. Indeed, the digital architecture of many SMPs can arguably force users to play a central role in the wider dissemination of disinformation (whether intentionally or otherwise). Not only do the major SMPs present users with feeds featuring personally relevant content, the digital interactions that users make act as ranking signals indicating where said content should be subsequently placed within the algorithmically ordered feeds of other users (Lada et al., 2021). Therefore, each interaction contributes to a ranking algorithm, signalling to the

SMP whether the content is something others may find interesting. As such, relatively effortless actions such as “liking” will influence the total reach of content.

These same mechanisms are “gamed” by bad actors who seek to disseminate disinformation. Indeed, disinformation campaigns often utilise highly sophisticated dissemination strategies to appear on users’ feeds, such as those used by digital marketers in legitimate organisations. For instance, previous campaigns have utilised “micro-targeting” advertisements (Committee On Intelligence United States Senate, 2019; Digital Culture Media and Sport Committee, 2019), hashtag strategies (DiResta et al., 2019), influencer recruitment (Alderman, 2021), audience segmentation (François et al., 2019), and community management (DiResta et al., 2019; Nimmo, François, Eib, & Tamora, 2020). As a result of disinformation disseminators harnessing technological and commercial features of SMPs, users do not need to follow specific accounts for disinformation to appear on their feed.

While reliance on user-interactions to amplify disinformation does of course mean that posts need to appeal to the users who initially see it, recommender algorithms allow SMPs to show users relevant content based on their previous activity (Kalimeris et al., 2021). As in, people are more likely to be presented with content that is similar to content which they have already interacted with. Arguably then, the disinformation encountered on social media is likely to be more personally relevant to the viewer than disinformation encountered in the offline world (e.g., via printed flyers, television, etc). Furthermore, if users then go on to interact with this digital disinformation, then, as previously discussed, they also assist in amplifying its total reach. Given users play a central role in the amplification of disinformation, it is therefore important to understand why they make these digital interactions in the first place. The following section provides a brief overview of some of the key research to date.

1.1.2. Current Disinformation Research in Psychology

A key focus of psychological research in this area relates to whether or not users believe disinformation. Indeed, while some users may spread disinformation intentionally (as in, are aware that the information is untrue) others may be unaware of any inaccuracies. This is known as "misinformation" (Digital Culture Media and Sport Committee, 2019). Previous work has suggested that people may believe misinformation if it supports their political beliefs (e.g. Kim et al., 2019) or if they have been repeatedly exposed to it (e.g. Pennycook et al., 2018). In turn, people are more likely to spread misinformation that they believe (e.g. Buchanan, 2020), and so an inability to identify inaccurate information may present a barrier for reducing disinformation spread. However, recent research also indicates that people may be more morally accepting of sharing disinformation that "feels" true, even when they know it isn't (Effron & Raj, 2020). The first research objective of this thesis is therefore to determine the relationship between identity-related beliefs on perceptions and intentions to spread false and misleading information on social media.

Notably, people are also more likely to believe misinformation that supports their ingroup, although whether this is due to motivations to protect identity (e.g. Kahan, 2015) or because of "lazy thinking" (as in, not engaging with more strenuous thinking processes (e.g. Pennycook & Rand, 2019)) is unclear. At the same time, people do report caring that the information they share online is accurate (Pennycook, Epstein, et al., 2021), and may also refrain from spreading disinformation to protect their reputation (Altay et al., 2020). However, as research has found people may be more lenient about lying when it is seen to be pro-social (Levine & Schweitzer, 2014), it may be that refraining from spreading disinformation may be less to do with the content being "inaccurate" and more to do with perceiving spreading inaccurate information as "wrong to do" in specific contexts (such as when the content undermines the ingroup). Another objective of this thesis is therefore to understand how moral evaluations of spreading disinformation on social media are influenced by social identity.

1.1.3. Social Identity Theory

Social identity approaches focus on "the group in the individual" (Hogg & Abrams, 1998, p. 17), for instance, how group membership influences the decisions and behaviours of an individual. Group memberships form an important part of an individual's self-concept, providing not only information about themselves but also how they relate to other people (Hogg & Abrams, 1998). However, said social identity may only become salient in specific situations, such as in response to relevant stimulus cues and social contexts (Carvalho & Luna, 2014). For instance, presenting a user with content about their ingroup (e.g. political party) on their SMP feed may make said social identity salient. As a result, decisions about the content may be made in the context of said ingroup and may therefore differ from decisions made in relation to content about another ingroup (e.g. sports team).

Social Identity Theory (SIT) also proposes that individuals are motivated to achieve or maintain a positive self-concept (Tajfel & Turner, 2004). Specifically, people are motivated to perceive relevant ingroups as positive distinct as their evaluations of group membership are connected to their self-esteem. Three key strategies help people achieve this in relation to group memberships: social competition (e.g. ingroup bias), social creativity, and individual mobility. Social competition and social creativity strategies will mostly involve attempts to "positively differentiate" (or, at the very least, differentiate) an ingroup. This helps to ensure that the boundaries of said group are clear, and ideally position an ingroup above an outgroup on some dimension (Tajfel & Turner, 2004). Therefore, acts such as expressing unique attributes of an ingroup, or framing an outgroup in less-favourable terms may assist individuals in maintaining a positive social identity.

Such strategies are particularly relevant in the context of user-behaviour within SMPs. Research suggests that core motivations for spreading memes (a form of user-generated content) include self-expression and social identity (Leiser, 2022). For instance, digital interactions with memes may relate to identity-construction (Aronson & Jaffal, 2022; Ask & Abidin, 2018; Bucknell Bossen & Kottasz, 2020; DeCook, 2018) and the

sharing of beliefs and opinions (Leiser, 2022). Notably, certain (collectively held) beliefs allow individuals to express group membership, and help to define group boundaries (Bar-Tal, 1998). The expression of these "group beliefs" also signal group membership to others, and therefore can hold a unifying purpose. Indeed, Leiser (2022) found that sharing memes allowed individuals to feel a sense of connectedness with like-minded others. Unfortunately, disinformation disseminators are not only aware of this; they have been known to actively produce and copy user-generated content (including memes) that could facilitate such identity expression, and strategically disseminate them to reach specific social groups (François et al., 2019).

The motivation to maintain a positive self-concept can also help explain individuals' reactions when faced with identity-related threats. For instance, people may question the credibility of information that threatens the value of an ingroup, e.g. a group-directed threat (Ellemers et al., 2002). Moreover, they may engage in normative behaviour if alternatives could lead to exclusion from an ingroup (a self-directed threat). For instance, if sharing certain information (e.g. a social media post) within a social context (e.g. an SMP) violates social norms, individuals may refrain from doing so. Thinking or behaving in ways that protect a specific social identity from potential threats can therefore help individuals to maintain a positive self-concept. Under the right circumstances, identity-protective responses may arguably also occur in response to disinformation and misinformation also. Therefore, the final research objective is to understand whether such threats to identity influence evaluations and intentions to spread.

1.2. Research Objectives and Questions

The impact of social identity on users' moral evaluations of disinformation has not been widely explored. The research presented here aims to explore the influence of social identity on users' evaluations and intentions to spread misinformation. Specifically, the research objectives are to:

1. Determine the relationship between identity-related beliefs and how individuals perceive and spread related disinformation / misinformation through a series of online correlational and experimental studies (RQ1).
2. Investigate whether moral judgements of disinformation / misinformation are context dependent and / or influenced by social identity (RQ 2, 3, 4).
3. Understand whether threats to social identity influence interactions and judgements of disinformation (RQ 3, 4).

The present thesis therefore addresses the following four research questions:

1. Are individuals more likely to contribute to the spread of disinformation (e.g. through digital interactions or inaction) when the message appeals to group-related beliefs, attitudes or values? (Studies 1, 3)
2. Do moral judgements of spreading identity-related disinformation differ according to the content's potential impact on achieving positive distinctiveness? (Studies 2, 4, 5)
3. Do perceived identity threats influence the moral judgements of spreading identity-related disinformation? (Studies 4, 5)
4. Do moral processes play a role in user contributions to disinformation spread? (Studies 3, 4, 5)

1.3. Outline of Thesis

The present chapter has provided context and outlined the objectives and research questions for the thesis, which consists of nine chapters in total.

Chapter Two provides an overview of the background literature on people's perceptions of disinformation, their ability to identify disinformation, reasons for interacting with disinformation, and the influence of social identity on disinformation

spread. Chapter Three then outlines several of the core methodological decisions and approaches used in this thesis. Here, the epistemological position is discussed, justifying the quantitative approach adopted throughout the five studies. The methodological approach is also outlined, addressing the use of internet research methods, approaches to statistical analysis, and engagement with open research practices.

Chapter Four contains the first research study, looking at whether SMP users are more likely to "interact" with "real-life" misinformation when it is consistent with their beliefs. Participants were also asked how morally acceptable they felt sharing the content was after learning it was untrue. Additionally, two sets of misinformation stimuli featured in this study focused on opposing beliefs about the UK government, while two sets focused on the risk of COVID-19. To understand how people use and evaluate misinformation in the context of their identity, exploratory analysis therefore compared responses given by Conservative and Labour voters.

Chapter Five contains the second study, an experimental design where Conservative and Labour voters were presented with "misinformation" that supported or undermined either their own political party (e.g. ingroup) or the political opposition (e.g. outgroup). Participants were asked how morally acceptable they felt sharing their assigned post was, before and after learning it was untrue. For exploratory analysis, they also rated how likely they were to "report" the content to a platform, which formed the basis for study three.

Chapter Six contains the third study, which was a partial replication of study one. One aim of this study was to test a "Social Media Spread" scale (created for this thesis) that incorporated participatory SMP interactions which may amplify or reduce the spread of content. Moral acceptability of spreading posts were collected before learning they were untrue, allowing the relationship between moral acceptability and spread to be analysed.

Chapter Seven contains the fourth study, an experimental design where fans of five English Premier League teams saw posts that either supported or undermined their team,

and also contained a “fact-check tag” (e.g. disinformation) or did not (e.g. misinformation). Notably, selecting social groups unrelated to politics here helped to rule out any confounding effects related to political ideology. Intentions to spread and moral acceptability responses were collected using both scales and free text (from which moral domain was quantified using linguistic content analysis. Responses in each condition were also analysed in relation to identity-strength, moral foundations, and political orientation.

Chapter Eight contains the final study, where interventions that tailored moral appeals to “individualising” and “binding” foundations were tested alongside “accuracy nudges” to understand their potential impact on moral evaluations and intentions to spread identity-benefitting misinformation. The effects of political orientation and identity-strength were also analysed.

Chapter Nine discusses the key findings from the present research in relation to the individual research questions, as well as findings related to political ideology and moral foundations theory. This is followed by Methodological implications, and implications for current interventions designed to help reduce user-contributions to disinformation spread. Limitations and recommendations for future research are also addressed.

Chapter 2. Literature Review

This chapter aims to provide an overview of the existing literature on disinformation in the context of social media users. It begins with an outline of the search strategy. The literature review is itself divided into four sections. First is a review of the literature related to perceptions of disinformation, specifically; what people interpret disinformation to be, whether they feel it is right or wrong to spread, and whether they perceived themselves to be vulnerable to it. Next, research on whether people are able to identify disinformation is explored, including factors which may aid or undermine their ability. This is followed by a review of the literature about why people might make digital interactions with disinformation. Finally, research related to the influence of social identity on disinformation spread is considered. This includes the role of the audience, strategies for achieving positive distinctiveness, and the impact of identity threat on disinformation spread.

While both "disinformation" and "misinformation" refer to false or misleading information, each has a specific meaning. Disinformation is created and spread with an underlying intention to deceive, either to cause harm or for personal, political, or financial gain (Digital Culture Media and Sport Committee, 2019). However, when people unknowingly spread false information, this is referred to as "misinformation". Empirical chapters within this PhD will distinguish between the two but, due to the nature of research in this area, for the purpose of this literature review the terms may be used interchangeably here.

2.1. Search Strategy

The initial searches for this literature review took place through July-September 2021 using the PsycInfo database. These focused on peer reviewed journals exclusively. To ensure relevance, it was also specified that search terms must appear in the abstract. At the time of searching, PsycInfo had 1,742 peer-reviewed papers that contained the words

"Disinformation", "Misinformation" or "Fake News" in the abstract. As the purpose of this thesis is to understand why individuals spread disinformation on social media, a series of search strings were defined to narrow down this search. This approach was taken for a number of reasons. As the 2017 word of the year, "Fake News" appears in many abstracts unrelated to disinformation research specifically. Moreover, there is an ever-growing supply of research about disinformation, however, this is not always specific to humans and their digital interactions with it. Therefore, searches were designed with four questions in mind: what do people think about disinformation; are they able to identify disinformation; what leads them to interact with disinformation on social media; does social identity have an influence on the online spread of disinformation.

Two of these questions were also allocated a series of search goals to ensure the searches picked up all relevant papers. Firstly, aspects of the risk evaluation process may be valuable for understanding what people think about disinformation. Therefore, the Threat Vulnerability Consequence framework (Willis, 2005) was used to help establish additional search terms. For example, measuring "threats" require specificity (Willis, 2005) and so identifying what individuals actually understand disinformation to be may be important. Next, as initial searches relating to social identity and disinformation produced only 8 results, broadening the search goals to include terms related to identity expression, partisanship, group regulation (including self-regulation in relation to groups) and social comparison strategies helped further ensure any potentially valuable papers would not be missed. An outline of all search strings can be found in Table 2.1.

In total, the search produced 544 results. After deduplication and screening this was narrowed down to 42 papers. These papers were then supplemented with items found elsewhere (e.g. Web of Science, Google Scholar, Research Rabbit). New papers were regularly monitored for following September 2021, and relevant work published after this date has been incorporated into the literature review and empirical chapters as appropriate.

Table 2.1

List of Search Strings Used for Literature Searches

Search goal	Keywords 1	Keywords 2	Keywords 3	Exclusions	PsycInfo #
		What do people think about disinformation?			
Citizen understanding of disinformation	(Disinformation OR Misinformation OR "Fake News")	(conceptuali* OR interpret* OR understand*)	(Public OR Citizen OR participant*)	NOT "Public Health"	130
Feelings about disinformation	(Disinformation OR Misinformation OR "Fake News")	(moral* OR approv* OR disapprov* OR unethical OR severity)			55
Perceived vulnerability to disinformation	(Disinformation OR Misinformation OR "Fake News")	(vulnerab* OR "third-person effect")			76
		Are people able to identify disinformation?			
Ability to spot disinformation	(Disinformation OR Misinformation OR "Fake News")	(spot* OR identif* OR detect* OR discern* OR recogni?e* OR judge*)	("dual-path" OR "reasoning" OR "skepticism" OR "deliberative" OR "heuristic" OR "cognition")		31
		What leads people to interact with disinformation on social media?			
Interactions with disinformation on social media	(Disinformation OR Misinformation OR "Fake News")	("social media" OR "social networking sites" OR SNS OR "online social network" OR Facebook OR Twitter OR Instagram OR Reddit OR Snapchat OR Tiktok OR Pinterest OR Tumblr OR WhatsApp OR Myspace OR "Second Life" OR Quora OR Weibo OR Vkontakte OR Douyin OR Kuaishou OR WeChat OR Telegram OR LinkedIn OR YouTube OR QQ OR Qzone OR "Baidu Tieba" OR Clubhouse OR Bebo OR microblogging OR blogging)	(share OR comment* OR like OR liking OR spread* OR viral OR "organic reach" OR impressions)		79

Search goal	Keywords 1	Keywords 2	Keywords 3	Exclusions	PsycInfo #
		How does social identity influence the online spread of disinformation?			
Social Identity and disinformation	(Disinformation OR Misinformation OR "Fake News")	("social identity" OR "group identity" OR "political identity" OR "self concept" OR intergroup OR "intergroup dynamic" OR "group dynamic" OR ingroup outgroup OR "collective identity" OR self-categori?ation)			8
Partisanship and disinformation	(Disinformation OR Misinformation OR "Fake News")	(partisan* OR republican OR democrat OR hyperpartisan)			32
Group norms and disinformation	(Disinformation OR Misinformation OR "Fake News")	("social norm" OR "group norm" OR "social influence" OR "subjective group dynamics" OR pro-norm OR anti-norm)			12
The digital audience and disinformation spread	(Disinformation OR Misinformation OR "Fake News")	(privacy OR disclos* OR "self-disclosure" OR anonymity OR deindividuation)			38
Identity expressions and disinformation	(Disinformation OR Misinformation OR "Fake News")	("self expression" OR "identity management" OR "identity expression" OR "impression management" OR "self-presentation" OR prosocial OR "self-affirmation")			8
Social comparisons and disinformation	(Disinformation OR Misinformation OR "Fake News")	(prejudice OR "affective polari?ation" OR "social status" OR "positive social identity" OR differentiation OR discrimination OR favo#ritism OR "ingroup bias" OR "social competition" OR "social creativity" OR "status hierarchy" OR "social dominance" OR "outgroup derogation" OR "outgroup hate")			75

Note. Number of papers reflect those found in searches run between July-September 2021.

2.2. Perceptions of Disinformation

Citizens are becoming increasingly aware of the issue of disinformation. How the public perceive and understand disinformation as a subject may offer insights for understanding its spread. First, citizens' own interpretations of what constitutes "disinformation", "misinformation" and "fake news" are addressed, before exploring research that offers insight into how disinformation is conceptualised in terms of morality (e.g. "right" and "wrong"). Finally, work on the third-person effect will be discussed to understand how citizens perceive their own vulnerabilities to disinformation in relation to others.

2.2.1. Citizens' Interpretations of "Disinformation"

While discussions surrounding definitions and interpretations of disinformation continue from academic and journalistic perspectives, little work addresses citizens' understanding. As with much recent disinformation research, the focus of these papers revolves heavily around the 2017 word of the year, "Fake News".

Disinformation may be more readily associated with outgroups than ingroups (Axt et al., 2020; Lyons et al., 2020; Michael & Breaux, 2021; Tong et al., 2020; van der Linden et al., 2020). For instance, van der Linden et al. (2020) found a majority of US-based participants made top of mind associations for the term "Fake News" that were associative (e.g. a specific news outlet or political leader) rather than descriptive. Almost three times more participants provided a "Trump" related response than descriptions of false information. Of those who gave media-related responses, 75% of conservatives associated the term "fake news" with "CNN", while 59% of liberals associated the term with "Fox News". Such politicised associations with "Fake News" have also been linked to stronger partisanship (Axt et al., 2020; Tong et al., 2020). Therefore, strength of identity may influence whether individuals associate this arguably undesirable label with outgroups.

Indeed, Social Identity Theory proposes that individuals are motivated to achieve or maintain a positive social identity (Tajfel & Turner, 2004). One strategy for achieving this is through social comparison, where ingroups are positively compared against outgroups on some dimension. Stereotyping outgroups as the main disseminators or consumers of "Fake News" may be one example of this. For instance, strong partisans who actively dislike an outgroup appear more likely to associate them with "fake news" (Tong et al., 2020). Others suggest these accusations could also help restore a sense of structure following a threat to the self (Axt et al., 2020). Indeed, as moral superiority is one way individuals can enhance their self-esteem (Dong et al., 2019) associating an outgroup with disinformation spread may allow people to view their ingroup as comparably more moral.

Another line of research suggests people may be more likely to associate "fakeness" with information that conflicts with their own attitudes and beliefs about an issue (Bago et al., 2020; Tsang, 2022). Furthermore, analysis of "fake news" themed memes sourced from social media platforms (SMPs) suggest users may direct the term towards ideologically opposing media outlets (Al-Rawi, 2021; C. A. Smith, 2019). Notably, such discourses may be more commonly found within right leaning communities to describe liberal media (Hameleers, 2020, 2020; C. A. Smith, 2019). Yet research prior to the 2020 US presidential election also found liberals were more likely than conservatives to direct the label towards politics and politicians (Tong et al., 2020; van der Linden et al., 2020). This adds to the argument that labels surrounding disinformation may be utilised in social comparison strategies, specifically as a means with which to negatively stereotype others.

2.2.2. Morality and Disinformation

People report concern about disinformation (Paisana et al., 2020) and feel it is harmful (J. W. Cheng et al., 2021). They can also be less trusting of people and media outlets who spread even one "fake news" story (Altay et al., 2020). If morality relates to

what is “good” and “bad”, then perhaps predictably, citizens’ perceptions of disinformation appear to fall into the latter.

Yet, members of groups targeted by disinformation campaigns may perceive the issue to be more severe than others (Chang, 2021). Furthermore, when their own reputation is undermined in some way, people may also judge harm to be greater for the ingroup than for others. For instance, scientists who are aware of science “fake news” may perceive harm to be greater for scientists (e.g. the ingroup) compared to the general public (Ho et al., 2020). Notably, personal norms about tackling “fake news” were also more strongly associated with perceived harm to scientists (e.g. the ingroup) compared to the public. In other words, norms reducing the spread of disinformation may be more likely to form when our ingroup is directly impacted by disinformation, at least compared to when disinformation impacts society generally.

Certain factors may also alleviate the perceived immorality of disinformation. Firstly, the actual intention behind spreading disinformation may play an important role in how the action is evaluated by others. Young et al. (2023) observed that individuals who spread disinformation unintentionally were seen by others not as moral actors but moral patients who themselves were being harmed. The discursive strategies utilised by participants when discussing these “moral patients” indicated reduced culpability. This suggests when someone accidentally spreads disinformation (i.e. “misinformation”) they may be viewed as a “victim” by others.

Furthermore, intention need not simply relate to the accidental sharing of disinformation. Indeed, qualitative studies report “fun” and entertainment to be key factors in the forwarding of disinformation (X. Chen et al., 2015; A. Duffy et al., 2020; Madrid-Morales et al., 2021). As one participant stated upon viewing a piece of Fake News: “When I’m forwarding it, I’m not forwarding news but a joke. As long as it’s a joke, I don’t need to verify” (Madrid-Morales et al., 2021, p. 1213). When disinformation has humorous appeal for interpersonal communications, the “intention” may therefore not be

to harm but instead to entertain. Therefore, the underlying motivation for spreading disinformation may be important for interpreting whether it is perceived as an "immoral" act.

Moreover, the spread of disinformation can be viewed as a lesser evil. For instance, in the context of other issues such as losing freedom of speech (J. W. Cheng et al., 2021; Jang & Kim, 2018; Melro & Pereira, 2019). Furthermore, the perceived severity of disinformation may also be influenced by context (A. Duffy et al., 2020), the framing of any disinformation threat (Sun, Oktavianus, et al., 2022, 2022) and previous experience with disinformation (Chang, 2021; J. W. Cheng et al., 2021). Therefore, disinformation may be viewed broadly as "problematic", however, moral evaluations of disinformation may remain malleable to external influences. In certain circumstances, the spread of disinformation may also be weighed up against other issues.

Notably, a limited number of studies have specifically explored the relationship between moral judgements of disinformation and its spread. These suggest that people may perceive it is more ethical to share false information that they think could have been (Effron, 2018) or may become true (Helgason & Effron, 2022). It has also been found that repeatedly encountering false information may lead people to perceive it is more ethical to share, even when they know it to be untrue (Effron & Raj, 2020). Furthermore, it is also suggested 'moral condemnation' may act as a mediator, and help explain intentions to "like" or share the content (Effron & Raj, 2020). Therefore factors such as prior exposure to disinformation content may lead individuals to be more lenient towards its spread, both in terms of moral judgements but also their own actions, even when aware of inaccuracies.

2.2.3. Perceived Vulnerability of Self and Others to Disinformation

It may be useful to consider whether users enter SMPs with an accurate understanding of risks related to disinformation. For instance, their perceived ability to detect disinformation threats. A body of research addresses whether SMP users see

themselves as being vulnerable to disinformation. Such work may provide useful context surrounding user expectations, particularly from a methodological perspective.

Research suggests greater confidence in identifying disinformation may have little or no effect on actual ability (Endresen et al., 2020; Lyons et al., 2021) or engagement in verification behaviour (Khan & Idris, 2019). Concerningly, however, Lyons et al. (2021) found almost 90% of participants reported being above average at identifying disinformation. If this finding applies to the wider population, most social media users may underestimate their own vulnerability to disinformation.

As belief in their own ability to detect disinformation increases, however, so might concern about the potential influence disinformation could have on others (Y. Cheng & Chen, 2020; J. Yang & Tian, 2021). Similarly, the Third-Person Perception (TPP) hypothesis proposes that people perceive themselves to be less vulnerable to media effects than others (Davison, 1983), potentially for self-enhancement purposes (Gunther & Mundy, 1993; Zhang, 2010). Indeed, research suggests encountering fact-checks on social media may potentially amplify TPP by reducing perceived influence on the self but not necessarily others (Chung & Kim, 2021). Therefore people may make downwards comparisons between what they perceive as their own and other people's ability to detect disinformation as a means of improving or maintaining the self-concept.

Such perceptions of "others" may expand to whole groups, allowing individuals to make intergroup comparisons. Indeed, research suggests while people view themselves as less vulnerable to disinformation than fellow ingroup members, the outgroup is also seen as being influenced the most (Corbu et al., 2020; P. L. Liu & Huang, 2020; Ștefăniță et al., 2018). This may be particularly the case for strong identifiers (Jang & Kim, 2018). Perceiving disinformation as undesirable (Y. Cheng & Chen, 2020; Jang & Kim, 2018) can also influence intergroup evaluations of TPP, suggesting these evaluations may also allow individuals to negatively stereotype outgroups.

While higher levels of disinformation-related TPP have been found to increase support for media literacy interventions, it may also reduce support for regulatory efforts to tackle the disinformation issue (Jang & Kim, 2018; F. Yang & Horning, 2020). Jang & Kim (2018) suggest that when outgroups are is seen as the source of disinformation spread, individuals may perceive educating those groups to be a better solution than infringing an entire population's freedom of speech. Therefore, individuals may begin to disengage from any collective responsibility in reducing disinformation spread by attributing greater blame to an outgroup.

There may arguably be problems if a majority of citizens perceive themselves to be relatively immune from the impact of disinformation. Research suggests individuals who feel sufficiently informed about online harms may perceive themselves as being less vulnerable, and in turn may be less likely to engage with protective behaviours online (De Kimpe et al., 2022). Yet, when this knowledge leads to greater perceived severity, they may also perceive themselves as more vulnerable and in turn, increase their intentions to take protective measures. Similarly, the effects of TPP on support for government interventions may also be reduced when people are able to perceive themselves to be similarly vulnerable to others (Baek et al., 2019). Therefore, the attitudes people hold about disinformation may be important for understanding perceived vulnerability as well as perceived responsibility.

Altogether, these findings indicate that users may feel confident about their ability to detect disinformation, but that confidence is not necessarily reflective of actual ability or their protective behaviours. Furthermore, a substantial body of work suggests that they may view disinformation spread and any solutions to be the responsibility of outgroups. The potential implications of this are discussed further within the methodology chapter of this thesis.

2.3. Identifying Disinformation

What distinguishes disinformation from other content circulating within social media platforms, other than malicious underlying intentions? From some perspectives, very little. Arguably, any links to untrustworthy websites are presented to users in the same visual format as legitimate websites. Moreover, disinformation often mimics (or even steals) user-generated content (UGC) that widely circulates within platforms. It is therefore important to understand whether users have the ability to distinguish disinformation from other content to best understand any intentions to spread it further.

To date, the identification of disinformation has been a major focus of research in this area. After addressing why the design of SMPs can make disinformation difficult to detect, this section will then discuss research looking at the influence of pre-existing attitudes and beliefs on ability to identify disinformation that confirms one's own beliefs. Findings suggesting that deliberative thinking processes can allow individuals to identify disinformation without bias, and subsequently limitations of this approach in the context of established knowledge, will be discussed. Research on motivated reasoning will then be explored before looking at the influence of identity in relation to accuracy judgements of disinformation.

2.3.1. Platform Design and Post Features

Disinformation circulating within social media platforms is not necessarily clearly identifiable. Therefore users may look for cues within a post to help gauge whether information is credible. For instance, the UK government encourages users to consider the source of information before trusting a post (UK Government, n.d.). However, research looking at the effects of source cues present mixed findings.

For instance, many social media platforms present links to external websites in a standardised format (containing website name, preview image, etc). Several studies using this format suggest the presence and position of source information may help individuals

identify unreliable sources (Bauer & Clemm von Hohenberg, 2021; Nadarevic et al., 2020), improve ability to detect disinformation (A. Kim & Dennis, 2019) and reduce disinformation sharing intentions (Di Domenico et al., 2021). However, others suggest source information may only have a limited influence on identification of disinformation (Dias et al., 2020; Schaewitz et al., 2020; Sterrett et al., 2019; Tsang, 2021). It is therefore unclear how well people are able to identify disinformation hosted on external websites (such as "fake news" links) within SMPs.

Yet, SMPs are spaces where user generated content (UGC) such as photos and videos are also exchanged and consumed (Kaplan & Haenlein, 2010). However, source information contained within image-based disinformation is much easier to manipulate or delete. Concerningly, when source information is not provided individuals may be no better at identifying disinformation than chance (Endresen et al., 2020). Another study found people may be more likely to accept source-less disinformation as accurate than factual information from a major media outlet (Clayton et al., 2019). Therefore, when no source information is displayed people may rely on other features to gauge accuracy. For instance, this may include a post's engagement counts (e.g. number of "likes") (Ali et al., 2022; Luo et al., 2022), although other findings suggest this may not be the case (Mena et al., 2020). The impact of other users will be discussed in further detail in the final section of this literature review.

As demonstrated thus far, identifying disinformation may not always be straightforward. However, while posts containing false or inaccurate information may rarely be taken down by SMPs, they may sometimes be "tagged" with a fact-check provided by an official, impartial organisation. Yet, findings regarding the effectiveness of such tags are mixed. Some research indicates that they may slightly reduce the perceived accuracy of the tagged content (Clayton et al., 2020; Pennycook, Bear, et al., 2020), while others found no such effect (Moravec et al., 2019; Oeldorf-Hirsch et al., 2020). A notable concern, however, is that such warnings about disinformation may also decrease belief in

legitimate news (Clayton et al., 2020; Moravec et al., 2019; Pennycook, Bear, et al., 2020) or even increase belief in other false information (Pennycook, Bear, et al., 2020). Therefore, platform features such as fact-check information also has the potential to influence accuracy judgements of disinformation more widely.

2.3.2. Pre-Existing Attitudes and Beliefs

A major focus of research addressing disinformation susceptibility looks at the influence of personal beliefs and attitudes. Beliefs can be thought of as probability assessment of a particular outcome being 'true' (Huber, 2009), while attitudes are object evaluations with affective, cognitive and behavioural components. Posts within SMPs which support a person's beliefs or attitudes may be judged as more accurate or credible (Huntington, 2020; Moore et al., 2021). Belief-confirming news content may also be perceived as more believable and objective than neutral or conflicting news (Kelly, 2019). People may therefore interpret information presented within social media in line with their beliefs (e.g. confirmation bias).

Disinformation appears to be treated similarly. Previous work on "fake news" suggests belief-consistent headlines are perceived as more credible (Moravec et al., 2019; Pennycook, Epstein, et al., 2021) and evaluated more positively (C. N. Smith & Seitz, 2019) than belief-conflicting headlines, whether true or false. Therefore, misinformation that reflects what a person sees as "true" may be more likely to be accepted as such. For instance, research suggests people may be more likely to believe misinformation that is consistent with their attitudes towards abortion (A. Kim et al., 2019; A. Kim & Dennis, 2019), feminism (Murphy et al., 2021), government (Vegetti & Mancosu, 2020) and political policy (Tsang, 2021).

Similarly, several studies have found relationships between closely related attitudes and disinformation susceptibility. For example, positive attitudes towards complementary and alternative medicines (Scherer et al., 2021) and low trust in medical institutions or

scientists (Agley & Xiao, 2021; Roozenbeek et al., 2020; Scherer et al., 2021; Su, 2021) have been found to predict health-related disinformation susceptibility. However, higher trust in science has been found to predict greater belief in disinformation containing false scientific claims (O'Brien et al., 2021). These opposing findings provide further support for the argument that people may interpret the accuracy of disinformation in the context of their attitudes.

One proposed explanation is that people engage with motivated scepticism. This is where information that aligns with prior attitudes is readily accepted without criticism but information that rejects or undermines attitudes is challenged (Taber & Lodge, 2012). For instance, research suggests people spend more time responding to politically incongruent information, potentially so as to generate counterarguments (Moore et al., 2021; Schaffner & Roche, 2017). Conversely, when congruent disinformation is retracted or corrected, individual's may continue to believe the content (Hameleers, 2019; Lewandowsky et al., 2005) or even increase their level of belief (Nyhan & Reifler, 2010; Schaffner & Roche, 2017).

However, other research has not found pre-existing beliefs to produce such "backfire effects" in the face of disinformation corrections (Ecker et al., 2014). Ecker et al. (2014) suggests that the distinction may be whether retractions ultimately challenge the underlying belief. Specifically, corrections highlighting flaws in belief-consistent disinformation may be viewed as a criticism of the belief itself and therefore may receive greater levels of scepticism than the disinformation itself. Overall, how information (including disinformation) relates to a person's pre-existing beliefs may influence the reasoning processes they then engage with.

Finally, members of a group may collectively hold similar beliefs and attitudes as part of their prototype. As such, specific groups may be more susceptible to certain types of disinformation. For instance, certain beliefs and attitudes may be closely connected to ideology and choice of political party. Several studies to date have found that people are

more likely to believe content that aligns with their political ideology (Allcott & Gentzkow, 2017; Bago et al., 2020). This might indicate that users within more homogenous digital spaces are less likely to spot disinformation and could therefore mean that disinformation has the potential to circulate for longer in certain digital environments.

2.3.3. Classical Reasoning

Much of the research looking at how people identify disinformation argues that engagement with deliberative thinking processes may reduce susceptibility to disinformation. Such research utilises the dual-process theory of judgements, which proposes that cognitive processes occurs via one of two systems (Kahneman & Frederick, 2012). "System 1" (e.g. intuitive thinking) is quick and effortless but relies heavily on heuristics such as affective cues, while "System 2" (e.g. deliberative thinking) is slower and effortful but more controlled. Kahneman & Frederick's (2002) dual-process model proposes that "System 2" still lightly supervises intuitive judgements however the efficacy of this supervision may be disrupted by pressures such as stress or distraction. Research taking this approach therefore assumes that when users judge the accuracy or reliability of content within SMPs, they would have greater engagement with either intuitive or deliberative thinking.

Several studies have found that people who more readily engage with deliberative thinking may be better able to discern between legitimate and "fake" news headlines pulled from social media (Bago et al., 2020; Bronstein et al., 2019; C. M. Greene et al., 2021; S. Lee et al., 2020; Pehlivanoglu et al., 2021; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019, 2020; Ross et al., 2021; Salvi et al., 2021). Such studies commonly employ the Cognitive Reflection Test (CRT) to measure tendency to engage in deliberation by asking questions such as "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" (Frederick, 2005). For instance, in the example provided here the intuitive answer would be "100", however, if system 2 stepped

in to override this judgement participants would give the correct answer (i.e. “5”). This increased tendency to override intuition may then explain why high deliberators appear to be better able to distinguish between “real” and “fake” headlines. In contrast, Martel et al. (2020) found that participants who were encouraged to make intuitive judgement were more likely to incorrectly judge “fake” headlines as accurate compared to a control group. This suggests relatively effortless, intuitive judgements may indeed make people more vulnerable to disinformation.

Furthermore, classical reasoning approaches state that deliberative thinking should lead to unbiased judgements. Indeed, several studies have found associations between greater deliberation and reduced or eliminated biased judgements of political disinformation (Bago et al., 2020; Bronstein et al., 2019; S. Lee et al., 2020; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019, 2020; Ross et al., 2021). However, others have argued this may be due to the use of ‘discernment’ calculations (the sum of accuracy ratings for both true and false headlines) to measure bias as well as accuracy of judgements (Batailler et al., 2021). Reanalysing data from (Pennycook & Rand, 2019), Batailler et al. (2021) suggests high deliberators may be more likely to accurately distinguish between real and “fake” headlines, but that deliberation may not reduce ingroup bias. The reanalysed data suggested that regardless of CRT score, participants were more likely to judge incongruent headlines as “fake” than congruent headlines, even when they were actually true.

Moreover, there is some evidence to suggest deliberative thinking may not always be as beneficial as hoped, and at times may actually increase susceptibility to disinformation. For instance, research suggests high deliberators may be more vulnerable to disinformation that connects previously seen “real” news stories (associative inference) than other disinformation strategies (S. Lee et al., 2020). Furthermore, strong deliberators were found to be more prone to believing false claims from a deepfake video when the video was not disclosed as being fake (Ahmed, 2021). This may suggest that any benefits

of deliberation may be dependent on context. Indeed, other research suggests that the benefits associated with scoring highly on the CRT may also only reduce susceptibility within specific cultures (e.g. the US) (Salvi et al., 2021) and for certain disinformation themes (Scherer et al., 2021). Given the serious impact that disinformation can have, it is important to ensure that undue blame is not attributed to individuals for simply not thinking hard enough in spaces not designed for strenuous thought. The following subsections will highlight alternative explanations for these findings, including the influence of social identity on system 2 reasoning.

2.3.4. Accessibility and Identification – Influence of Plausibility, Exposure, and Prior-Knowledge

Within the current literature, several findings support the idea that people judge disinformation that complements their pre-existing knowledge to be more accurate. For example, Pennycook & Rand (2019) identified a relationship between deliberative thinking and plausibility judgements which may explain why high deliberators more accurately identified "fake news". Therefore, judgements of disinformation may be made against a spectrum of plausible and implausible outcomes, and so may relate to knowledge activation. Indeed, others have found exaggeration or sensationalism within disinformation content may have no impact on individual's ability to detect disinformation (Einav et al., 2020; Schaewitz et al., 2020). If individuals do not hold the correct knowledge to base deliberative plausibility judgements against, they may not be able to gauge if a claim is overstated.

As in Pennycook & Rand (2019), gauging plausibility may help people detect certain styles of disinformation. However, arguably, disinformation can often be entirely plausible and, at times, may even involve legitimate content being used out of context. With this in mind, the use of plausibility judgements to detect disinformation may ultimately not be a reliable method for user-identification. To date there are no known

studies looking at the influence of disinformation plausibility in relation to detection ability.

However, previous work suggests people may rely heavily on plausibility judgements to detect lies (Hartwig & Bond, 2011). Moreover, Duran et al. (2020) found participants demonstrated a level of ability to detect emotion or opinion-based lies. However, their ability to detect lies based on "actions" appeared to be no better than chance. This may be because action-based lies were structured in a manner that was more closely based on reality (albeit adapted for the lie) and therefore may appear more plausible. If such findings also apply to disinformation, users may be better able to more accurately detect content that is opinion-driven or emotive. However, disinformation that has a potential to be "real" may be less likely to be detected.

Furthermore, several studies have compared ability to identify disinformation against specific types of issue-specific knowledge. Firstly, overclaiming knowledge has been linked to greater vulnerability to disinformation (Pennycook & Rand, 2020; Salvi et al., 2021). Therefore, those who attempt to identify disinformation against incorrect (or simply missing) knowledge may be less likely to make accurate judgements. However, knowledge of politics has been associated with ability to identify political disinformation (Vegetti & Mancosu, 2020). This also appears to be the case for Brexit knowledge and Brexit disinformation (C. M. Greene et al., 2021). However, Scherer et al. (2021) found that while participants who took statin medications were less susceptible to disinformation about statins, this did not apply for disinformation about other treatments. These findings suggest that a pre-existing body of accurate knowledge may help reduce the influence of disinformation, but preferably when it relates to the specifically targeted issue.

Levels of plausibility and comprehension of disinformation are also related (Abendroth & Richter, 2020). If individuals lack the skills or knowledge to interpret complex (yet accurate) information from reputable sources, they may perceive any more easily comprehensible alternatives proposed by disinformation sources to be more

plausible. With this in mind, research suggests that developing individual's broader knowledge may then also help with disinformation identification. For example, several studies have found a connection between education level and accuracy judgements of disinformation (Allcott & Gentzkow, 2017; Baptista et al., 2021; Scherer et al., 2021). Health literacy (Scherer et al., 2021), digital media literacy (A. M. Guess et al., 2020), critical media literacy (Xiao et al., 2021), information literacy and skills (Dabbous et al., 2022; Jones-Jang et al., 2021) and numeracy (Roozenbeek et al., 2020) have also all been connected with reducing the perceived accuracy of disinformation. Furthermore, a virtual game which teaches users about the production and techniques of disinformation has also been found to reduce susceptibility (Basol et al., 2020; Roozenbeek, Maertens, et al., 2021; Roozenbeek & van der Linden, 2019). Therefore, the more easily individuals are able to evaluate information (but also question cues of errors within disinformation against their pre-existing knowledge) the better they may make accurate judgements in relation to identifying disinformation.

Finally, prior exposure may also influence the perceived accuracy of disinformation. Three studies found demonstrate that repeated exposure to a piece of disinformation may increase individuals' perceptions of accuracy (Effron & Raj, 2020; Nadarevic et al., 2020; Pennycook et al., 2018). As such, each time that an individual encounters a specific piece of disinformation, the more likely it may be that they perceive it to be "true". This 'illusory truth effect' also appears to be unaffected by plausibility (Fazio et al., 2019). In other words, even if individuals repeatedly encounter implausible disinformation, they may begin to perceive it to be more accurate due to the increased quantity of such exposure. This effect may also occur when individuals are presented with disinformation taking a similar stance on an issue to legitimate information they had previously consumed (Abendroth & Richter, 2020). Concerningly, this may mean an individual may be more vulnerable to disinformation that is similar enough to legitimate information they have previously heard.

2.3.5. Motivated Reasoning

Motivated reasoning is not simply making biased evaluations, but accounts for adjustments in cognitive strategies that are consciously or unconsciously employed to achieve goals including accuracy or reaching desired directional outcomes (Leeper & Slothuus, 2014). Which strategy is employed may be influenced by external context such as the political environment, or goal-related norms (Creyer et al., 1990). When social media users (or indeed participants) are presented with disinformation, how they evaluate "accuracy" may therefore change in relation to their most pressing goal (e.g. identifying disinformation vs. protecting identity) and this in may turn be influenced by external events or what is perceived as desirable. For example, political environments may influence individuals to make judgements and decisions against partisan goals (Leeper & Slothuus, 2014).

Indeed, individual concerns such as promotion or prevention may be triggered by external cues (Molden et al., 2008). Therefore, within SMPs the goals may constantly shift due to everchanging environments and contexts. The impact of such switches on goal-related outcomes may be apparent in research looking at the influence of digital "echo chambers" on identifying disinformation. Rhodes (2022) found Democrats (but not Republicans) were better at identifying "fake news" within heterogenous news environments. However, when presented with only Democrat headlines, these participants were much more likely to believe false headlines. This suggests environments where partisan goals are more salient may lead people to make more biased judgements of disinformation, even when they tend not to otherwise. As previous work suggests high homogeneity in instant messaging (IM) groups may increase personal tolerance to spreading disinformation (Gill & Rojas, 2020), SMP environments which only provide a limited range of beliefs and opinions may also reduce how severe people perceive disinformation to be.

Interestingly, research on motivated reasoning indicates that drawing attention to accuracy may also improve participant performance on accuracy-related tasks (Creyer et al., 1990). This is notable given much research on disinformation identification requires participants to make accuracy ratings (Martel et al., 2020; Pehlivanoglu et al., 2021; Pennycook & Rand, 2019; Salvi et al., 2021). Arguably, in the context of motivated reasoning, whether such tasks inadvertently prime cognitive strategies for achieving factual-based accuracy goals may be an important methodological consideration.

Finally, while classical reasoning suggests that judgements prone to bias are quick and effortless (due to reliance on heuristics), motivated reasoning supposes biased judgements may involve additional effort. Specifically, when disconfirmation bias occurs, people may dedicate greater resources to defend against attitude-incongruent arguments, and yet still may form invalid conclusions (Taber & Lodge, 2012). For example, EEG research found people dedicated greater cognitive resources when shown belief-consistent headlines containing fact-check flags compared to headlines that undermined their beliefs (Moravec et al., 2019). Despite this, participants continued to prioritise their beliefs in accuracy judgements. In other words, engaging with more effortful thinking did not necessarily lead to a reduction in bias or accurate identification of disinformation as proposed by the classical reasoning account. Such findings could even indicate that people may not necessarily make judgements about ‘known’ disinformation that confirms their beliefs using the same processes as disinformation that undermines them.

2.3.6. Influence of Social Identity on User-Identifications of Disinformation

Social identity (as opposed to personal identity) refers to the collection of group categories that an individual has internalised to become part of their self-concept (Tajfel & Turner, 2004). While a collection of individuals who, for example, hold similar views on an issue may be cognitively categorised as a ‘group’, for an identity to form a person must have internalised said group themselves. Social identity is therefore thought to be an

affective component of group membership, whereby strong identification with an ingroup involves an increased emotional connection. Whether a group is evaluated positively or negatively compared to other groups therefore has a direct influence on an individual's self-concept. When an ingroup is evaluated negatively, individuals may either leave the group or engage with positive distinctiveness strategies to shift the position of the ingroup. Alternatively, they may even reject the credibility of information that negatively frames their ingroup as a means of undermining the threat to the group's value (Ellemers et al., 2002).

When presented with content that affirms or threatens one's identity, individuals may engage with identity-protective cognition leading them to make more polarised judgements (Kahan, Jamieson, et al., 2017). Disinformation that fulfils this criterion could lead to similar results. For example, a number of studies have found individuals may be more susceptible to disinformation that benefits an ingroup or undermines an outgroup (Faragó et al., 2020; Neyazi & Muhtadi, 2021; Pereira et al., 2023). Furthermore, people may be more willing to believe information that suggests a political ingroup upheld rather than undermined values, even when the values conflict with their politics (Pereira et al., 2023). Therefore, when identity is salient within disinformation, how the ingroup and outgroup are framed in the context of moral status may be more important for perceived accuracy than other factors.

Identity-cues within disinformation content may only be salient to certain ingroup members based on their personal knowledge. For example, two rival groups evaluating the same information (identity-affirming for one and undermining for the other) may come to different conclusions about accuracy (Schaffner & Luks, 2018; Schaffner & Roche, 2017). Research presenting pictures comparing former Presidents' Obama and Trump's inaugurations found that, when asked to identify which had larger attendance, those with high approval of Trump and higher education or political interest were more likely to give an incorrect answer (e.g. Trump's) than those with lower education and interest. However,

when asked to label each photo to a president, the level of inaccuracies made by Trump approvers with lower education or political interest matched or overtook judgements by those with more education and interest. In other words, for those with higher education and political interest, the identity-context may have been more salient than for other Trump approvers in the original condition, without the explicit cues of leader names within the question. Given that, for impartial observers at least, the question of which photo has a larger crowd should be clear, protection of identity may be prioritised over acknowledging reality. Acceptance of disinformation may therefore be viewed as a valuable tool for identity protection.

The emotional aspect of social identity may explain why identity strength has been linked to increased disinformation susceptibility in certain contexts. For example, strength of political identity has been linked to belief in political fake news (Anthony & Moulding, 2019). Sanchez & Dunning (2021) also found that strength of emotional investment in a political ingroup was related to greater belief in ingroup friendly disinformation, as well as reduced belief in disinformation unfriendly to the ingroup. For friendly disinformation, emotional investment of partisans was also a stronger predictor of belief than cognitive ability. Additionally, how strongly someone identifies as “anti-vaxx” has also been found to predict decision making surrounding vaccination mandates (Motta et al., 2023). When a group forms around collective misbeliefs, the added social and emotional dimensions may therefore make acting on disinformation more likely.

Finally, a line of research related to motivated reasoning suggests that those who more readily engage with deliberative thinking use their abilities to selectively engage with different strategies in a way that may benefit their goals, and notably, protect their identity. For example, when participants low and high in religiosity were asked to identify whether the statement “human beings, as we know them today, developed from earlier species of animals” was true or false, there was greater polarisation in responses as science comprehension increased (Kahan, 2015, p. 8). However, when “according to the theory of

evolution" was added to the statement, correct answers positively correlated with science comprehension scores, regardless of religiosity. Where necessary, participants appeared to shift strategies between protecting identity or accuracy, dependent on the question framing. It has also been suggested that certain issues, such as climate change, fracking, gun control and immigration may be tightly linked to identity and may produce more polarised judgements in high deliberators (Kahan, 2013, 2015; Kahan et al., 2012; Kahan, Jamieson, et al., 2017; Kahan, Peters, et al., 2017). By this account, those who more readily engage with system 2 may quickly adapt their reasoning strategy in relation to the immediate goal. Disinformation unrelated to identity (or indeed in accuracy-focused situations) may be more accurately identified, however, within an identity-salient SMP environment, the needs of the ingroup may instead be favoured.

2.4. Amplifying the Spread of Disinformation via Social Media Interactions

While malicious actors may be the initial disseminators of disinformation on social media, its further spread relies on other users interacting with it. While reposting content is one way of contributing to the reach of disinformation, any user interaction can ultimately positively impact its spread. Due to the introduction of algorithmically ordered feeds across most major SMPs, all actions (whether "liking", commenting or even "saving") help determine the reach of a post (Facebook, n.d.; LinkedIn, n.d.; Mosseri, 2021; TikTok, 2020). This section addresses research exploring the impact of belief in the message content, trust in the digital environment and ideological alignment to explore the similarities between interactions with disinformation and any other type of UGC. Finally, the impact of affect and outrage on increasing disinformation reach is discussed.

2.4.1. Believing the Content

While only a minority of individuals may deliberately spread disinformation within social media platforms, many spread it thinking it is true. Several studies suggest users are more likely to interact with disinformation when they believe the content to be accurate

(Ahmed, 2021; Baptista et al., 2021; Buchanan, 2020; A. Kim et al., 2019; A. Kim & Dennis, 2019). Research focusing on identification of inaccurate headlines (e.g. "fake news") also suggests people who correctly identify more "fake" headlines are much less likely to share (T. Hopp, 2022). Therefore, disinformation that is believed (whether believable, or simply confirms what an individual or their social group perceives to be "true" or normative) is more likely to be spread on SMPs.

As previously discussed, accurate identification of disinformation is not, however, straightforward. Disinformation that affirms a person's beliefs about the world may be more believable simply by nature of probability. At times, disinformation can even present real situations out of context. Yet, it has been argued that users may simply not be considering accuracy when interacting with content on SMPs. Although people claim that the accuracy of the information they share is important, research suggests they are more likely to prioritise their politics when sharing (Pennycook, Epstein, et al., 2021). However, Pennycook et al. (2021) found asking participants to rate the accuracy of an unrelated headline improved sharing discernment. Others have also explored the effect of drawing attention to accuracy, finding varying degrees of success (Capraro & Celadin, 2022; Epstein et al., 2021; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2022; Roozenbeek, Freeman, et al., 2021). Therefore, simply having doubts about the accuracy of unverified information may help reduce the spread of potential disinformation in some way.

2.4.2. Trust Within Social Media Platforms

A number of studies demonstrated that trust may play an important role on the onward spread of disinformation in a number of ways. Firstly, relationship with the digital environment may play a role. Indeed, people who report greater trust in information circulated online and within SMPs may be more likely to share disinformation themselves (Apuke & Omar, 2021; Laato et al., 2020; Talwar et al., 2019). Furthermore, trusting a

message source (e.g. URL) may also increase the likelihood of engaging with disinformation (Sterrett et al., 2019). It may be the case that users do not expect to be presented with disinformation within environments they place trust in.

How much people trust individual users or accounts may also influence whether they go on to share disinformation posted by them. Indeed, higher trust in the sharer account (as opposed to original message source) is associated with greater likelihood of interacting with false content (Bringula et al., 2022; Buchanan & Benson, 2019; Di Domenico et al., 2021; Sterrett et al., 2019). Notably, when disinformation travels across SMPs, it can become disconnected from the original poster (OP) or indeed platform. As such, users may be more willing to spread disinformation unwittingly spread by a friend or family member than had they been presented with the OP post.

Moreover, if a user sees that someone they trust has "liked" a post they may perceive it to be more credible (Mena et al., 2020) and as previously noted may then be more willing to spread it themselves. Additionally, research has suggested high trust in the sharer account can reduce the benefits of interventions intended to help reduce disinformation spread (Di Domenico et al., 2021). Therefore, the relationships people have with other users and accounts may be important for understanding why people spread disinformation. Yet notably trust in another user does not need to be based on in-person or even reciprocal relationships. People may be more likely to perceive disinformation to be more credible when it is shared by a trusted public figure (Mena et al., 2020), and also may be more likely to spread it further (Sterrett et al., 2019). Therefore, "trust" does not only come from knowing the account owner personally.

Furthermore, people may be more susceptible to disinformation when posted by an account that generally shares belief-consistent information (Bauer & Clemm von Hohenberg, 2021). Concerningly, however, research suggests whether the account owner is an expert or not may not be important to users when gauging the accuracy of their posts (Hameleers, 2019). Therefore, disseminators of disinformation who can develop user trust

through legitimate posts may be able to take advantage of their trust, even if users do not see them to be an "expert".

Finally, it has been suggested that individuals may put greater trust into accounts owned by government departments or official organisations for health related information than individual users (Trivedi et al., 2020). Participants were then more likely to believe this content. However, it is unclear whether people may be able to differentiate between legitimate or imposter organisations. Additionally, higher trust in science has been associated with a greater willingness to share disinformation that contains scientific references (O'Brien et al., 2021). If disinformation is presented as "official information" from institutions users trust, then people may be more likely to believe and spread it further.

2.4.3. Ideological Differences or Simply Appealing to Political Ideology?

A number of studies have suggested that political conservatives may be more susceptible to spreading disinformation (Baptista et al., 2021; Garrett & Bond, 2021; A. Guess et al., 2019). Notably, in their Portuguese-based study, Baptista et al. (2021) found politically right-leaning participants were more willing to spread fake headlines regardless of stance, suggesting more than a simple confirmation bias. Others suggest political conservatives may be more tolerant of spreading disinformation (De Keersmaecker & Roets, 2019) and be less likely to process retractions of attitude-consistent disinformation (Ecker & Ang, 2019). Conversely, prior research suggests "individualising" moral foundations (more commonly associated with political liberals) predict reduced acceptance of COVID-19 disinformation (Ansani et al., 2021). The findings from these studies may therefore support some of the political asymmetries observed for disinformation spread within SMPs.

However, Ryan & Aziz (2021) argue that illusionary truth effects may instead explain political asymmetries found in studies using disinformation that has previously

circulated in the real world (e.g. Garrett & Bond, 2021; A. Guess et al., 2019, etc). By presenting three pieces of original 'disinformation' in their study, Ryan & Aziz (2021) found participants on the political left and right to be equally vulnerable to accepting disinformation. Whether or not people are simply more likely to encounter right-leaning disinformation on social media (and therefore appear more susceptible when presented with the same narratives) is therefore an important methodological consideration.

The tactics used in the creation dissemination of disinformation may also mean that the categorisation of harmful content into two ideological categories is not always appropriate. Arguably, such approaches may overlook the nuanced strategies previously observed in some disinformation campaigns (DiResta et al., 2019; François et al., 2019). For example, previous work has categorised disinformation targeting social movements including Black activists and the LGBTQ+ community under a broad "liberal" category (e.g. Helmus et al., 2020). However, such content tends not to be disseminated by generic "liberal" accounts (François et al., 2019) or targeted to broad "liberal'" audiences (DiResta et al., 2019). More often than not, such "left leaning" disinformation has previously been disseminated instead to specific groups. Arguably, dichotomising by ideology in this context may risk overlooking the potential influence of beliefs and social identities, particularly in those who are politically liberal.

Indeed, analysis of known "sock puppet" accounts on Twitter active in the run up to the US election found half the number of left-leaning accounts compared to right-leaning, and of these left-leaning accounts, half inauthentically presented as Black activists (Freelon et al., 2022). These accounts posted just a third of the number of tweets that right-leaning accounts had, however, tweets from "Black activist" accounts also attracted more engagement. Similarly, it has been found that while the IRA invested the greatest amount of money targeting Facebook adverts at an audience segment called "Conservative Politics and Culture", adverts targeted at this segment were less effective than those targeted at the segments "African American Politics and Culture" and "Latin American Culture" (Howard

et al., 2018). While a similar number of adverts were found to target the Latin American and Conservative segments, the latter was around ten times more expensive to target yet adverts received around half the number of impressions and clicks. Such findings demonstrate that in the context of social media factors such as total number of posts or financial investment are not the same as efficacy or even reach.

Arguably, it may therefore be difficult to distinguish any "true" effects related to political ideology from those arising from disinformation content and strategies. Notably, spreading disinformation may provide users with a way to express their opinions (X. Chen et al., 2015) and may also be influenced by higher levels of investment in any topics featured within the content (Bringula et al., 2022; A. Kim et al., 2019; A. Kim & Dennis, 2019; Osmundsen et al., 2021; Schaewitz et al., 2020; Sterrett et al., 2019; Valenzuela et al., 2019). With that in mind, people may spread disinformation because it is similar to what they usually share (and therefore arguably may align with ideology-related beliefs).

2.4.4. The Influence of Emotion and Affect

Disinformation can often be framed in a way that may elicit negative emotion (Acerbi, 2019; Cheung-Blunden et al., 2021). For example, disinformation content may often be related to threat, sex or disgust (Acerbi, 2019). One reason for this is that compared to neutral content, emotional and moral content captures more visual attention which, in turn, may increase the likelihood users share the content (Brady et al., 2020). This may explain why certain topics attract more engagement when inducing anger (Brady et al., 2017) and fear (Ali et al., 2019). Therefore, one reason people may be more likely to spread emotion-inducing disinformation is that they are more likely to notice it. Additionally, if experiencing an emotion such as anger people may be more likely to believe disinformation that affirms their ingroup (Weeks, 2015). As such, emotion-inducing disinformation could create serious challenges when disseminators seek to undermine less emotive, official information.

Moreover, emotions experienced in response to viewing disinformation may also help explain how people go on to interact with it. For instance, compared to "real" news posts, language in user responses to disinformation contains higher levels of disgust (Vosoughi et al., 2018) and anger (Barfar, 2019; Pulido et al., 2020). Another study found disinformation about Ebola contained higher risk perception frames and also induced more user discord than "real" tweets. It has also been found that when users view negative tweets about vaccines (including disinformation) they may be more likely to subsequently post negative tweets about vaccines themselves (Dunn et al., 2015). Such findings also correspond with reports that researchers at Facebook identified anger as an important driver of disinformation spread within the platform (Merrill & Oremus, 2021).

People may also respond to disinformation as a means of expressing positive emotions. It may even be the case that in some instances people are more willing to spread content on divisive political issues when it is positively framed than when it is negative (Brady et al., 2017). This may also explain why one study found participants were more likely to spread disinformation when they experienced strong and positive reactions (Helmus et al., 2020). Moreover, research looking at disinformation content posted by the IRA and Iran on Twitter suggested users may be more likely to "like" and share positively-framed disinformation compared to the more commonly found disinformation that was negatively-framed (Cheung-Blunden et al., 2021). Rather than simply the type of emotion, these findings indicate that the strength of affective response elicited upon viewing disinformation may be important for understanding users' interactions with it.

Yet others have found that disinformation featuring fear and anxiety appears to receive fewer engagements on Twitter (Cheung-Blunden et al., 2021). Other work suggests that "sad" disinformation tweets spread more slowly and less far than other types of disinformation (Vosoughi et al., 2018). However, it is thought that when people experience emotions such as fear, anxiety, and sadness they may feel an urge to avoid or withdraw

from a situation (Lazarus, 1991). It may be that certain emotions encourage people to refrain from spreading disinformation in some circumstances.

Moreover, in other situations people may even take an active role in reducing disinformation spread. When people consider the potential harms of disinformation they may experience emotions such as anger (Myrick & Erlichman, 2020; Sun, Chia, et al., 2022; Sun, Oktavianus, et al., 2022). This anger may then help to increase a person's support of regulatory interventions (Sun, Chia, et al., 2022). Other work suggests that people are less likely to spread disinformation when they are more morally condemning of it (Effron & Raj, 2020). Therefore, disapproval through emotions such as anger may play a role in reducing the spread of disinformation but may need to be directed at disinformation itself rather than towards any narratives featured within the content.

So far, these studies have illustrated the influence of emotional responses to disinformation content. However, the anticipation of affective responses can play an important part in behavioural regulation. For instance, when people violate a moral norm they may experience the distressing emotion of guilt (Lazarus, 1991). Therefore, a person may experience guilt after learning a post they previously shared on social media contained false information. To avoid distress in the future, rapid and unconscious affective cues then help guide them to adjust their behaviour in similar situations (Baumeister et al., 2007). As such, higher levels of "anticipated guilt" are thought to promote increased intentions to correct disinformation on social media (Sun, Chia, et al., 2022; Sun, Oktavianus, et al., 2022). Therefore, disinformation may not need to elicit a "strong" emotional response to influence a person's behaviour. Instead, affective processes may unconsciously guide whether or not a person interacts with disinformation in line with their moral norms and experiences in similar situations.

2.5. Social Identity and the Spread of Disinformation

As discussed in chapter one, Social Identity Theory proposes that individuals are motivated to have a positive social identity (Tajfel & Turner, 2004), which may be impacted by factors such as violations of moral norms (Ellemers et al., 2013). When identity is threatened people may engage in certain strategies to maintain or achieve a positive social identity (Tajfel & Turner, 2004). These include "social competition", where downwards comparisons are made against a relevant outgroup, allowing people to perceive ingroup superiority. Within digital environments this might manifest as prejudice towards an outgroup (Ahmed et al., 2021) or affective polarisation (Bliuc et al., 2021).

Social media platforms also provide spaces in which people can connect with other group members without physical barriers, as well as play with and express their identity. However, certain groups have also been targeted on SMPs by those who disseminate disinformation (François et al., 2019). The following section will therefore explore the potential impact of digital audiences on the spread of disinformation, specifically the influence of echo chambers and group norm regulation. Next, the potential use of disinformation in strategies for achieving positive distinctiveness is discussed. Finally, research addressing the links between identity threats and disinformation spread are addressed.

2.5.1. The Digital Audience – Norm Conformity and Echo Chambers

The application of both injunctive and descriptive norms (e.g. what others think should be done vs. what others actually do) have been explored in relation to disinformation spread on SMPs. For instance, disinformation content may express that people should behave in a certain way, such as engaging in behaviour that may be dangerous for one's health. By increasing perceptions that others would approve of the behaviour, these injunctive norms may increase the likelihood people engage in said harmful behaviour (Myrick & Erlichman, 2020). However, injunctive norms have also

been harnessed within interventions that encourage reporting of disinformation, helping to reduce their spread (Andı & Akesson, 2020; Gimpel et al., 2021). Therefore, people's perceptions of how others feel they should behave may be important for understanding how they interact with disinformation.

However, the effect of descriptive norms appears to be less straightforward. Recent studies suggest that awareness that other users marked a post as misleading (Pretus et al., 2022) or chose not to interact with disinformation (C. M. Jones et al., 2021) may reduce intentions to spread. Yet, Gimpel et al. (2021) found including a "count" of how many other people reported a post only had an impact on participant's own reporting intentions when used in conjunction with an injunctive norm intervention. Additionally, increasing the number of users who had supposedly reported the content improved the probability of participants reporting, but only to a point. When presented with the largest value (3,125 users) participants were least likely overall to report. These findings suggest that perceiving others to act in a manner which may reduce the spread of disinformation may influence whether a person is willing to intervene themselves; but only if they are aware this is what is expected of them also. However, if too many others are seen to be involved then people may feel they do not need to get involved themselves.

Furthermore, other work has looked at how disinformation exists and travels within online communities. For instance, echo chambers within SMPs may impact disinformation spread if users perceived there to be majorities of opinions. Within SMPs opposing the group consensus can be viewed as identity-subversion, and may attract derogation from ingroup members and potentially even exclusion (Ditrich & Sassenberg, 2017). Indeed, research suggests one reason people refrain from spreading disinformation is to protect their reputation (Altay et al., 2020). As such, people may adjust their evaluations (Jahng et al., 2021) and intentions to interact with (Boot et al., 2021; Colliander, 2019) disinformation in line with comments made by other users (e.g. that it is "fake", etc).

Therefore, cues from other users may help discourage people from spreading a piece of disinformation.

However, similar processes may also assist in amplifying the spread of disinformation. A related body of research based on the spiral of silence suggests that users may self-censor online when they perceive their opinions are in the minority (Woong Yun & Park, 2011), potentially due to fear of isolation (H. T. Chen, 2018; Fox & Holt, 2018; Wu & Atkin, 2018). As such, while the act of digitally "spreading disinformation" may be a proscriptive norm (e.g. behaviour one should refrain from) other norms such as ones which discourage people from undermining commonly held beliefs (or indeed other group members) may at times be stronger.

Such fears may explain why one disinformation study found users appeared to only rarely criticise the accuracy of shared claims within political Facebook groups (Hameleers, 2020). Another study looking at posts from a private anti-vaccination Facebook group found users who criticised majority opinions experienced bullying and even membership removal (Bradshaw et al., 2021). However, this is not simply due to removal of those who make serious norm violations, but also through regulating the behaviour of other group members. Indeed, Bradshaw et al. (2021) found users who inadvertently violated norms sometimes made efforts to reinstate harmony in their subsequent comments. This may be because individuals are aware that standing by their own beliefs in the face of other's conflicting opinions can be viewed negatively by others (Bonetto et al., 2019). If individuals desire to be perceived as warm and reduce psychological distance they may adjust how resistant to persuasion they appear (Bonetto et al., 2019). Individuals may therefore be willing to sacrifice beliefs in the face of conflicting disinformation to present a positive self-image to the ingroup. As such, digital communities who punish users for challenging majority opinions may create environments where disinformation that supports said opinions is allowed to circulate with relatively little criticism.

Moreover, users may be rewarded by the group for behaving in a normative manner, which may at times include spreading disinformation. It may even be a way that some people feel they are able to develop their status (Apuke & Omar, 2021). For example, analysis of an AIDS-denialist group with 15,000 members on VK.com found that those who adhered to AIDS-denialism were not only more likely to post, they were also more likely to receive "likes" from others (Rykov et al., 2017). Others have observed new users within anti-vaccination Facebook groups may be rewarded with positive treatment if appearing open to advice (Bradshaw et al., 2021). Therefore, digital communities may encourage and even reward the spread of disinformation if it is deemed as abiding with group norms.

Finally, it is also likely that some disinformation disseminators are aware of how communities regulate norms within SMPs. Indeed, research analysing posts by IRA sock puppet accounts on SMPs such as Twitter (Xia et al., 2019) and Tumblr (Neill Hoch, 2020) observed constructed performances which conformed to certain social norms, possibly to avoid detection by other users. For example, some of the most successful IRA sock puppet accounts may have demonstrated a strong understanding of the nuanced digital communication norms within individual platforms (Neill Hoch, 2020). Certain accounts also utilised screen-grabbed content posted by real users to reduce the chance of detection through non-normative language use and behaviour (Neill Hoch, 2020). How well disinformation conforms to the norms of SMP communities it is intended to spread within may therefore have an important role in its spread.

2.5.2. Disinformation as a Means for Expressing Positive Distinctiveness

Social identity theory proposes that people are motivated to seek or maintain a sense of positive distinctiveness for the group (Tajfel & Turner, 2004). They may achieve this by engaging with social creativity strategies which can occur within SMPs. For example, some users use SMPs to express or play with identity and through harnessing

pre-packaged UGC, users may also be able to do this with relatively little effort. However, disseminators of disinformation have been known to hijack the self-presentational affordances that SMPs provide users to create compelling sock puppet accounts. One infamous IRA sock puppet, “Jenna Abrams”, had a Twitter account amassing 70 thousand followers, a blog, and an active email account. Analysis of her account identified a deliberate performance to present the political and national identities of a strong conservative and an American citizen (Xia et al., 2019). Xia et al. (2019) also suggest that by exploiting the functionality within SMPs to express identity, “Abrams” was able to achieve a sense of authenticity. Additionally, it has been suggested that people who share content within SMPs as a means for self-expression may be more likely to spread disinformation (Apuke & Omar, 2021). As active participation is encouraged within SMPs, engaging with identity-expressive disinformation (such as disseminated by sock puppet accounts) may be a way in which other users can express what appears to be a shared identity.

Another way people can enhance or maintain a positive self-concept is to engage with social comparison strategies. This involves making favourable evaluations of an ingroup compared to an outgroup, or negative evaluations of an outgroup. Indeed, analysis of Twitter data indicates that people may be more likely to spread disinformation that benefits the ingroup compared to disinformation that puts the ingroup in an unfavourable position (Osmundsen et al., 2021). Such interactions may even utilise featured within a platform, such as “love” reactions towards posts about the ingroup and “angry” reactions for posts about outgroups (Rathje et al., 2021). In other words, when disinformation provides an opportunity for users to engage in social comparison strategies, they may interact with it in a way that helps achieve or maintain positive distinctiveness.

It may also be that people may be more likely to resort to spreading disinformation when legitimate content does not allow people to engage in strategies that support positive distinctiveness. Indeed, Osmundsen et al. (2021) found that US partisans were most likely

to select “politically useful” news to share on Twitter. However, pro-Democrat, centrist and pro-Republican “real” news headlines were found to be comparably or more negative towards Republican versus Democrat elites. It was therefore suggested that sharing pro-Democrat and centrist “real” news would facilitate strategies for achieving positive distinctiveness for Democrats, however, few “real” headlines would do this for Republicans. Indeed, only pro-Republican “fake news” headlines truly fulfilled this role for Republicans, which Osmundsen et al. (2021) suggested may be a reason as to why it is more common. Therefore, people may care less about whether the information they spread within SMPs is “accurate” or not, but whether it helps achieve a directional goal. Yet, when the only content available to help fulfil said goal is itself untrue, then any concerns for accuracy may be dismissed in favour of achieving positive distinctiveness for the ingroup.

There is also evidence to suggest that people may be more willing to share disinformation that is critical of the outgroup than disinformation that highlights a positive ingroup identity (Pereira et al., 2023), which is also supported by work on the spread of SMP content generally (Rathje et al., 2021). Others have found a relationship between negative feelings towards the outgroup and increased likelihood of spreading disinformation (Osmundsen et al., 2021). This is something that disseminators of disinformation may take advantage of, for instance, by targeting pre-existing tensions to encourage hate-driven interactions. For example, in Denmark extreme examples of hostility prejudice were found on fake Facebook pages which pretended to be authored by radical Islamists (Farkas et al., 2018, 2018). By creating fraudulent profiles playing into racist stereotypes, those behind the accounts may have been able to reinforce inaccurate, racist beliefs as well as provide some users with false content that could be utilised in expressions of prejudice.

Finally, research also suggests certain people may be more prone to spreading disinformation that facilitates social comparison or social creativity strategies. For

instance, how strongly they identify with a group may also influence whether they endorse or interact with content in line with social comparison strategies (F. J. Jennings et al., 2020; Sanchez & Dunning, 2021). Additionally, exaggerated beliefs about the importance of the ingroup may influence belief in disinformation, and in turn, support for collective action (Mashuri et al., 2022). Collective narcissism has also been linked to increased support of misleading government campaigns about the environment (Cislak et al., 2021) and belief in outgroup related conspiracy theories, but not ingroup related (Cichocka et al., 2016). The relationship between a person and their ingroup may therefore influence whether they amplify identity-related disinformation further within an SMP.

2.5.3. The Impact of Identity Threats on Disinformation Spread

According to Social Identity Theory, threats to identity can negatively impact the self-concept and may lead people to engage with strategies such as social creativity and social competition to help restore self-esteem (Ellemers et al., 2002). However, how threats manifest in the context of disinformation and SMPs may vary across specific circumstances. First, the impact of an external crisis on identity and the spread of targeted disinformation will be considered. Next, identity-threats within disinformation are discussed before addressing the threat of the "disinformation crisis" itself on its SMP spread.

Disseminators of disinformation have been known to engage in coordinated SMP posting during and following crisis events such as terrorist attacks (Innes, 2020). It may be that the "realistic" threats created by such incidents lead people to be more vulnerable to disinformation. For instance, a study run six months before and four days after 9/11 found American students experienced increased identification with their country and university directly following the attacks (Moskalenko et al., 2006). As strength of identification may also influence how people evaluate and spread content within SMPs (F. J. Jennings et al., 2020), such changes may influence their vulnerability to disinformation. Furthermore,

disinformation has also been targeted at divisions within society (DiResta et al., 2019). Research suggests that when the moral value of the ingroup is threatened (a "symbolic" threat), people may be more willing to share articles which are critical of the outgroup over those which favour the ingroup (Amira et al., 2021). Therefore, people may adapt their spread-related behaviours within SMPs in relation to threat-related situations.

Interactions with disinformation may also allow individuals to resolve negative emotions experienced as a result of identity threat. The Social Identity Model of Collective Action suggests affective injustice (often measured by group-based anger or resentment) is an important predictor of group-based action (van Zomeren et al., 2008). Furthermore, digital environments such as SMPs allow individuals to engage in collective action at lower costs (economic, and less effortful) than traditional political actions. One study suggested that publicly tweeting as a form of collective action in response to sexism leads to higher levels of hostility initially, but eventually leads to lower levels of hostility and better psychological wellbeing when compared to other affected group members (Foster, 2015). Therefore, people may be able to resolve feelings of affective injustice by engaging in SMP-based expression, which could arguably also include interacting with relevant disinformation content.

Another potentially important consideration is that "action" driven by disinformation during a time of crisis need not necessarily be antisocial. In a four-wave study during the COVID-19 pandemic, Ohme et al. (2021) found that holding higher levels of certain misinformation beliefs were related to individuals taking part in prosocial political participation (e.g. volunteering, donations, etc). While this will certainly not be the case for all types of misinformation beliefs, such findings suggest that disinformation may influence a person's beliefs about the state of the world, which could lead to action in times of crisis. Individuals may therefore engage in prosocial ways within SMPs in response to disinformation, which, given examples of fraudulent crowdfunding linked to disinformation, is another way that users may be vulnerable.

If, however, some users are aware that identity-congruent content is inauthentic during a time of crisis, expressing this within group spaces on SMPs may present issues. Identity-threats increase defensiveness to ingroup criticism (Adelman & Dasgupta, 2019) and may lead members to prioritise group loyalty over foundations of harm or fairness (Leidner & Castano, 2012). Therefore, individuals who violate group norms by drawing attention to the inaccuracies of ingroup disinformation during this time may experience greater repercussions.

However, group differences may play a role in how individuals respond to a fellow group member spreading disinformation in the face of a theat. In their study, Maxey (2021) found that Republicans increased support for a hypothetical Republican president when the truthfulness of their justification for military intervention in response to a security threat was challenged by experts. Democrats, instead, gave higher levels of support when the words of their own leader were supported by experts. Therefore, there may not be a universal response to how individuals interpret others spreading disinformation in the face of a crisis. Instead, other norms may be prioritised over the violation of "lying".

Moreover, information may itself threaten identity and influence how people respond to it. For example, it has been suggested that whether a group is threatened or affirmed by scientific research may influence evaluations of the research (Nauroth et al., 2017; Salvatore & Morton, 2021). Said evaluations are also thought to be related to the strength of emotional response to the research (Salvatore & Morton, 2021). This emotional connection may also explain why, for instance, strong group identifiers appear more likely to express their negative evaluations of identity-threatening science online (Nauroth et al., 2015). Such findings may also extend to disinformation content that threatens identity.

Prior experience of an identity-threat caused by disinformation may, however, lead people to be more conscious of its potential impact. People may perceive disinformation to be more prevalent and severe if an ingroup has been previously targeted (Chang, 2021). Furthermore, when known disinformation threatens ones' social identity it may itself be

viewed as hostile and, in turn, increase likelihood of commenting that that article is “fake” (E. L. Cohen et al., 2020). Additionally, focus group research with young Americans found that those who identified as “anti-Trump” felt an obligation to counter disinformation through their digital actions (Penney, 2020). Knowing that a piece of content is false or misleading may therefore attract reactions in some people that may ultimately influence its spread.

2.6. Conclusion

People may generally feel spreading disinformation is “wrong” to do, however, there may also be tendencies to associate disinformation susceptibility with other people; especially outgroups (e.g. political opposition). Furthermore, the perceived morality of spreading disinformation appears dependent on factors such as intention, perceived accuracy, as well as potential harm (to the self, others, and in relation to other issues). The flexibility of moral evaluations of disinformation in relation to identity will be addressed throughout the five studies in this thesis.

One question that has perhaps gained the most research attention in recent years is how and whether people can identify disinformation. People may be more susceptible to disinformation which supports their attitudes and beliefs or aligns with their pre-existing knowledge. In turn, several studies have indicated that perceiving the content to be accurate may increase the likelihood of users spreading it further. The relationship between beliefs and intentions to spread will be explored in both studies one and three of the present thesis.

Other research in this area has highlighted the tendency for people to make biased evaluations of disinformation in favour of an ingroup. In particular, there continues to be disagreement surrounding the reasoning processes underlying this, and whether such biases may be due to factors such as people forgetting to prioritise accuracy or making evaluations in an identity-protective manner. However, many of the studies discussed here

also illustrate how norms about disinformation, as well as emotional and behavioural responses to the content may be situation specific. Studies two, four and five will therefore explore how threats to identity presented within disinformation (e.g. no threat, threat to self, threat to group; as outlined by Ellemers et al. (2002)) help to explain intentions to spread the content further.

Chapter 3. Methodology

The following chapter outlines some of the core methodological decisions and approaches for this thesis. The epistemological position is first discussed, with a particular focus on how digital environments (e.g. social media platforms (SMPs)) may conflict with traditional definitions of "objective reality". Next, the chosen methodological approach is outlined. Here, the use of internet research methods is discussed, as well as the statistical analysis and open research practices used during the course of this thesis.

3.1. Epistemological Position

For information to be false or misleading it must, in some way, be untrue; therefore suggesting it deviates away from reality. But which reality? As the present thesis focuses on behaviour occurring within manufactured, digital environments, it is worth considering how this aligns with the concept of a singular, objective reality consisting of "physical objects, events, and forces" (APA Dictionary of Psychology, n.d.).

Social media use consists mostly of behaviour primarily performed within a digitised reality (Kaye et al., 2022). These environments may not reflect the "offline world", nor may users expect them to. Indeed, objects which exist in these digital environments are not "real" in a physical sense, although some be "simulated"(Krämer & Conrad, 2017). Users may also have their own perceptions of what (if anything) is "real" within these environments. However, due to ontological uncertainty (Brey, 2014) users may feel actions such as "sharing" are "real" (which may then influence behaviour). Furthermore, these manufactured environments are also algorithmically shaped by user feedback, and therefore can only ever be, at most, a subjectively framed simulation of reality. Differences in ontological status may therefore influence how content (including disinformation) situated within SMPs is perceived compared to information received via the "offline" world.

However, that is not to say that social media environments are entirely detached from an objective reality. User behaviour is itself initiated physically in the offline world before manifesting virtually. Brey (2014) suggests "virtual actions" can produce "intravirtual" effects that exist exclusively within the digital space (e.g. a Facebook "poke" does not involve the physical prodding of another human). Arguably, as an algorithmic weighting will often be assigned to these actions, the strength of related effects may be unstable and difficult to observe. However, Brey (2014) also suggests virtual actions may produce "extravirtual" effects which cross over into the "real" world. For instance, acts of deception hold real world significance, even when occurring within a digital environment. Additionally, virtual actions can produce physical effects, such as influencing the behaviours and emotions of others (Brey, 2014). By this reasoning, SMPs are not entirely separated from an objective reality.

For something to be disinformation it must deviate away from the truth, that is, what occurred in reality. Yet, what a person perceives to be truthful is not necessarily the same as that which is factual. As this research combines this with the epistemological complexity of digital environments, it could be argued that a purely positivist approach may not be appropriate. With that in mind, it would be reasonable to ask why mixed methods were not used here. This was mainly due to the need to address questions relating to moral evaluations within this thesis.

Notably, people are motivated to be seen as moral and so may make situation-dependent adjustments to expressed judgements (Rom & Conway, 2018). Participants may therefore attempt to anticipate what an interviewer perceives as favourable to provide a suitable (but not necessarily genuine) response. Specifically, the risk here was that participants may avoid giving interview responses that accurately reflect their judgements about disinformation so as not to be perceived as "immoral". Arguably, this may be further influenced by the tendency to underestimate ones' own vulnerability to disinformation (Y. Cheng & Chen, 2020; Corbu et al., 2020; Jang & Kim, 2018). However, there is evidence

to suggest that the use of online surveys (as have been used here) can help reduce social desirability bias compared to other mediums of data collection (Burkill et al., 2016; M. K. Jones et al., 2016). Given that 'spreading disinformation' is likely to be seen as an 'immoral' act, there may be some value in minimising researcher-participant contact via the use of online survey methods if it helps to reduce moral motivations.

Another consideration relates to the suggestion that the processes behind moral judgements and moral justification are not necessarily the same. Both the dual-path model of moral judgement (J. D. Greene et al., 2004) and the social intuitionist model (Haidt, 2001) suppose that moral judgements can occur independently from moral reasoning. Interviews may therefore pose problems for two reasons. Firstly, individuals are sometimes not able to explain their moral judgements, but an interview would require them to try (which could arguably result in manufactured responses). Secondly, as will be illustrated in study four, users may not necessarily engage in moral reasoning processes when using SMPs. However, an effective interview would require this from participants. The external validity of any 'moral'-related findings from an interview may therefore be questioned. As such, the studies presented here use self-report surveys where (as with using SMPs) participants may engage in either moral reasoning or moral intuition.

3.2. Methodological Approaches

3.2.1. Internet Research Methods

Data collection for this thesis took place on the online survey platform Qualtrics, with participants recruited via the crowdsourced recruitment platform Prolific. While in-person data collection would not have been possible during the earlier stages of this thesis (due to COVID-19 restrictions), there is a strong argument for using digital platforms to collect data on the judgements and intended behaviour of users within digital environments (e.g. SMPs) regardless.

Although survey platforms such as Qualtrics can only measure attitudes and perceptions as opposed to actual behaviour, there are important similarities as a medium to SMPs. The use of online surveys and SMPs are both considered "online exclusive" behaviours (Kaye et al., 2022): both are predominantly located within an online environment, require similar technological functions (e.g. clicking, scrolling), and the sharing of information with a remote audience. Additionally, participants may have accessed the studies using the same devices and been located within the same "offline" (e.g. physical) environments that they access SMPs on. From this perspective, online surveys may provide participants with an experience which is closer to that of using SMPs than other alternative data collection methods.

Furthermore, a key consideration here was the pace at which internet research methods can facilitate data collections. Public opinion, news cycles, and social media trends can rapidly change or decay, making data collection timeframes a priority. Crowdsourced recruitment platforms such as Prolific provide an efficient means for data collection in terms of both cost and time, particularly in the context of recruiting specific participant groups.

There are, of course, valid criticisms of recruitment platforms. For instance, rival platform Amazon's MTurk has previously been criticised for diminished data quality (Chmielewski & Kucker, 2020). However, evidence suggests that Prolific (the platform exclusively used here) has a considerably better quality of data than MTurk (Gupta et al., 2021; Peer et al., 2022; Uittenhove et al., 2023). Indeed, Prolific-sourced participants may produce data of a comparable quality to lab-based (Gupta et al., 2021) and web-based student samples (Uittenhove et al., 2023). Participant screening functions on Prolific and security settings on Qualtrics (including bot checkers) were also used to help further improve data quality here. Additionally, while an important ethical criticism of MTurk is their low participant payment rates (Williamson, 2016), the minimum rates of pay on Prolific are considerably higher (Palan & Schitter, 2018). Prolific and Qualtrics are not, of

course, the only methods of data collection which could have been used here or in the future. The use of alternative approaches is therefore explored in the general discussion.

3.2.2. Statistical Analysis and Open Research Practices

The main types of analyses used in this thesis were ANCOVA and multiple regression. Sample sizes were calculated using power analysis prior to data collection to ensure that relevant effect sizes were detectable. These were generally based on Ferguson's (2009) minimum recommendations, as effects of these size are thought to have potential real world implications. Assumptions for inferential tests were also checked, with any violations noted. Furthermore, several advanced analysis techniques (and corresponding software) were picked up during the course of this PhD. Firstly, PROCESS plug in for SPSS was used for mediation analyses in study three and conditional process analyses in studies 4 and 5. Study four also included linguistic analysis using the extended moral foundations dictionary (F. R. Hopp et al., 2021), requiring Python to generate the data. Additionally, analysis using R and Jamovi were also conducted.

The use of open research practices has developed alongside the thesis. The general aim of these practices is helping improve the accessibility and transparency of research. While all hypotheses, sample sizes, and planned analyses were decided ahead of data collection, studies 3-5 were pre-registered on AsPredicted. Furthermore, all data and syntax have been shared in line with open research data principles. Finally, the first paper arising from this thesis has been published in an open access journal, with data and syntax available via the Open Science Framework.

Chapter 4. Study One

This chapter begins by discussing the social media environments that disinformation[1] campaigns often target and some of the research that looks at why users may interact with strategically disseminated disinformation. The relationship between disinformation spread and beliefs is discussed, before exploring how people may prioritise the expression of such beliefs above accuracy. Finally, the potential influence of moral evaluations is considered. The relationship between beliefs, interaction and moral judgements of belief-consistent disinformation is tested through a series of multiple regressions. Exploratory analysis then looks at the influence of belief-consistency on individual digital interactions before testing for effects of group membership on interactions and moral judgements.

4.1. Introduction

The algorithmic ordering of feeds and microtargeting of advertising ensure individuals are presented with highly relevant information. Social Media Platforms (SMPs) are therefore commercial environments designed with users' attention in mind in addition to connectivity. As such, a typical feed may feature content from a mix of other users, businesses, and media organisations. Notably, almost half the UK adult population get their news through SMPs (Ofcom, 2022). SMPs are also the most popular news source for 16-35 year olds. These are also locations within which users can make purchasing decisions, discuss health issues, or find advice on employment. The use of SMPs as locations for information seeking may therefore mean people are particularly susceptible to disinformation within this context.

[1] While both "disinformation" and "misinformation" refer to false or misleading information, the former's creation and spread has an underlying purpose of deception (Digital Culture Media and Sport Committee, 2019). This could be for personal, political, or financial gain, or specifically to cause harm. If individuals unknowingly interact with this information, it is referred to as "misinformation". For the purpose of this chapter the term misinformation is used when either may apply.

Creators of disinformation take advantage of SMP spaces and features to disseminate misleading content to their intended audiences. At the most organised levels, a single disinformation campaign may produce a number of competing narratives around a single issue to target the common, yet specific, interests of different groups (Diresta et al., 2019). In 2020 (a year marked by major events including a global pandemic, police brutality protests and a US presidential election) the sheer variety of disinformation observed during this time demonstrates how disinformation, even around a single event, is not homogenous. Disinformation may also subtly blend in amongst genuine content and can be tailored to appeal to different beliefs or groups. There is arguably value in developing our understanding of how pre-existing beliefs may lead people to interact with belief-consistent misinformation, particularly during a crisis.

Recent critical analysis of disinformation research and policy identified a need for issue-specific approaches to disinformation research (Colley et al., 2020). Colley et al. (2020) highlighted links between the COVID-19 5G conspiracy theory and pre-existing, unsubstantiated beliefs around health risks of radio waves and microwaves. On one hand, many people may judge content featuring the 5G conspiracy to be “implausible” or extreme. However, the safety of radio and microwaves have been questioned in British society for around 40 years and prior to the pandemic, more than a third of the UK population believed electromagnetic frequencies to be carcinogens (Colley et al., 2020; Shahab et al., 2018). Concern surrounding the lack of official guidance regarding 5G implementation was also flagged by fact-checking organisations in early 2019 (Full Fact, 2020). For individuals who already felt these technologies are unsafe, seemingly harmless posts implying connections between 5G and COVID-19 may not only appear legitimate, but also important information to share. As such, pre-existing beliefs may aid the dissemination of belief-consistent misinformation during times of crisis.

4.1.1. Social Media Platforms as Social Environments

When using a platform, users may be more prone to thinking and acting like a group member when identity-relevant cues are present. The conditions of physical isolation and, under certain circumstances, anonymity afforded by these digital spaces may shift a user's focus to their social identities (Spears et al., 1990) and increase the likelihood of them conforming to group norms (Coppolino Perfumi et al., 2019). These deindividuation effects may lead judgements to become more polarised in favour of an ingroup (Spears et al., 1990), potentially even if users are exposed to both sides of an argument (E.-J. Lee, 2007). Individuals may then be more vulnerable to sharing or interacting with identity-related misinformation within SMPs than when offline.

Social media use is often categorised into "active" and "passive". The former describes behaviour where people interact with other users or accounts, for instance, by sharing information, leaving comments, or "liking" another user's post (Kross et al., 2021). When users consume information within an SMP but do not interact this is described as passive use. When users encounter content within SMPs their decisions to interact with it may be influenced by numerous factors. Arguably, these may also extend to interactions with misinformation. For instance, people are more likely to interact with misinformation posted by someone they trust (Buchanan & Benson, 2019) which may be expected given that SMPs are digital locations for building and sustaining relationships.

Furthermore, users may also decide to share a post based on how personally relevant it is (C.-C. Huang et al., 2009) or their attitudes towards it (J. Huang et al., 2013). They may also use low-cost actions such as "liking" to signal agreement or enjoyment (R. A. Hayes et al., 2016; S.-Y. Lee et al., 2016; Lowe-Calverley & Grieve, 2018; Sumner et al., 2017). While this mode of feedback is relatively effortless compared to commenting, "likes" are highly valued by some users (Chua & Chang, 2016) and can produce a sense of reward in the recipient (R. A. Hayes et al., 2016). People are also motivated to reward those who share similar beliefs to them (Allen & Wilder, 1975) and may avoid interacting

with posts by friends if the information conflicts with their beliefs (Sumner et al., 2017). Therefore, identity-related concerns may be prioritised above personal relationships within SMPs when required.

Compared to likes, actions such as sharing are a relatively more visible interaction. Research suggests SMP users may be less willing to express beliefs they perceive to be in the minority, particularly if they are controversial (Y. Liu et al., 2017). However, within more intimate digital settings (e.g. private groups, direct messages) these identity-concerns may be alleviated. For example, research suggests that people perceive fellow ingroup members to have similar attitudes towards childhood vaccinations than themselves, and that attitudes of outgroup members are more likely to differ (Rabinowitz et al., 2016). When a user perceives an audience to hold similar beliefs (e.g. a private group) any concerns for sharing belief-consistent posts may be reduced.

4.1.1.1 Strategic Targeting of Disinformation on Social Media. Disinformation campaigns have been known to curate and disseminate materials which target opposing sides of an issue (François et al., 2019) or even reframe a message to appeal to different political attitudes (DiResta et al., 2019). Using examples of memes that had previously been spread by the Internet Research Agency (IRA), Helmus et al. (2020) found participants who were politically left or right leaning were more likely to "like" memes when they were politically-concordant than politically-discordant. They were also most likely to "like" content which supported their country (e.g. the USA). This suggests that decisions regarding interactions may have been influenced by participants' social identity (e.g. national, political).

However, a proportion of participants also reported they were likely to interact or experience positive emotions in relation to politically-discordant memes (Helmus et al., 2020). Another consideration is that while certain stances may be associated with the values and attitudes of one end of the political spectrum, it does not necessarily reflect the

specific beliefs and priorities of every left-leaning or right-leaning individual. Models of political ideology may provide a way of explaining en-masse how individuals organise their political beliefs, but may not accurately represent individuals own personal belief systems (Feldman, 2013). For instance, Helmus et al. (2020) categorised pro-gun memes as "right-leaning". While this may reflect the issue's political leaning, it arguably does not reflect the actual beliefs of every politically right-leaning American. Indeed, around 20% of Republicans prioritise gun control over gun rights, while more Republicans also believe that gun control laws are "not strict enough" rather than "too strict" (31% vs 20%) (Schaeffer, 2019). While perhaps a minority, it does highlight the potential issues with conflating political-leaning or affiliation with specific beliefs.

Furthermore, the IRA has previously employed strategies to target specific communities and sub-cultures within SMPs. For instance, networks of inauthentic "sock puppet" accounts were found on several platforms that appropriated the identities of subcultural and activist groups in an attempt to generate trusting followings (François et al., 2019; Nimmo, François, Eib, & Ronzaud, 2020; Nimmo, François, Eib, Ronzaud, et al., 2020). These accounts often focused on specific issues, for example feminism, LGBTQ+ rights, the environment, pro-police, or southern confederacy. In a study looking at data based on interactions with these sock puppet accounts, Freelon et al. (2022) found accounts which inauthentically presented as Black American activists received more likes, retweets and comments than other categories (e.g. right-leaning, news accounts, hashtag gamers). They also indicated that when these accounts are merged with a general left-leaning category (as in previous studies) they potentially falsely inflate interactions. This again suggests that focusing on broadly defined ideological groups may not accurately represent relationships between misinformation and interaction behaviour.

However, the use of categorical (e.g. political party) (Helgason & Effron, 2022) and self-reported political orientation placement (Faragó et al., 2020; Helmus et al., 2020) as a representation of "political belief" is not uncommon in this area. Arguably, it may be

more effective to focus on the specific attitudes and beliefs of interest. Indeed, research that compared personal stance on abortion in relation to the position of misinformation about abortion found that individuals were more likely to believe and interact with stance-consistent misinformation, particularly when they felt the issue was important (A. Kim et al., 2019). Therefore, capturing and measuring specific beliefs that relate to the messaging within misinformation content may more accurately represent a users' likelihood of interacting than broader ideological categories or groups alone.

4.1.2. Personal Beliefs and the Spread of Misinformation

While research suggests that engaging in deliberative reasoning may help users identify disinformation, perceived plausibility appears to play an underlying role (Pennycook & Rand, 2019). However, disinformation can be extremely plausible and may be difficult to detect not least without motivation, skills, or relevant knowledge. To complicate things further, factually accurate but misleading content is increasingly being spread in attempts to counter AI detection systems (Wardle, 2019). Imposter, fabricated and manipulated content also make judgements based on "quality" cues more difficult, but also a growing reality with the rise of affordable and user-friendly image and video editing software. As such, it may be difficult for citizens for use their own judgement to verify misinformation and in some instances may lean towards perceiving it as legitimate.

Research suggests people are more likely to believe misinformation that supports their politics (Allcott & Gentzkow, 2017; A. Kim et al., 2019), and may generally be less sceptical of memes that align with their political views (Huntington, 2020). Moreover, other types of beliefs may increase susceptibility to misinformation. For instance, research suggests levels of trust in scientists may be an important predictor of increased susceptibility to COVID-19 related misinformation (Agley & Xiao, 2021; Roozenbeek et al., 2020; Su, 2021). Roozenbeek et al. (2020) found participants from five countries who reported lower trust in scientists judged misleading statements about COVID-19 to be

more reliable than individuals with higher trust. Comparably, other trust-related predictors (such as trust in government) only met significance in specific countries and were less important to the overall models. This is notable, given that the misinformation statements focused on the virus origin, contagion and cures which may clearly relate to "science". Furthermore, each country would have a specific "government" (each with its own approach to the pandemic) and therefore distinctions between countries may be more likely.

However, in contrast, others have found that higher trust in scientists may be related to increased acceptance and intentions to spread pseudoscience (O'Brien et al., 2021). Taken together, these findings would suggest that as with political beliefs, the relationship between beliefs and misinformation susceptibility is not necessarily one-way, but dependent on the "misinformation" being presented. Additionally, people are also more likely to share misinformation when they believe the content (Buchanan, 2020; Halpern et al., 2019; A. Kim et al., 2019; Pennycook, McPhetres, et al., 2020). Therefore, when people are presented with misinformation supporting specific beliefs they hold about the world, an issue or an event, they may not only be more likely to feel it is true, they may also spread it further.

4.1.3. Self-Expression, Disinformation and "The Truth"

It is somewhat important to consider that users spread information on SMPs for a variety of reasons, not simply because they know it to be accurate. Indeed, SMPs provide users with spaces within which they may express their personal realities with other likeminded individuals and limited gatekeeping. For instance, users may use imagery such as memes to express identity (Ask & Abidin, 2018; DeCook, 2018; Mahoney, 2020) or prejudice towards a person or a group (Andreasen, 2021; DeCook, 2018; Nee & De Maio, 2019; Stassen & Bates, 2020). Memes can also be used to express feelings towards a specific issue (Stassen & Bates, 2020). Therefore if users encounter misinformation that

supports their beliefs about the world, they may spread it further as a means of self-expression.

Beliefs can, however, be subjective. Indeed, beliefs may be understood as a particular outcome a person perceives to be "true" to varying levels of certainty (Huber, 2009). However, the specific degree of said outcome may not itself be objective as it will exist across a spectrum of potential outcomes. Arguably then, belief-consistent disinformation may represent something that feels true, even when people know it is factually inaccurate. As such, people may interpret belief-consistent disinformation as different in some way from "disinformation" generally. For instance, this might explain why people can associate concepts related to disinformation (e.g. "fake news") with those who hold political beliefs which are different to their own (Michael & Breaux, 2021; Tong et al., 2020; van der Linden et al., 2020). People at the extreme ends of the political spectrum are also thought to perceive their own beliefs to be superior or simply "correct" (Harris & Van Bavel, 2021; Toner et al., 2013). Therefore, the way that beliefs are held could influence people to adapt their evaluations of spreading disinformation that is belief-consistent, even if they know it is factually untrue.

Moreover, while individuals do generally report that the accuracy of information they share is important to them (Pennycook, Epstein, et al., 2021), they may still consider sharing false information after being informed it is factually inaccurate, for instance, through the presence of fact-checking tags (Oeldorf-Hirsch et al., 2020). Indeed, while such tags appeared to influence other judgements (such as perceptions of a website), Oeldorf-Hirsh et al. (2020) found fact-checking tags may do little to reduce credibility assessments of disinformation. This suggests there may be instances where the factual basis of disinformation is itself perceived as less important than the overall message being depicted and any personal desire to spread this further.

4.1.4. Morality and Misinformation

Accuracy may not be the sole reason why a person may refrain from interacting with misinformation. People may not interact with information they know to be false (and arguably, also other types of content) because they feel it is not moral to do so. Indeed, recent work has highlighted the potential importance of perceived morality (e.g. "right" and "wrong") for misinformation research. For instance, Effron & Raj (2020) found that repeatedly encountering disinformation reduces any moral condemnation of sharing it and, in turn, increases intentions to share (even when aware it is untrue). Thinking about whether disinformation could have been true (Effron, 2018) or might become true (Helgason & Effron, 2022) has a similar effect. Effron & Helgason (2022) suggest repetition and imagination may lead its "gist" (e.g. general idea) to seem truer (even if the information is still acknowledged as being otherwise factually incorrect) and as such is perceived as more acceptable to share. Arguably then, if people feel that the "gist" of belief-consistent disinformation is true, then this may lead them to be more morally lenient about spreading it further.

Morality also plays an important role in the self-regulation of behaviour. Specifically, the social cognitive theory of moral thought and action suggests violations of personal moral standards may lead to self-condemnation and, as such, people may regulate their behaviour to avoid this (Bandura, 1991a). As such, if people feel it is "wrong" to spread a piece of content they may refrain from doing so. However, while a person may feel spreading disinformation is "wrong" they may not always be informed or able to identify that disinformation is "disinformation". As such, they may not perceive any potential moral violations surrounding sharing belief-consistent misinformation. Indeed, users may not feel it is "untruthful" to spread misinformation if they perceive it to be accurate. Without any active intention to mislead others they may also be unlikely to feel it is "dishonest" (see Barber, 2020 for an overview). In contrast, the less consistent

misinformation is with a person's beliefs, the less "true" it might feel and therefore the more likely it may be perceived as being potentially "wrong" to spread.

Yet, as previously discussed, people may also be more morally lenient towards disinformation when it has been made to feel more "true" in other ways and, in turn, more likely to share it with others (e.g. Effron, 2018; Effron & Raj, 2020; Helgason & Effron, 2022). Other research suggests people may not necessarily view "dishonesty" as "lying" when doing so benefits others, perhaps because they view prosocial acts of dishonesty as more morally acceptable (Cantarero & Szarota, 2017). Therefore learning a post is "inaccurate" may not necessarily always mean it will be viewed as "wrong" to spread, even by people who feel that about disinformation generally. Indeed, people are able to minimise violations of moral standards when required. For instance, moral disengagement strategies, such as redefining the behaviour or disregarding the impact of actions, may allow users to protect their self-esteem when taking part in morally questionable behaviour (Bandura, 1999) such as cyberbullying (Meter & Bauman, 2018), cyberstalking (Fissel et al., 2021) and hacking (R. Young et al., 2007). Additionally, Heering et al. (2020) found people may legitimise potentially harmful digital behaviours if they feel the target is unresponsive to their own needs . Therefore, when users believe a certain outcome is true, they may find ways to justify the sharing of false information that otherwise supports said belief to help ensure it does not violate their moral standards.

4.1.5. The Present Study

The aim of the present study is to explore whether the consistency between issue-specific beliefs and misinformation can predict moral judgements and intentions to interact with the content. Data were collected in January 2021 while COVID-19 restrictions were still in place within the UK. Therefore, study materials were related to the pandemic. Misinformation from two over-arching topics were selected, and both of which then being divided into two opposing categories. These were misinformation that were "Favourable"

or “Unfavourable” towards the UK Government’s handling of COVID-19 and misinformation that sought to either “Minimise” or “Maximise” the threat of the virus itself.

This approach was selected as during the easing of the first lockdown restrictions within the UK Duffy & Allington (2020) observed social divisions across dimensions of government trust in relation to the handling of COVID-19 (high vs low) and risk perceptions of COVID-19 risks (high vs. low). One group associated with perceiving COVID-19 to be lower risk was also most likely to believe authorities had exaggerated mortality rates. Furthermore, Smith et al. (2020) found lower levels of concern and risk perception of COVID-19 were also associated with reduced adherence to self-isolation and lockdown guidelines. Notably, official fact-check websites identified a number of popular SMP posts which attempted to understate the risk of COVID-19, particularly in regards to restrictions (Allen-Kinross, 2020a; Krishna, 2020) and so it may be plausible that this type of misinformation could appeal to those who perceive COVID-19 to be lower risk.

Duffy & Allington's (2020) findings also indicate that a group associated with perceiving COVID-19 to be higher risk were also most likely to believe authorities were deliberately underreporting mortality rates. Notably, much misinformation identified as circulating on SMPs, particularly early on in the pandemic, did portray the virus as an extreme risk (AAP Factcheck, 2020; Panjwani, 2020b). Therefore it may be plausible that people who perceive COVID-19 to be a higher risk could also be more likely to share this type of information on SMPs, if only to communicate what they believe to be an accurate level of risk.

There were also examples during the course of the pandemic of misinformation that either positively framed or undermined the government. For instance, disinformation aimed at Boris Johnson (Fisher, 2020; Rana & O’Neill, 2020) but also examples of incorrect statistics and other information quoted by UK Government officials (Allen-Kinross, 2020b; Panjwani, 2020a). However, while there were clear partisan differences in

trust levels over the UK Government's handling of COVID-19 (B. Duffy & Allington, 2020; W. Jennings et al., 2020), individual trust levels within-each party were neither homogenous (W. Jennings et al., 2020) or stable (YouGov, 2021). It is because of this that the hypotheses will take a belief, as opposed to group, based approach to understand why individuals interact with misinformation.

It was predicted that individuals who hold issue-specific beliefs would be more likely to engage with misinformation that relates to aforementioned belief:

H1. Individuals who have lower trust in the UK Government's handling of COVID-19 (COVID-19) would report greater likelihood of interaction with misinformation that is unfavourable towards the government.

H2. Individuals who have higher trust in the UK Government's handling of COVID-19 would report greater likelihood of interaction with misinformation that is favourable towards the government.

H3. Individuals who believe COVID-19 to be lower risk would report greater likelihood of interaction with misinformation that minimises COVID-19 risk

H4. Individuals who believe COVID-19 to be higher risk would report greater likelihood of interaction with misinformation that maximises COVID-19 risk.

It was also predicted that individuals presented with disinformation that aligned with their beliefs would judge it more morally acceptable to share on social media than individuals for whom the disinformation opposed their beliefs:

H5. Individuals reporting lower trust in the UK Government's handling of COVID-19 would judge the sharing of disinformation that is unfavourable towards the government as more morally acceptable than individuals reporting higher trust in the government.

H6. Individuals reporting higher trust in the UK Government's handling of COVID-19 would judge the sharing of disinformation that is favourable towards the government as more morally acceptable than those reporting lower trust in the government.

H7. Individuals who believe COVID-19 to be lower risk would judge the sharing of disinformation that minimises COVID-19 risk as more morally acceptable than those who believe COVID-19 to be higher risk.

H8. Individuals who believe COVID-19 to be higher risk will judge the sharing of disinformation that maximises COVID-19 risk as more morally acceptable than those who believe COVID-19 to be lower risk.

4.2. Method

4.2.1. Development of Stimuli

To ensure that stimuli reflected relevant, real world disinformation, two overarching topics were selected that had been identified as disinformation narratives circulating on social media during the COVID-19 pandemic by fact-checkers (Brennan et al., 2020; John, 2020; Poytner, 2021). Disinformation themes that had been a major focus of media coverage (e.g. conspiracy theories, alleged cures, or anti-vaccine) were avoided. This was to reduce the likelihood of participants being already familiar with these disinformation narratives. In reality, it is usually up to the individual user to distinguish whether information posted within SMPs is factually correct or not (the exception being when a post is accompanied by a fact-check).

The first topic focused on the Government's handling of the crisis (i.e. "political"). The second theme was around the severity of the threat posed by the virus. A total of four stimuli sets were required for the study. These were "Favourable" or "Unfavourable" towards the UK Government and "Minimised" or "Maximised" the risk of COVID-19. A

small pilot study was therefore carried out to assign materials to each of these four categories.

4.2.1.1 Pilot Study

4.2.1.1.1 Materials. A selection of 16 items across 2 key topics were piloted. All items featured false or misleading information at the time the study took place (e.g. imagery used out of context or incorrect statistics). These either related to the performance of the UK Government or COVID-19 risk. The need for a balance of opposing stimuli for each theme was also considered during this initial selection stage (for example, four items depicting COVID-19 as high risk and four items as low risk). The stimuli were images sourced through fact-checking resources such as Full Fact, the Associated Press and Reuters, or taken directly from publicly available social media pages. For the latter, the information either closely mirrored another fact-check or a reputable source had been consulted to cross-check for accuracy.

4.2.1.1.2 Participants. 23 participants (6 males) aged 20-65 (M = 28.91, SD = 12.32) were recruited for the study via social media and were required to be residents of England. Ethical approval was obtained from the University's Psychology Ethics Committee ETH2021-0737 (Appendix A).

4.2.1.1.3 Procedure. The study was hosted online using the survey platform Qualtrics. Participants first answered basic demographic questions (age, gender, and location). Eight images related to the UK Government were presented (Figure 4.1), with participants asked to rate how favourable the images were towards the UK Government across an 11-point scale, from "Very unfavourable" to "Very favourable". Next, eight images relating to COVID-19 were presented (Figure 4.2), with participants asked to rate how high a risk the image made COVID-19 appear. This again used an 11-point scale from "Not at all risky" to "Very high risk". Finally, participants were thanked and debriefed.

Figure 4.1

Piloted Political Stimuli for Study One: "Favourable" and "Unfavourable" Towards the UK Government.

A – FG1

B – FG2

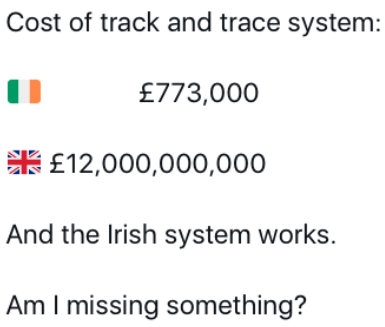
C – FG3

D - FG4

E – UG1

F – UG2

G – UG3



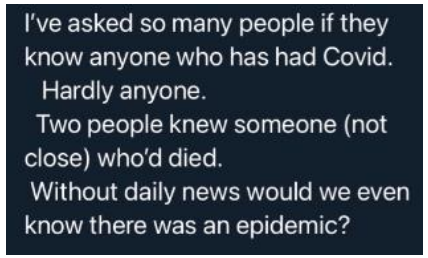
H – UG4

Note. Panels A-D feature misinformation "Favourable" towards the UK Government. Panels E-H feature misinformation "Unfavourable" towards the UK Government.

Figure 4.2

Piloted Virus Stimuli for Study One: "Minimising" and "Maximising" the Threat of the COVID-19 Virus.

A – MinCV1

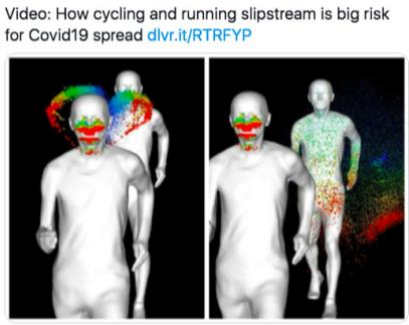


B – MinCV 2

C – MinCV 3

D - MinCV 4

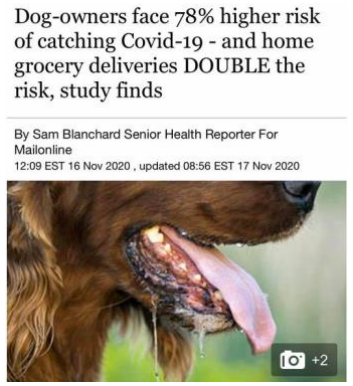
E – MaxCV1



F – MaxCV2

G – MaxCV3

H – MaxCV4



Note. Panels A-D feature misinformation that "Minimises" the threat of COVID-19. Panels E-H feature misinformation that "Maximises" the threat of COVID-19.

4.2.1.1.4 Results. Mean favourability scores for government-related stimuli are displayed in Table 4.1. Scores below 6 indicate content that was rated as unfavourable while scores over 6, favourable. One image fell into an unexpected category (FG3). The remaining images with mean scores above 6 were selected for the "Favourable" stimuli set. Of the remaining items, the three with the lowest mean favourability scores were selected for the "Unfavourable" set of stimuli.

Table 4.1

Mean Favourability Ratings of Political Stimuli

Item	Mean	*SD*	Minimum	Maximum
FG 1*	8.39	1.34	5	10
FG 2*	8.57	1.53	6	11
FG 3	3.91	2.23	1	9
FG 4*	8.57	1.62	5	11
UG 1*	2.04	1.02	1	4
UG 2*	1.96	1.61	1	6
UG 3*	1.96	1.30	1	6
UG 4	3.35	2.12	1	9

Note. $N = 23$. Abbreviation, FG = "Favourable towards Government", UG = "Unfavourable towards Government". *items in final selection.

For COVID-19 related images the three images with the highest and lowest mean risk perception scores were selected for "Maximising" and "Minimising" sets of stimuli (Table 4.2).

Table 4.2

Mean Risk Ratings of Virus-Related Stimuli

Item	Mean	*SD*	Minimum	Maximum
MinCV 1	3.43	1.88	1	8
MinCV 2*	3.09	2.15	1	10
MinCV 3*	3.04	2.14	1	8
MinCV 4*	3.22	1.70	1	7
MaxCV 1*	7.04	1.64	4	10
MaxCV 2	6.48	2.11	1	10
MaxCV 3*	9.35	1.58	5	11
MaxCV 4*	7.87	1.58	2	10

Note. $N = 23$. Abbreviations, MinCV = "Minimising COVID-19", MaxCV = "Maximising COVID-19". * items in final selection.

To check that the final pairs of stimuli sets were significantly different from one another in terms of "favourability" or "risk", the mean scores for each of the final stimuli set were calculated. A paired sample *t*-test showed that the difference between favourability ratings of Unfavourable ($M = 1.99$, $SD = 1.52$) and Favourable ($M = 8.51$, $SD = 1.33$) stimuli was significant ($t(22) = 13.99$, $p < .001$, $d = 2.91$). Furthermore, the difference between risk ratings for Minimising ($M = 3.12$, $SD = 1.59$) and Maximising ($M = 8.09$, $SD = 1.11$) misinformation was also significant ($t(22) = 11.83$, $p < .001$, $d = 2.47$. The large effect sizes observed here (J. Cohen, 1992) indicate that messages within the final selection of stimuli sets are distinctly different from their counterpart (e.g. in terms of favourability or risk).

4.2.1.1.5 Discussion. The final selection of images were identified as either the most Favourable or Unfavourable towards the UK Government, or presented COVID-19 as either the highest or lowest risk (and have been allocated as such). Across the four stimuli sets, each topic pair presented a message that was significantly different from its counterpart. Therefore this combination of stimuli grouping is suitable for use as four distinct "themes" in the main study.

4.2.2. Main Study

4.2.2.1 Participants

218 participants (85 males) aged 19-81 (M = 40.98, SD = 14.34) were recruited through Prolific to take part in the study. Ethical approval was obtained from the University's Psychology Ethics Committee ETH2021-0777 (Appendix B). Sample size was determined through a power analysis using G*Power, which indicated that 191 participants were needed to detect R^2 = .04 with 80% power. An effect size of r2 = .04 is a recommended minimum within social science research (Ferguson, 2009).

All participants were required to have an active Facebook account and currently be residing in England. This is because other nations within the United Kingdom have devolved governments who managed their own COVID-19 response. Furthermore, recruitment was split across four equal-sized groups using Prolific's screening tools to best ensure a balance of political views. Places on the study were therefore allocated based on political affiliation (e.g. identifying self as being left or right side of political spectrum) and Brexit vote (e.g. voted either to "Leave" the European Union or to "Remain"). Of the participants who took part in the study, the majority reported they would either vote for the Conservatives (*N* = 87) or Labour Party (*N* = 93) if the election were held the following day. For analysis where political parties are compared, participants who indicated anything other than these two parties were excluded from analysis due to small samples (Table 4.3).

Table 4.3

Participant Demographics for Study One

	N	%
Total	218	100.0
Gender		
Female	132	60.6
Male	85	39.0
Non-Binary	1	0.5

	N	%
Education completed		
Less than GCSEs	2	0.9
GCSEs	30	13.8
A-Levels	42	19.3
Bachelor's Degree	102	46.8
Master's Degree	38	17.4
Doctoral Degree	3	1.4
Other	1	0.5
Political Party		
Conservatives	87	40.0
Labour	93	42.7
Liberal Democrats	5	2.3
Other	16	7.3
Unsure	17	7.8

4.2.2.2 Materials and Measures.

4.2.2.2.1 Citizen Trust in Government Organisation Scale, Appendix C. Beliefs surrounding trust in the UK Government's handling of COVID-19 were assessed using the Citizen Trust in Government Organisation Scale (Grimmelikhuijsen & Knies, 2015). The original scale is provided as a template allowing questions to be tailored for specific issues. For example, *"When it concerns the handling of… 'the COVID-19 pandemic in the UK',… the government are capable".* A total of nine statements are included, with three statements relating to each dimension of trust (Competence, Benevolence and Integrity). Participants rated the level to which they agreed or disagreed with each statement using a 7-point Likert scale and an overall "Trust" score was created from the mean of all nine items summed.

When combined, the full scale had acceptable reliability (M = 3.40; SD = 1.60; α = .97). Across the two major political parties, Conservative voters reported significantly higher trust levels than Labour voters with a large effect size ($t(178) = 14.95$, $p < .001$, d = 2.22). This is to be expected as the Conservative party formed the Government at the time

of data collection, and reflects other findings that Conservative voters were more trusting of the Government's pandemic effort (B. Duffy & Allington, 2020; W. Jennings et al., 2020). This suggests the scales are valid in their ability to distinguish levels of trust.

4.2.2.2.2 COVID-19 Perceived Risk Scale, Appendix D. To measure participants' beliefs surrounding the risk of COVID-19, participants answered eight questions from the COVID-19 Perceived Risk Scale (Yıldırım & Güler, 2020), for example "*What is the likelihood that you would catch COVID-19?"*. Responses were collected using a Likert scale of 1 (Negligible) to 5 (Very High). This scale measures perceived risk across both cognitive and emotional dimensions. The full perceived risk scale had acceptable reliability (M = 3.02, SD = 0.66, α = .83).

Female participants also scored significantly higher than male participants on the full risk scale ($t(214) = 2.32$, $p < .05$, $d = .32$). Again, this was to be expected as both the original study (Yıldırım & Güler, 2020) and original SARS risk scale (Brug et al., 2004) reported similar findings.

4.2.2.2.3 Social Media "Interactions" with Misinformation. Participants were presented at random with each of the 12 stimuli selected from the pilot study. In total participants saw four categories of stimuli – "Favourable" or "Unfavourable" towards the UK Government, or "Minimising" or "Maximising" in regard to COVID-19 risk. However, at this stage of the study participants were not informed of the misleading nature of the information. To avoid social desirability effects, the participant invitation letter also made no reference to the veracity of the information presented in the study, only mentioning that the images were drawn from social media and not created by the researcher.

For each item, participants were asked "*If a Facebook friend posted this image, how likely is it that you would have "liked', shared privately (e.g. send to a friend or a private Facebook group) or shared publicly (e.g. to your own wall or newsfeed)".* Liking

and sharing are distinct behaviours with differing levels of effort required to interact with each (C. Kim & Yang, 2017). Furthermore, SMP users may engage with self-presentational strategies on social media, particularly surrounding divisive topics (Y. Liu et al., 2017). Users may therefore choose to share some content in more exclusive spaces (e.g. direct messaging, private groups). Ensuring that these private options for sharing are included in analysis may capture interactions that could otherwise be missed. Each behaviour ("liking", sharing privately, sharing publicly) was rated separately on a 7-point Likert scale from "Extremely unlikely" to "Extremely likely". For each stimuli set the responses for each behaviour were summed and a mean score created. Finally, combined, and averaged scores of all three behaviours were produced to create overall "interaction" scores.

4.2.2.2.4 Moral Judgements of Sharing Misinformation. For the final stage of the study participants were informed that the materials had been flagged as problematic by independent fact-checkers for being untrue or taken out of context. All 12 stimuli items were presented again, and participants were asked to judge how morally acceptable it was for others to share the post. Responses were given using a 7-point Likert scale from "Extremely unacceptable" to "Extremely acceptable". A mean score for each stimuli set was created.

4.2.2.3 Data Analysis

The present study employed multiple regressions to test the hypotheses. All tests applied an α level of .05. The dependent variable for H1-H4 was the relevant combined Interaction scores and for H5-H8 the corresponding Moral Judgement scores. To check the findings with non-parametric tests, Spearman's correlations were run following each regression. "Trust in Government", "Perceived Risk", gender and age were used as predictors in all regressions.

4.3. Results

Qualtrics data were exported into Excel for data cleaning before importing into SPSS. Four participants were removed from analysis due to not having Facebook accounts or lack of variance in answers. Responses for “Trust in Government”, “Perceived Risk” and moral judgements for each stimuli category were summed and mean scores calculated. To create interaction scores, each stimuli category had mean scores calculated for each interaction type (“Like”, “Privately Shared”, “Publicly Shared”) as well as a pooled mean score consisting of all three interaction types.

Descriptive statistics for all variables are shown in Table 4.4, with histograms and QQ plots of predictor variables provided in Appendix E.

Table 4.4

Summary of Descriptive Statistics by Misinformation Category

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skewness	Kurtosis
Age	218	40.94	14.37			19-81	0.49	−0.76
Trust in Government	218	3.40	1.60	.97	1-7	1.00-6.89	0.10	−1.22
COVID-19 Perceived Risk	218	3.02	0.66	.83	1-5	1.13-4.63	−0.23	−0.30
Favourable								
All - Interaction	218	2.09	1.32		1-7	1.00-6.56	1.33	1.00
Like	218	2.55	1.74		1-7	1.00-7.00	0.91	−0.42
Privately Shared	218	1.93	1.30		1-7	1.00-6.33	1.43	1.26
Publicly Shared	218	1.77	1.36		1-7	1.00-7.00	2.08	3.67
Unfavourable								
All - Interaction	217	2.29	1.48		1-7	1.00-7.00	1.37	1.42
Like	217	2.48	1.77		1-7	1.00-7.00	0.85	−0.32
Privately Shared	217	2.29	1.64		1-7	1.00-7.00	1.32	0.96
Publicly Shared	217	1.90	1.52		1-7	1.00-7.00	1.96	3.16
Minimising								
All - Interaction	218	1.82	1.29		1-7	1.00-6.67	1.78	2.27
Like	218	2.06	1.58		1-7	1.00-7.00	1.45	1.02
Privately Shared	218	1.82	1.31		1-7	1.00-6.67	1.72	2.21

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skewness	Kurtosis
Publicly Shared	218	1.59	1.28		1-7	1.00-6.67	2.44	5.23
Maximising								
All - Interaction	218	2.09	1.18		1-7	1.00-6.33	1.18	0.90
Like	218	2.22	1.31		1-7	1.00-6.67	0.93	0.15
Privately Shared	218	2.20	1.38		1-7	1.00-6.67	1.07	0.30
Publicly Shared	218	1.84	1.27		1-7	1.00-7.00	1.66	2.19
Moral Acceptability								
Favourable	218	3.48	1.65		1-7	1.00-7.00	0.20	−0.91
Unfavourable	218	3.02	1.55		1-7	1.00-7.00	0.57	−0.42
Minimising	218	2.44	1.51		1-7	1.00-6.67	1.01	0.09
Maximising	218	3.00	1.37		1-7	1.00-6.67	0.29	−0.45

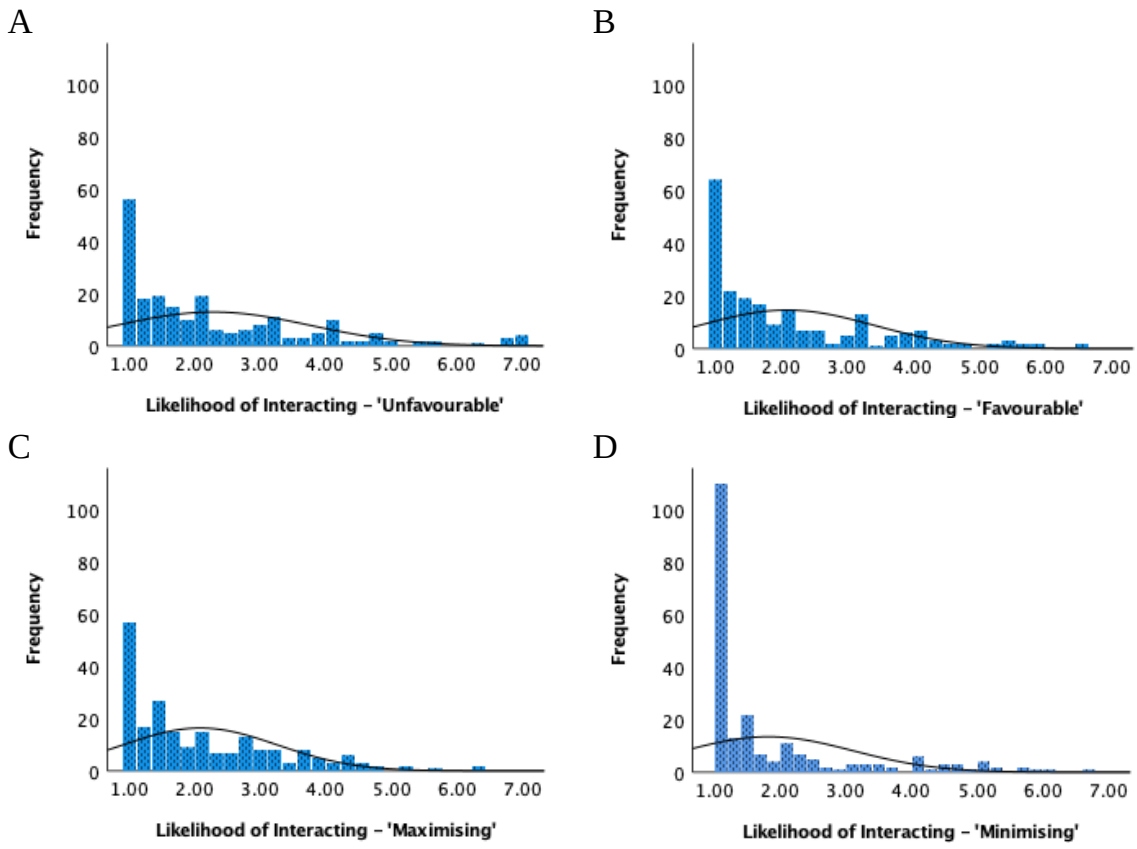
Variables were screened for reliability, normality, and homogeneity. There was evidence of skewness and kurtosis in a number of variables. Of the predictor variables, only Trust in Government showed evidence of negative kurtosis. However, while kurtosis may lead to an underestimate of variance within multiple regressions, Tabachnick & Fidell (2013) suggest this is only an issue with smaller samples (e.g. below 200) so may not impact the results here.

Several of the “interaction” variables and the moral acceptability score for “Minimising” disinformation showed evidence of positive skewness. Examination of the histograms (Figures 4.3 & 4.4) demonstrated floor effects, where high numbers of participants indicated it was “extremely unlikely” that they would interact with any of the stimuli presented from that set, in any manner. This pattern reflects not only the supposed sharing of misinformation (A. Guess et al., 2019) but also industry figures for interactions on SMPs generally (Kemp, 2020). As only 19 participants indicated they were “extremely unlikely” to interact with any of the 12 stimuli items in any manner (but provided responses with sufficient levels of variance in their moral judgements), one may assume that these are not simply floor effects but reflect a lack of intention to interact. Assessments of histograms for the interaction variables suggest that an inverse

transformation would not be appropriate given the proportion of participants who gave the lowest possible response across each set. To ensure that the distribution does not influence the findings, non-parametric tests will be used where possible to support the results.

Figure 4.3

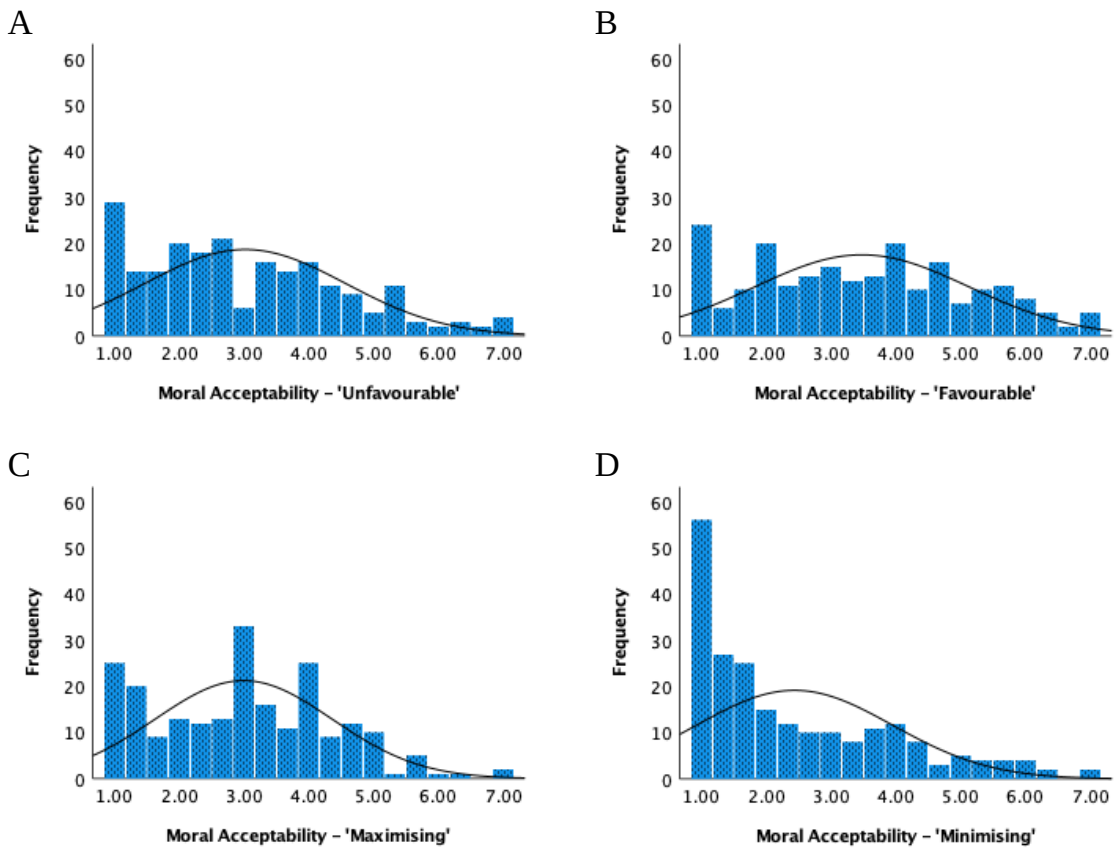
Histograms of Interaction Variables for Individual Misinformation Categories



Note. Panels A & B show “Unfavourable” and “Favourable” interaction scores respectively. Panels C & D show “Maximising” & “Minimising” interaction scores.

Figure 4.4

Histograms of Moral Judgement Variables for Individual Misinformation Categories



Note. Panels A & B show "Unfavourable" and "Favourable" moral acceptability scores respectively. Panels C & D show "Maximising" & "Minimising" moral acceptability scores.

4.3.1. Planned Tests

4.3.1.1 Effects of Belief Consistency on Interactions with Misinformation

To ensure the assumptions for multiple regression were not violated, preliminary analyses were conducted to assess normality, linearity, multicollinearity, and homoscedasticity. Any violations are noted within the results.

First, multiple regressions were carried out to assess whether people were more likely to interact with misinformation about the UK Government when it was consistent with their beliefs. Two models were run using "Trust in Government", "Perceived Risk",

age and gender as predictors. The first model predicted intentions to interact with "Unfavourable" misinformation, while the second predicted intentions to interact with "Favourable" misinformation. Assessments of P-P plots (Appendix F) suggest that the residuals for both models may not be normally distributed and therefore the results should be taken with caution.

The first model significantly predicted intentions to interact with Unfavourable misinformation, $F(4, 210) = 13.55$, $p < .001$, adj. $R^2 = .19$. While "Trust in Government", "Perceived Risk" and gender all added significantly to the model, Trust was the strongest predictor ($\beta = -.40$, $t(214) = -6.23$, $p < .001$) with an effect size above the recommended minimum effect sizes (RMPE) recommended by Ferguson (2009). As the relationship was negative, this suggests that lower levels of trust in the government's handling of the pandemic was associated with increased intentions to interact with misinformation that undermined the UK Government. H1 is therefore accepted.

The second model also significantly predicted interaction with Favourable misinformation, $F(4, 211) = 10.80$, $p < .001$, adj. $R^2 = .15$. "Trust in Government" and "Perceived Risk" added significantly to the model, however, again Trust was the strongest predictor ($\beta = .40$, $t(215) = 6.17$, $p < .001$). The effect size of Trust was similar to the first model; however, the relationship was instead positive, suggesting the level of consistency between belief and the message expressed by misinformation may be important for understanding intentions to interact. H2 is therefore accepted. Regression coefficients for both models can be found in Table 4.5.

Table 4.5

Summary of Regressions Predicting Interactions with Political Misinformation

	Unfavourable			Favourable		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	2.53***	.51		0.52	.46	
Age	0.01	.01	.11	−0.01	.01	−.06
Gender	−0.63**	.19	−.21	−0.24	.17	−.09
Trust	−0.37***	.06	−.40	0.33***	.05	.40
Risk	0.31*	.14	.14	0.27*	.13	.14
R^2		.21			.17	
Adj. R^2		.19***			.15***	
F		13.55			10.80	

Note. Gender coded as dummy variable, M = 0, F = 1.

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

Spearman correlations were run to support the findings. These confirmed that Trust had a significant, negative relationship with intentions to interact with Unfavourable misinformation ($r = -.40$) and that the relationship with Favourable misinformation was positive ($r = .39$). Both relationships had a medium effect size (J. Cohen, 1992).

Next, the same regression models were used to predict intentions to interact with "Minimising" and "Maximising" misinformation. Assessment of P-P plots for both models suggest that the residuals may not be normally distributed (Appendix F) and therefore again the results should be taken with caution. The first model did not significantly predict interaction with Minimising misinformation, $F(4, 211) = 1.19$, $p = .32$, adj. $R^2 = .004$. H3 is therefore rejected. However, the second model did significantly predict interactions with Maximising misinformation, $F(4, 211) = 3.20$, $p < .05$, adj. $R^2 = .04$. Only "Risk" added significantly to this model and was above RMPE (β = .21, $t(215) = 3.09$, $p < .01$). As the relationship was positive, this suggests that greater perceived risk was associated with

increased intentions to interact with misinformation that presented COVID-19 as a higher risk. H4 is therefore accepted. Regression coefficients can be found in Table 4.6.

Table 4.6

Summary of Regressions Predicting Interactions with Virus Misinformation

	Minimising			Maximising		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	1.83	.49		0.73	.44	
Age	0.01	.01	.10	0.003	.01	.04
Gender	−0.16	.18	−.06	−0.19	.16	−.08
Trust	0.03	.06	.03	0.06	.05	.08
Risk	−0.13	.14	−.07	0.38**	.12	.21
R^2		.02			.06	
Adj. R^2		.004			.04*	
F		1.19			3.20	

Note. Gender coded as dummy variable, M = 0, F = 1.

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

Spearman correlations again confirmed that there was a small but significant relationship between perceived Risk and interaction with Maximising misinformation (r = .18). There was no significant relationship with Minimising misinformation.

4.3.1.2 Effects of Belief Consistency on Moral Judgements of Disinformation

Further multiple regressions were run to understand whether people are more morally lenient towards false information that is consistent with their beliefs. The previous models were used to predict moral judgements of sharing “Unfavourable” and “Favourable” disinformation. Assessment of the P-P plots suggest that the residuals for the Unfavourable disinformation model may not be normally distributed (Appendix F).

The first model significantly predicted moral acceptability ratings of Unfavourable disinformation, $F(4, 211) = 12.75$, $p < .001$, adj. $R^2 = .18$. Trust in Government and gender

both added significantly to the model, with Trust in Government being the strongest predictor (β = −.38, $t(215) = -5.92$, $p < .001$). Again the relationship was negative, suggesting that lower levels of trust were associated with higher ratings of moral acceptability for spreading disinformation that undermined the UK Government. H5 is therefore accepted.

The second model also significantly predicted moral acceptability ratings of Favourable disinformation, $F(4, 211) = 4.12$, $p < .01$, adj. $R^2 = .06$. Here, only Trust in Government added significantly to the model and was above Ferguson's (2009) recommended minimum for effect sizes (β = .24, $t(215) = 3.50$, $p < .01$). As before, the relationship was positive, suggesting that belief consistency may be important for understanding how people make moral evaluations of false information. H6 is therefore accepted. Regression coefficients for both models can be found in Table 4.7.

Table 4.7

Summary of Regressions Predicting Moral Judgements of Political Disinformation

	Unfavourable			Favourable		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	5.55***	.53		3.50***	.61	
Age	−.01	.01	−.08	−0.01	.01	−.09
Gender	−.49*	.20	−.15	−0.38	.23	−.11
Trust	−.37***	.06	−.38	0.25**	.07	.24
Risk	.21	.15	−.09	−0.07	.17	−.03
R^2		.20			.07	
Adj. R^2		.18***			.06**	
F		12.75			4.12	

Note. Gender coded as dummy variable, M = 0, F = 1

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

Moral acceptability ratings of Unfavourable disinformation sharing significantly correlated with Trust with medium effects ($r = -.38$). Furthermore, Trust positively

correlated with the moral acceptability ratings of Favourable disinformation with small effects (r = .22).

Two final regression models were carried out to assess whether perceptions of COVID-19 risk could predict the moral judgements of spreading Minimising and Maximising disinformation. Assessment of the P-P plots suggest that the residuals may not be normally distributed (Appendix F). The model significantly predicted moral acceptability ratings of Minimising disinformation, $F(4, 211) = 6.34$, $p < .001$, adj. $R^2 = .09$. Only "Perceived Risk" added significantly to the model and this was above recommended minimums ($\beta = -.30$, $t(215) = -4.42$, $p < .001$). This suggested that levels of perceived risk of COVID-19 are negatively associated with moral acceptability judgements of spreading disinformation that attempted to minimise the risk of COVID-19. H7 is therefore accepted. However, the model did not significantly predict moral acceptability ratings of Maximising disinformation $F(4, 211) = 1.66$, $p = .15$, adj. $R^2 = .01$. H8 is therefore rejected. Regression coefficients can be found in Table 4.8.

Table 4.8

Summary of Regressions Predicting Moral Judgements of Virus Disinformation

	Minimising			Maximising		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	4.66***	.55		3.69	.51	
Age	−0.003	.01	−.03	−0.01	.01	−.11
Gender	−0.25	.21	−.08	−0.40	.19	−.14
Trust	0.03	.06	.04	0.002	.06	.002
Risk	−0.68***	.15	−.30	−0.01	.15	−.01
R^2		.11			.03	
Adj. R^2		.09**			.01	
F		6.34			1.66	

Note. Gender coded as dummy variable, M = 0, F = 1

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

Spearman correlations confirmed that Risk had a medium sized significant relationship with moral acceptability of sharing Minimising disinformation ($r = -.32$) but no significant correlational relationship was found for Maximising disinformation.

4.3.2. Exploratory Analyses

4.3.2.1 Effects of Belief Consistency on Distinct Interaction Types.

To understand how belief consistency influenced intentions to engage with misinformation in specific ways (e.g. liking, sharing privately, sharing publicly) a series of multiple regressions were run. Again, this data had skewness and P-P plots suggest that the residuals were not normally distributed and therefore the results should be taken with caution. For misinformation about the UK Government, the models predicting the likelihood of Liking "Unfavourable" or "Favourable" misinformation accounted for 22% and 19% of variance respectively. This reduced to 15% and 9% for sharing privately, and 11% and 9% for sharing publicly. Trust remained the strongest predictor across all the models, and again the direction of its relationship with the interaction type was dependent on the misinformation being viewed (e.g. Unfavourable or Favourable). Regression coefficients are displayed in Tables 4.9 and 4.10.

Table 4.9

Summary of Regressions by Interaction Type with Unfavourable Misinformation

	Like			Share - Privately			Share - Publicly		
	B	*SE B*	β	*B*	*SE B*	β	*B*	*SE B*	β
Model									
Constant	3.71***	.59		2.28***	.57		1.16**	.54	
Age	0.001	.01	.01	0.01	.01	.12	0.02**	.01	.19
Gender	−0.78**	.22	−.22	−0.62**	.21	−.19	−0.50*	.20	−.16
Trust	−0.48***	.07	−.43	−0.36***	.07	−.35	−0.27***	.06	−.29
Risk	0.33*	.17	.12	0.35*	.16	.14	0.23	.15	.10
R^2		.24			.17			.13	
Adj R^2		.22***			.15***			.11***	
F		16.37			10.75			7.69	

Note. Gender coded as dummy variable, M = 0, F = 1.

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

Table 4.10

Summary of Regressions by Interaction Type with Favourable Misinformation

	Like			Share - Privately			Share - Publicly		
	B	*SE B*	β	*B*	*SE B*	β	*B*	*SE B*	β
Model									
Constant	0.49	.59		0.77	.47		0.30	.49	
Age	−0.01	.01	−.10	−0.01	.01	−.06	0.001	.01	.01
Gender	−0.24	.22	−.07	−0.27	.18	−.10	−0.20	.18	−.07
Trust	0.50***	.07	.46	0.25***	.06	.30	0.25***	.06	.30
Risk	0.34*	.17	.13	0.24	.13	.12	0.24	.14	.12
R^2		.21			.10			.10	
Adj R^2		.19***			.09***			.09***	
F		13.82			6.08			6.02	

Note. Gender coded as dummy variable, M = 0, F = 1.

*p < .05. ** p < .01. *** p < .001.

Furthermore, Risk was the only significant predictor in all three "Maximising" models. However, the model predicting the public sharing of Maximising misinformation was significant, accounting for only 5% of variance. None of the models predicting interactions with "Minimising" misinformation were significant. Regressions coefficients are displayed in Tables 4.11 and 4.12.

Table 4.11

Summary of Regressions by Interaction Type with Maximising Misinformation

	Like			Share - Privately			Share - Publicly		
	B	*SE B*	β	*B*	*SE B*	β	*B*	*SE B*	β
Model									
Constant	1.14*	.49		0.87	.52		0.16	.47	
Age	0.00	.01	−.004	0.00	.01	.004	0.01	.01	.10
Gender	−0.19	.18	−.07	−0.18	.19	−.06	−0.20	.18	−.08
Trust	0.08	.06	.09	0.04	.06	.05	0.07	.06	.09
Risk	0.32*	.14	.16	0.43**	.15	.20	0.40**	.13	.21
R^2		.03			.04			.07	
Adj R^2		.02			.03			.05**	
F		1.86			2.38			4.08	

Note. Gender coded as dummy variable, M = 0, F = 1.

*p < .05. ** p < .01. *** p < .001.

Table 4.12

Summary of Regressions by Interaction Type with Minimising Misinformation

	Like			Share - Privately			Share - Publicly		
	B	*SE B*	β	*B*	*SE B*	β	*B*	*SE B*	β
Model									
Constant	2.49***	.59		1.96***	.49		1.05*	.48	
Age	0.01	.01	.07	0.01	.01	.06	0.01*	.01	.16
Gender	−0.09	.22	−.03	−0.23	.19	−.08	−0.18	.18	−.07
Trust	0.06	.07	.06	0.01	.06	.01	0.01	.06	.01
Risk	−0.30	.17	−.13	−0.09	.14	−.05	0.01	.14	.004
R^2		.03			.02			.03	
Adj R^2		.01			−.004			.02	
F		1.50			0.78			1.83	

Note. Gender coded as dummy variable, M = 0, F = 1

*p < .05. ** p < .01. *** p < .001.

4.3.2.2 Relationships Between Interactions and Moral Judgements

A series of Spearman's correlations (Table 4.13) were run to explore the relationships between reported likelihood of interacting with misinformation (with no prior accuracy knowledge) and moral judgements of spreading disinformation (upon learning the information was false or misleading).

Table 4.13

Correlations between Interactions and Moral Judgements by Misinformation Category

Interaction Likelihood	Moral Acceptability			
	Unfavourable	Favourable	Minimising	Maximising
Unfavourable	.48***	.03	.09	.20**
Favourable	−.02	.36***	.20**	.18**
Minimising	.21**	.31***	.51***	.22**
Maximising	.17*	.19**	.07	.32***

* p < .05. ** p < .01. *** p < .001.

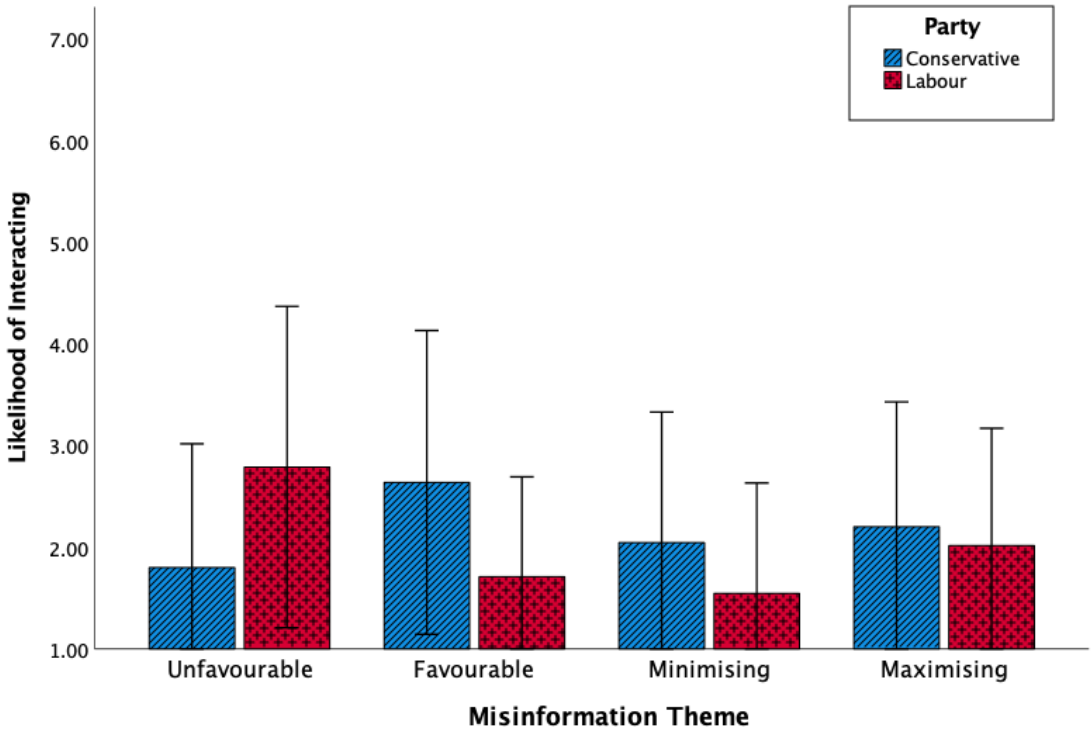
There were significant correlations between the interaction scores and moral judgements of corresponding stimuli, both with medium and large effect sizes (Cohen, 1992). This suggests people who indicated a greater likelihood of interacting with one category of misinformation may have also been more morally lenient towards sharing it upon learning it was untrue. Furthermore, there were significant relationships between intentions to interact with Minimising misinformation and moral judgements of category of disinformation, although the effect sizes varied from small to large. However, this may have been in part driven by floor effects in moral acceptability scores for Minimising disinformation.

4.3.2.3 Political Differences in Interactions and Moral Judgements

There were a number of differences between responses provided by Conservative and Labour voters. Firstly, Conservative voters were significantly more likely to intend to interact with Favourable misinformation (M = 2.64, SD = 1.49) than Labour voters (M = 1.71, SD = 0.98), with medium effects ($t(147.09) = 4.89$, $p < .001$, $d = .74$). In turn, Labour voters reported significantly greater likelihood of interacting with Unfavourable misinformation (M = 2.79, SD = 1.80) than Conservative voters (M = 1.80, SD = 1.22), again with medium effects ($t(171.53) = -4.71$, $p < .001$, $d = .70$). Additionally, Conservative voters reported a greater likelihood of interacting with Minimising misinformation (M = 2.05, SD = 1.28) than Labour voters (M = 1.54, SD = 1.08), although the difference was small ($t(168.88) = 2.81$, $p < .01$, $d = .42$). Any difference between the likelihood of interacting with Maximising misinformation for Conservative (M = 2.20, SD = 1.23) and Labour voters (M = 2.02, SD = 1.15) was not significant ($t(178) = 1.04$, p = .30). Means and standard deviations are displayed in Figure 4.5.

Figure 4.5

Mean Likelihood of Interacting with Misinformation Split by Political Party.



Note. Means with standard deviations displayed.

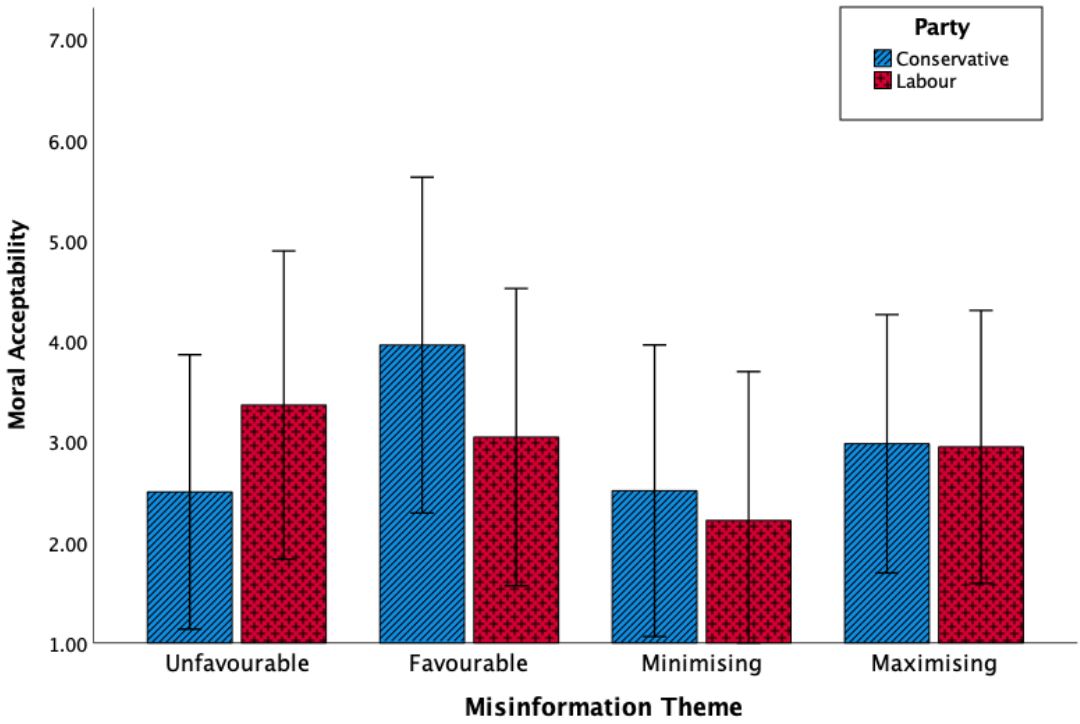
A similar pattern emerges for the moral judgements of disinformation. Labour voters judged spreading Unfavourable disinformation (M = 3.37, SD = 1.53) to be significantly more acceptable to spread than Conservative voters (M = 2.50, SD = 1.36), $t(178) = -3.98$, $p < .001$, $d = .59$. In turn, Conservative voters judged Favourable disinformation to be more morally acceptable to spread (M = 3.96, SD = 1.67) than Labour voters (M = 3.05, SD = 1.48), $t(178) = 3.90$, $p < .001$, $d = .58$). This suggests both groups judged ingroup benefitting disinformation to be more acceptable to spread than their opposition did. No other between-group differences in moral judgements were significant.

However, paired t-tests revealed that Conservative voters reported the sharing of Favourable disinformation (M = 3.96, SD = 1.67) to be significantly more acceptable to spread than Unfavourable disinformation (M = 2.50, SD = 1.36), ($t(86) = 8.45$, $p < .001$, $d = .91$). Yet, any difference between the moral judgements of Favourable (M = 3.05, SD =

1.48) and Unfavourable disinformation (M = 3.37, SD = 1.53) for Labour voters was not significant ($t(92) = -1.94$, $p = .06$). Means and standard deviations are found in Figure 4.6.

Figure 4.6

Mean Moral Acceptability of Disinformation Split by Political Party.



Note: Means with standard deviations displayed.

4.4. Discussion

The primary aim of this study was to determine the influence of belief-consistency on the likelihood of interacting with misinformation. Additionally, the study explored whether these issue-specific beliefs influence moral judgements of disinformation. The findings support both H1 and H2, in that trust in the UK Government's handling of the COVID-19 pandemic significantly predicted interaction with government related misinformation. Specifically, higher levels of trust predicted increased likelihood of interacting with "Favourable" misinformation, while lower levels of trust predicted increased likelihood of interacting with "Unfavourable" misinformation. It was also observed that perceived levels of COVID-19 risk played a smaller but significant role in

predicting interaction, with higher levels of perceived risk predicting increased likelihood of interacting in both instances. Altogether, the models accounted for 15% and 19% of variance in increased likelihood of interaction with Favourable and Unfavourable misinformation respectively. The findings also supported H4, in that heightened perceived risk of COVID-19 positively predicted greater likelihood of interacting with "Maximising" misinformation. However, H3 had stated that lower levels of perceived risk would predict increased likelihood of interacting with "Minimising" misinformation. The model was found not to be significant.

Both H5 and H6 were also supported by the findings. Trust negatively predicted moral judgements of sharing Unfavourable disinformation. The model, which also included gender as a significant factor, accounted for 18% of variance. Additionally, trust positively predicted moral judgements of sharing Favourable disinformation and accounted for 6% of variance. Furthermore, as predicted in H7, lower perceptions of COVID-19 risk positively predict increased acceptance of sharing Minimising disinformation. In other words, participants who perceived COVID-19 as lower risk may also have viewed the sharing of disinformation supporting this belief as more morally acceptable than those who perceive COVID-19 as high risk. Higher levels of perceived risk did not predict moral judgements of sharing Maximising disinformation as suggested in H8.

In line with previous findings, the present results show that social media users may be more likely to share or engage with misinformation when it confirms or supports their beliefs (A. Kim et al., 2019). Furthermore, it also supports findings about interactions with social media content generally, whereby users are more likely to engage with social media content when it is personally relevant (R. A. Hayes et al., 2016; S.-Y. Lee et al., 2016; Lowe-Calverley & Grieve, 2018; Sumner et al., 2017). Therefore, there may arguably be similarities between how users interact with misinformation and content generally.

Importantly, the present study highlights challenges in treating relationships between specific beliefs and interactions with misinformation as one-way. Instead these

findings indicate that the sentiment within misinformation messages matter. While causation cannot be inferred due to correlational nature of the study, those with beliefs either end of the trust spectrum saw increased likelihood of interacting with misinformation that appealed directly to those beliefs. This may help to bring context to previous findings that showed trust in scientists to have a negative relationship with susceptibility to COVID-19 misinformation (Agley & Xiao, 2021; Roozenbeek et al., 2020; Su, 2021) but a positive relationship with susceptibility to pseudoscience (O'Brien et al., 2021). Rather than trust being a predictor of misinformation susceptibility generally, the present findings suggest that belief-consistency may also explain this divergence.

However, interpreting data in the context of belief-consistency may also prove useful in other ways. In the present study the relationships between beliefs surrounding the risk of COVID-19 and intentions to interact with Minimising and Maximising misinformation was less clear. While levels of belief-consistency appeared help explain interactions with Maximising misinformation, the actual variance accounted for was small (4%). Furthermore, perceived risk did not predict interactions with Minimising misinformation (however, interestingly it did predict moral judgements of Minimising disinformation, suggesting that belief-consistency was not entirely irrelevant here). This is where external factors may play a role, as data was collected almost a year into COVID-19 measures in the UK. Notably almost twice as many participants reported it was "extremely unlikely" they would interact with Minimising misinformation than for any other misinformation category. Moreover, upon learning the posts were misleading participants felt sharing Minimising misinformation was less acceptable than other misinformation types. It may therefore be the case that this model did not reach significance as people may have intended to refrain from interacting with Minimising misinformation, even if it was belief consistent, perhaps because they sensed others perceived the topic as potentially controversial. Indeed, previous work on the spiral of silence suggests that people avoid sharing their beliefs on social media when they perceive said beliefs to be held by the

minority (Y. Liu et al., 2017). Arguably, public awareness and education about misinformation narratives may therefore have an important, but potentially indirect, impact on reducing spread on social media. However, it is also important to acknowledge that the lower response rates of Minimising misinformation may be related to participant sampling. Indeed, certain groups distrusting of academic institutions may be more difficult to recruit for studies (J. C. Young, 2021). Therefore it may also be that those who may otherwise interact with information that minimises the perceived impact of COVID-19 are less likely to engage with academic research generally.

Moreover, such findings may have important methodological implications for misinformation research generally, as they highlight the importance of context. For instance, as previously discussed, the present work illustrates the potential influence of distinct misinformation narratives on findings, as well as the potential importance of public awareness of such narratives at certain points in time. As such, future work may wish to consider whether and how susceptibility to misinformation evolves across time. Moreover, not only may considerations of belief-consistency help develop understanding of why people spread misinformation beyond broad, group-based associations and political attitudes, it may help with interpreting culture-specific associations. For instance, Roozenbeek et al. (2020) found participants in the USA, Mexico and Spain who reported higher levels of trust in politician's response to COVID-19 were also more susceptible to COVID-19 misinformation. While the misinformation was essentially unrelated to politics, both presidents of the USA and Mexico (where the effect was strongest) had been heavily criticised for misleading their citizens about COVID-19 (Evanega et al., 2020; Human Rights Watch, 2020) which may explain the association between high trust in politicians and susceptibility to virus-related misinformation. In the context of the present findings, it could be argued that it is not necessarily "low trust" that makes people susceptible to misinformation, but rather that specific narratives within misinformation are perhaps

appealing to what a person perceives as being true (e.g. their beliefs, which may indeed include low levels of trust in an institution).

Indeed, on the whole, participants also felt it was more morally acceptable to spread disinformation (e.g. information they learnt was untrue) when it was more consistent with their beliefs. In other words, they may be more morally lenient about spreading belief-consistent disinformation than other people. As beliefs represent outcomes that a person perceives to be in some way true (Huber, 2009), belief-consistent disinformation may in some way "feel" accurate, even if the factual basis is not. This supports recent work suggesting that people may be more accepting of spreading disinformation when its "gist" (e.g. general idea) feels true (Effron & Helgason, 2022). Furthermore, while it has also been argued that one reason people spread misinformation is because they don't consider accuracy (Pennycook, Epstein, et al., 2021), here participants were asked to make moral judgements after learning that the previously viewed misinformation was inaccurate (e.g. disinformation). Therefore, while people may report that the accuracy of the information they share online is important to them (e.g. Pennycook et al., 2021), the present findings suggest that such concerns may potentially be selectively applied to moral judgements in relation to the degree of belief-consistency of disinformation.

Notably, there was no association between level of perceived risk and moral judgements of spreading Maximising disinformation. However, compared to the other misinformation categories, there may have been some moral ambiguity. While many participants judged the sharing of Minimising disinformation as being "extremely unacceptable", there was greater variance in responses regarding sharing Maximising disinformation. However, in contrast, the associations for political disinformation were much clearer. While not possible to ascertain from the present findings, one potential reason for this may be that the perceived benefits of spreading Maximising disinformation conflicted with other moral considerations, especially for those who perceived the severity

of COVID-19 to be high. For instance, some people who perceive COVID-19 to be higher risk may have also been concerned about the impact of disinformation during the crisis. However, as perceiving COVID-19 to be a greater risk was also associated with adherence to COVID-19 guidelines (L. E. Smith et al., 2020) perhaps others felt the Maximising disinformation, albeit false, may encourage others to do the same. Rather than suggest that people who perceived COVID-19 to be higher risk may be more willing than others to spread disinformation, this example illustrates how disinformation may have the potential to produce moral dilemmas. Specifically, when disinformation targets a specific concern that people feel is morally important it has the potential to outweigh moral concerns relating to accuracy. Future work is therefore needed to better understand the impact of moral dilemmas presented by disinformation on spread.

At first glance, it may appear that only a small proportion of participants reported that they may interact with misinformation, but this was somewhat to be expected. Industry figures estimate that for every 100 views an image receives on Facebook it will receive four engagements on average (e.g. likes, shares, etc) (Kemp, 2020). Similarly, typical UK-based Facebook users will only share one post and "like" 16 posts in an average month (Kemp, 2020). In the present study, the proportion of participants indicating they would likely interact varied across misinformation types, with between 11%-22% of participants reporting they may "like" a post (the most favoured interaction on the whole, reflecting normal social media behaviour). Unlike previous studies, here, sharing "privately" and "publicly" were distinguished. Given the rise of misinformation in private Facebook groups and direct messaging services, providing participants with distinct options may be valuable for external validity reasons. However, notably, when predicting interaction with Maximising misinformation only the model for "share publicly" was significant. Compared to the other misinformation categories this was somewhat unexpected but may suggest a desire to inform a wider audience driven by said belief, rather than simple "agreement" or need to inform a limited few. Future research may therefore wish to

explore how different misinformation categories influence engagement with specific digital interactions,

However, where beliefs could reasonably be assigned to groups (e.g. political party), over a third of participants may have been willing to interact with misinformation that may have benefitted the ingroup. For instance, more than half of participants who vote Labour reported some intention to like misinformation comparing the cost of track and trace systems in the UK and Ireland (using an incorrect value for the Irish system). Conversely, 45% of participants who vote for the Conservatives reported some intention to like a post claiming the UK was testing more than anywhere else in Europe (which at the time was incorrect). These findings support previous work suggesting people are more likely to interact with misinformation that aligns with their political leaning (Helmus et al., 2020). From a social identity theory perspective (Tajfel & Turner, 2004) such interactions may also allow users to express positive aspects of an ingroup (e.g. Conservative voters interacting with Favourable misinformation) or allow them to engage in social comparison strategies (e.g. Labour voters interacting with Unfavourable misinformation) as a means of achieving or maintaining a positive self-concept.

However, after participants learnt the content was untrue the symmetry found for identity-related intentions to spread was lost. On one hand, moral judgements between Unfavourable and Favourable disinformation were significantly different for Conservative voters but were not for Labour voters. As this disinformation was directed towards the UK Government, which was at the time Conservative, and there were no differences in moral judgements made of virus-related disinformation, this asymmetry in moral judgements may be driven by Conservative voters making identity-protective judgements. Indeed, the Subjective Group Dynamic model suggests that pro-norm deviants (e.g. those who share disinformation benefitting the ingroup) are not judged as harshly as anti-norm deviants (Abrams et al., 2000, 2002; Hichy et al., 2008). The anonymity provided by SMPs may also increase the likelihood that a person’s focus switches from their personal to their

social identity (e.g. deindividuation) and may lead judgements to become more polarised (Spears et al., 1990). This may explain why moral judgements made by Conservative voters were done so in an identity-protective manner. However, there is also evidence to suggest that outgroup cues are enough to make ingroup identity salient (Wilder & Shapiro, 1984), and so identity alone may not explain the political asymmetries in the moral judgements here. Subsequent studies within this thesis will explore this asymmetry in more detail.

Finally, the level of privacy that actions afford to the user had some influence over interaction behaviour. "Liking" was the most favoured interaction while "sharing publicly" was the least likely. While this mimics typical behaviour on social media platforms, "sharing" is often a single response within social media focused misinformation research. However, this may be useful to explore further in relation to misinformation shared in periods of crisis. Yet it should be noted that when these actions are looked at on an individual level the levels of skewness for some DVs may create an issue for individual regression models.

There are, however, potential limitations with the present study. Firstly, "interactions" were defined as only three actions. In reality, Facebook has a number of "reactions" in addition to liking as well as the option to comment. Many of these actions may be used in a negative sense and could have complex meanings. "Anger", for instance, may be a way of expressing anger about a situation, towards an individual or in response to the existence of the content itself. Other actions such as reporting or downvoting content may also impact the total reach of the content from an algorithmic perspective. Future studies may wish to introduce additional engagement measurements to cover a broader range of interaction types. Furthermore, while the skewed responses for "interactions" may reflect normal social media behaviours it does mean the results must be taken with caution. Indeed, skewness becomes more of an issue in larger samples the further the score is from 0 (Tabachnick & Fidell, 2013). A number of the individual interaction regressions in

particular may therefore be less reliable where skewness scores are higher. Finally, it may be that the certain relationships did not reach significant because of the scale selected to represent perceived risk of COVID-19. Indeed, when a single question from the scale (regarding COVID-19 emerging as a long-term health issue) was used, the model predicting interactions with Minimising misinformation was significant and accounted for 4% of variance. This ties in with Duffy & Allington's (2020, p.15) findings that one group emerging during the pandemic ("The Frustrated") believed there to be a greater need to prioritise the economy and that "too much of a fuss" was being made about the risk of COVID-19. Therefore, the way in which beliefs were captured may have influenced their role as predictors here.

While people may be aware of the potential consequences of misinformation and disinformation, if they perceive belief-consistent disinformation as "different" in some way, current interventions may not be as effective. Not only may this influence whether users interact with misinformation themselves, it may prevent them from holding fellow users who share misinformation accountable. As Bandura suggests, "The triumph of evil requires a lot of good people doing a bit of it in a morally disengaged way with indifference to the human suffering they collectively cause" (Bandura, 1999, p206). Understanding misinformation as a moral issue may be a useful approach for understanding its spread and developing future public-facing interventions.

4.4.1. Conclusion

The present study looked at two different types of beliefs specifically relating to COVID-19 - trust in the UK Government's handling of the pandemic and perceived risk of the virus. Participants were asked to rate the likelihood of interacting with a series of misleading posts, and upon learning they were misleading, how morally acceptable they would be for others to share. The misinformation related to the aforementioned beliefs and divided into four distinct categories. Trust in the Government's handling of the pandemic

positively predicted interactions with misinformation favourable towards the Government. However, trust also negatively predicted interaction with misinformation that was unfavourable to the Government. This pattern of findings was also reflected in the moral judgements of Favourable and Unfavourable disinformation. Perceived risk of COVID-19 positively predicted interaction with misinformation that maximised the threat of the virus but did not predict moral judgements of Maximising disinformation. Finally, perceived risk negatively predicted moral judgements of disinformation that minimised the threat of the virus but did not predict interaction with Minimising misinformation.

Chapter 5. Study Two

5.1. Introduction

This chapter expands on from the exploratory, group-based findings in study one looking at moral judgements of spreading disinformation. Study two moves away from focusing on the influence of belief-consistency and focuses on the role of social identity and group norms on moral judgements of spreading misinformation and disinformation. Notably, data for this study was collected during the 2021 London Mayoral and Assembly elections, a period when political identities should be more salient to voters in that area. This chapter will begin by discussing the relationship between morality and identity further. Firstly, the role of moral intuition in moral cognition will be discussed, followed by an overview of how moral violations can threaten the moral self and a person's social identity. The use of different moral principles when making decisions in the face of competing outcomes is then explored. Finally, work on Moral Foundations Theory is discussed in relation to political asymmetries in morality. The impact of message target (e.g. ingroup or outgroup) and stance (e.g. supportive or undermining) on moral judgements of misinformation are tested using ANOVAs. Next, a series of paired *t*-tests demonstrates the effect of learning that the post is untrue (e.g. disinformation) on moral judgements. Exploratory analysis looks at moral judgements of disinformation and intentions to report disinformation to a social media platform. Finally, moral judgements of misinformation and disinformation, and intentions to report are analysed in the context of group membership (e.g. Conservative and Labour voters).

Notably, the political asymmetry observed in the previous chapter requires further exploration. As the stimuli focused solely on the UK Government, who at the time of the study were Conservative, it may be the case that the stimuli caused political identity to become salient for Conservative supporters only, leading to deindividuation effects. This would suggest that any disinformation that in some way primed elements of one's own

social identity could lead individuals to make biased moral evaluations in relation to impact on the ingroup. Additionally, it could be that Conservative and Labour supporters simply approach moral judgements of identity-related disinformation differently, in a way that may not affect the judgements of other types of disinformation. The purpose of the present chapter is therefore to establish whether these differences in moral judgements of identity-related disinformation were primarily driven by content or by the groups themselves.

5.1.1. Automatic Intuitions in Moral Cognition

Moral intuitions allow people to quickly and automatically interpret "right" and "wrong" (Haidt & Kesebir, 2010). As affective responses, moral intuitions lead to fast, but often unconscious, evaluations (Haidt & Kesebir, 2010) and play a key role in impression formation (Wojciszke et al., 1998). Given media such as imagery (Bradley & Lang, 2002) and short news articles (E. J. Johnson & Tversky, 1983) can induce integral affect, when misinformation appears within a social media newsfeed, users may be able to gauge almost immediately whether they perceive it as "moral" or '"immoral", even when they are themselves unaware it is factually inaccurate. However, unlike the potentially objective evaluations that may be made through more effortful deliberation, moral intuitions are subjective. As such, the framing of the information (including whether it appears to be inaccurate) is likely to play a role.

While moral intuitions may quickly produce a sense of "wrongness", they also have the potential to guide related decisions and judgements. Moral reasoning is thought to be distinct from moral intuition, the former being a deliberative process and the latter automatic. However, moral intuition is still believed to have influence over moral reasoning (Haidt & Kesebir, 2010) in a manner than can shape real world outcomes. For example, the level of emotional outrage people experience in response to norm violations may influence the severity of any punishment they assign (Kahneman et al., 1998). Affect

is also generally considered a key heuristic in decision making (Västfjäll et al., 2016). Social media content that in some way presents a moral violation may therefore result in strong affective responses that guide users to “feel” that the content is ”wrong” and ultimately guide any judgements relating to it.

5.1.2. Protecting the Moral Self From Threats Posed by “Disinformation Spread”

Research suggests people may care more about being perceived as a moral than being seen as competent. Notably, judgements relating to morality are also thought to occur more readily than judgements of competence (Pagliaro et al., 2016; Wojciszke et al., 1998; Ybarra et al., 2001). Arguably, the need to evaluate whether someone intends to deliberately cause a person harm is likely more urgent than evaluating whether they are competent. From an evolutionary perspective at least, this rapid prioritisation of judging moral traits such as “warmth” over “competence” is believed to be beneficial for survival (Fiske et al., 2007). As such, people are motivated to behave in ways that are considered ”moral” (both by their own (Bandura, 1991a) and other people’s standards (Pagliaro et al., 2016)) and therefore should regulate their behaviour to avoid negatively impacting the self.

Research suggests people feel they care about the accuracy of information they share on social media (Pennycook, Epstein, et al., 2021). However, while the sharing of disinformation within social media platforms is still a relatively recent phenomenon, the act of digitally sharing false information is likely to violate well-established moral values. Indeed, people begin to make judgements about dishonesty and lying from a relatively young age (Bussey, 1999; Peterson et al., 1983). Yet, from a motivated reasoning perspective, perceptions of “accuracy” may not always have a factual basis and, instead, may relate to identity-related goals (Leeper & Slothuus, 2014). Indeed, research suggests people can sometimes be willing to make accuracy judgements that prioritise protecting their identity, even in the face of conflicting evidence (Schaffner & Luks, 2018). The sharing of identity-affirming disinformation may therefore not always be seen as

"inaccurate" in the same way as other types of disinformation. While people may report caring about sharing accurate information online, this may be a way for them to protect their moral self. Indeed, "prefactual virtues" allow people to express good intentions without having to carry them out (Effron & Conway, 2015). In turn, these expressions may license people to act immorally (Cascio & Plant, 2015). By expressing the importance of accuracy in relation to certain contexts (e.g. when it "feels wrong") people feel they are permitted to share disinformation when it benefits them.

The distinctions between "factually inaccurate" and "immoral" are potentially also important in the context of misinformation. Without the knowledge or evidence to suggest the information is correct, unverified content has the potential to violate moral standards surrounding spreading only "accurate" information. If unverified information is perceived as a potential threat to the moral self then, according to social cognitive theory (Bandura, 1991a), users should self-regulate their behaviour and therefore not spread it further. However, research suggests that much harsher judgements are also assigned to actions carried out deliberately rather than accidentally (Parkinson & Byrne, 2018). If sharing disinformation is likely to result in more negative consequences than spreading misinformation, then there may be less motivation for users to be truly and objectively sceptical of the information they are presented with within social media platforms.

Moreover, people may perceive themselves to be less vulnerable to misinformation than others (Jang & Kim, 2018). As such they may simply dismiss the notion that they may be exposed to misinformation or expect to be able to easily identify it. Yet in reality, within social media platforms (SMPs) there is, more often than not, a lack of transparency around the accuracy of information. Individuals may therefore base interaction decisions on different moral standards when the accuracy of a post is unknown (e.g. misinformation) versus when they are actively aware that the information is misleading (e.g. disinformation). However, even strongly held personal or collective moral standards

relating to the sharing of potential disinformation have the potential to be cast aside in favour of conflicting social norms within group contexts.

5.1.3. Groups and Morality: Norms, Identity Threats and Hypocrisy

While the concept of morality is not itself specific to humans, certain elements are unique within human morality. This includes the motivation to be perceived as moral by others as well as the use of abstract social norms to define specific expectations about what others should do (Tomasello & Vaish, 2013). Social Identity Theory outlines that negative evaluations of an ingroup can impact the self-concept (Tajfel & Turner, 2004). As such, violations of group norms may therefore not only threaten the image of the ingroup; they may also impact the self-concept of other group members. Group-defined norms therefore help guide individuals to act in ways that will be viewed as acceptable by fellow ingroup members (Ellemers & van den Bos, 2012).

Moreover, "moral norms" are a subcategory of social norms which discourage group members from engaging in selfish behaviour (FeldmanHall et al., 2018). Whether or not an ingroup is perceived to be complying with moral norms is also thought to be more important for group evaluations than other factors such as competence and sociability (Brambilla et al., 2013; Leach et al., 2007). As such, group members may judge violators of ingroup norms negatively (Abrams et al., 2002) and even distance themselves from an individual whose immoral actions threaten the ingroup's image (Brambilla et al., 2013). However, as people are motivated to maintain or achieve a positive social identity (Tajfel & Turner, 2004) they may also take care to behave in ways that fellow group members would perceive as morally acceptable (Pagliaro et al., 2016; Van Nunspeet et al., 2014). Indeed, affective responses are thought to help alert people to identity-related threats, both in the context of the group, but also the self, such as their position within the ingroup (Ellemers et al., 2002). From this perspective, "morality" at a group-level involves both a

framework of rules and standards that group members should individually abide by, but also a lens through which individuals and groups can be evaluated.

However, while it is important that group members behave in line with norms, people may judge moral violations committed by ingroup members less harshly than those committed by outgroup members. For example, while individuals may define their own actions and those of fellow group members to be "fair", they can judge the same actions to be significantly less fair if carried out by an outgroup member (Valdesolo & Desteno, 2007). It has also been suggested that when strong identifiers perceive an ingroup as morally superior, they are more lenient towards ingroup rule breaking (A. Iyer et al., 2012). Those scoring highly in ingroup glorification may also assign more moderate punishments for moral violations carried out by fellow ingroup members than for those committed by outgroup members (Leidner et al., 2010). Therefore, shared social identity may lead people to make more lenient evaluations of moral violations.

While any flexibility in these judgements may be a way to limit damage to an ingroup's reputation (e.g. a threat to the group's value), it has also been suggested that moral violations committed by outgroup members are processed differently. Rather than a threat to the group image, moral violations committed by outgroup members may instead be perceived as a threat to safety (Brambilla et al., 2013). Moreover, outgroup perpetrated violations can also result in stronger negative emotional responses than violations committed by fellow ingroup members (Walter & Redlawsk, 2019). Therefore, immoral acts committed by outgroup members may be judged as categorically "different" from moral violations committed by ingroup members. As such, it could be the case that ingroup members who seek to harm an outgroup by spreading disinformation may be perceived as less of a threat than outgroup members committing the same act directed towards the ingroup.

5.1.4. Moral Dilemmas and the Shifting of Moral Principles

While some situations will present relatively clear moral decisions, in others the potential outcomes may conflict (e.g. moral dilemmas). For instance, spreading false information may violate a person's own moral standards, whereas a piece of factually untrue information may in some way be beneficial in the context of identity. Research looking at moral decision making suggests that different moral principles may be used to help determine the "best" outcome (at least in the context of the moral self). Two major principles emerged post-Enlightenment: consequentialism (where judgements of "right" and "wrong" are made in relation to possible outcomes) and deontology (where level of harm is judged solely on the action itself) (Haidt & Kesebir, 2010). These principles are best illustrated in the context of the well-known "trolley problem" (Thomson, 1985), which sees participants presented with a scenario where a trolley is hurtling towards five people that will ultimately kill them all. However, if participants choose to pull a lever, the trolley will divert, only killing one person on a different track. Participants taking a deontologist stance would judge pulling the lever to be immoral, whereas those taking a consequentialist approach would perceive the consequences of sacrificing one person to saving five as "a greater good". Generally, in the traditional version of this dilemma, participants tend to choose the latter (J. D. Greene et al., 2001).

However, the principles against which people determine the morality of an action can change in relation to situational factors. For instance, when participants are instead told they would need to physically push one person off a footbridge in front of a train to save the other five, they are more likely to give deontological responses (J. D. Greene et al., 2001). The re-framing of the trolley dilemma question (e.g. "saved" instead of "killed") and number of potential victims has also been shown to influence responses (Cao et al., 2017). Moreover, the ability to visualise potential harm is also thought to be an important factor in determining whether people make deontological decisions. Indeed, Amit & Greene, (2012) found people with stronger visual over verbal cognitive styles were more

likely to make deontological judgements in footbridge-style moral dilemmas. Furthermore, they also found the footbridge dilemma was more likely to induce visualisation than the trolley problem, and that this increased visualisation helped to explain why people were more likely to make deontological judgements. This suggests that people rely on visualisation of harm when they are making "remote" moral judgements. However, whether they engage in visualisation may be determined not only by individual differences but also by situational factors (such as perceived severity of the presented information). As such, the employment of specific moral principles is not stable, but rather is highly sensitive to context.

One suggested explanation is that situational cues may influence affective processes, and as such, guide underlying moral decision-making processes. Indeed, studies employing fMRI suggest that people's emotional responses differ based on their specific role within a moral dilemma (Ciaramelli et al., 2007; J. D. Greene et al., 2001; Koenigs et al., 2007). For instance, while the fatal "action" in the original trolley problem is mediated by a lever, the footbridge problem requires an individual to physically engage in the act of pushing another human to their death. Moral judgements of the latter then appear to involve greater engagement with areas of the brain associated with emotions but also lower engagement with working memory (J. D. Greene et al., 2001). In other words, contemplating causing "direct" harm to a person may induce stronger affective responses that could lead to more emotionally driven decisions, likely based upon deontological principles. Yet, when the act is mediated, the affective response may be weaker and people may be more likely to engage in deliberation (and potentially, decisions based on consequentialist principles).

Furthermore, the basis of moral decisions may also be influenced by interactions between situational and person-level factors (e.g. social identity, etc). For instance, when political liberals were presented with a version of the trolley dilemma where sacrificing one character who was assumed to be White would save a group of 100 people who were

assumed to be Black, participants made consequential decisions, choosing to intervene and sacrifice the individual (Uhlmann et al., 2009). However, when the characters' assumed race was switched, politically liberal participants made more deontological decisions. In contrast, participants who reported being politically conservative made more deontological judgements for both scenarios. A subsequent study presented participants with scenarios regarding military action and collateral damage, where politically conservative participants were found to be more likely to judge collateral damage as acceptable (e.g. consequential) when victims were Iraqi compared to when they were American. However, this time politically liberal participants were more likely to make deontological judgements across both scenarios. Having also found that priming for patriotism within the military action scenario led to more consequential decisions (regardless of political orientation), Uhlmann et al. (2009) suggests such adjustments in moral principles may be influenced by whichever perceived outcome best supports a person's salient identity. Therefore, moral decision making when presented with "moral dilemmas" may be influenced by motivations. As such, the need to achieve a positive social identity may impact the judgements and decisions people make about identity-relevant disinformation.

Finally, when taken together, the research in this area indicates there may be important psychological impacts of technology on peoples' moral decision making. Firstly, certain digital environments (such as social media platforms) may act as mediators, distancing people from the harm they could potentially inflict on others. Unless they can visualise the potential harms[2] (e.g. Amit & Greene, 2012) they may not experience the same sense of "wrongness" that may otherwise dissuade them from spreading misleading information in face-to-face contexts. As such, these environments may dampen affective processes in a way that potentially encourages judgements based on factors such as

[2] To further complicate matters, any potential "harm" caused by disinformation may also be abstract (e.g. "undermine democracy") and therefore difficult to visualise. This may also require individuals to fully appreciate the role that their "micro-contributions" play in regard to disinformation spread.

perceived severity (e.g. consequential) rather than personal beliefs and moral standards about disinformation generally. Finally, algorithmic systems that prioritise identity-relevant content may also result in people making moral evaluations of information in a motivated manner. As such, even if people feel that spreading disinformation is "wrong" they won't necessarily evaluate any misleading information they encounter within social media using deontological principles. Indeed, unless automatic processes suggest that it is "wrong" to spread, they may instead consider the likely consequences and potentially even make exceptions for spreading it if it were to support "a greater good".

5.1.5. Political Orientation and Differences in Moral Cognition

Not only do situational factors influence moral cognition; individual differences can also play an important role. One of the notable theories in this area, Moral Foundations Theory (MFT), moves beyond thinking of morality solely in terms of consequentialist versus deontologist principles (which may themselves relate to "fairness" and "harm" respectively (Haidt & Kesebir, 2010)) and instead proposes that there are at least five universal foundations which underly morality: harm, fairness, loyalty, authority and sanctity (Haidt & Graham, 2007). By incorporating concepts of morality, emotion and evolution, MFT helps to explain moral differences through the concept of foundation prioritisation (Graham et al., 2013). That is to say, a person may have a greater tendency than others to prioritise certain moral foundations, for instance, if said person more readily upholds values relating to ingroup loyalty than other people.

It has also been argued that MFT can potentially explain apparent political differences in moral judgements and decision making. Indeed, previous research has proposed that those who are more politically liberal may be more likely to prioritise "harm" and "fairness" foundations in their judgements, while conservatives endorse all five equally (Graham et al., 2009). It has also been suggested that conservatives may give greater priority to "authority", "loyalty" and "sanctity" (Voelkel & Brandt, 2019), although

a recent meta-analysis suggests this may be more applicable to social rather than economic conservatives (Kivikangas et al., 2021). Notably, Haidt & Kesebir (2010) suggest that such differences in foundation prioritisation between ideological groups may reflect two different (but at times overlapping) attempts at building moral systems, with political liberals endorsing more secular, harm / fairness-based societies in contrast to the tighter knit communities that political conservatives often prefer. As such, two individuals at opposite ends of the political spectrum may look at the same situation and make vastly different moral judgements, which they both ultimately perceive to be "correct".

The readiness with which people engage with a specific foundation may then help explain apparent political differences in moral evaluations. Specifically, some people may be more sensitive than others to specific foundation-related moral violations and be more willing to uphold certain foundation-related moral values. However, it is also not the case that other people will never engage with these values. Indeed, situational factors may influence which foundations are prioritised. Research suggests when social identity is under threat people may be more likely to make moral judgements in relation to group loyalty and authority rather than harm and fairness (Leidner & Castano, 2012). In judging a moral violation carried out by a fellow member in terms of "loyalty" instead of "harm", individuals may help to limit damage to their self-concept. Therefore, arguably apparent differences in moral evaluations may be situational.

5.1.6. The Present Study

Study one produced a number of questions regarding the way in which people make moral judgements of identity-relevant disinformation. Specifically, moral judgements of government related disinformation differed for Conservative and Labour party supporters, but that was not the case for virus related disinformation. While Labour supporters judged both Favourable and Unfavourable disinformation similarly, Conservative supporters were significantly more likely to perceive the sharing of

Unfavourable disinformation as less acceptable than Favourable disinformation. Since the stimuli focused exclusively on the Conservative party, differences related to political orientation in moral evaluations of false information cannot be assumed. Instead, deindividuation effects may have led Conservative supporters alone to make more polarised judgements in line with their social identity (Spears et al., 1990). To better understand how people make moral evaluations of identity-relevant misinformation on social media, the present study therefore employs an experimental design, showing participants (Conservative and Labour supporters) a single item of misinformation that either supports or undermines their ingroup or a relevant outgroup.

Given the precedent of disinformation campaigns targeting elections, the 2021 London Mayoral and Assembly elections provided a unique opportunity to collect data. Indeed, Self-Categorisation Theory suggests identity-relevant situations may lead categorised social identities to become more salient (Turner & Reynolds, 2012). As such, people's political identities should be more readily available close to a heavily publicised election. While voter turnout may not be as high as for a general election, there is substantial media coverage (including TV debates) in the run up to the London Mayoral elections and traditionally voter turnout is higher than for local elections held on the same date (London Elects, 2021; The Electoral Commission, 2019). Therefore, the stimuli in this study specifically referenced the London elections and all participants were required to be current London residents. Data collection for the pilot study was collected a week prior to the election, with the main study data collected the day before the election itself.

People are motivated to achieve or maintain a positive identity (Tajfel & Turner, 2004), however evaluations of any particular ingroup are not stable and are instead influenced by social context (Spears et al., 1997). As such, people may have different perceptual, affective and behavioural responses to information that affirms their identity compared with information that in some way threatens it (see Ellemers et al., 2002). For instance, the moral foundations underlying their judgement may shift (Leidner & Castano,

2012). As such, when people view false information about their ingroup, their moral evaluations regarding spreading it further may not be focused on whether it is potentially inaccurate or generally harmful. Instead, evaluations may be based on whether it is specifically harmful or beneficial to said ingroup. It was therefore predicted that participants would make moral judgements about the sharing of misinformation that named their ingroup, based on whether it would affirm or threaten its reputation:

> **H1.** Individuals will perceive the sharing of misinformation that supports the ingroup as more morally acceptable than misinformation that undermines the ingroup.

Positive distinctiveness from other groups is also an important goal for achieving a positive identity (Tajfel & Turner, 2004). As an election may present a threat to status, people (especially strong identifiers) may be motivated to emphasise intergroup differences in an attempt to mitigate any such threat (Ellemers et al., 2002). Some misinformation may help people achieve this goal; however other misinformation may also further threaten the value of the ingroup. As such, rather than making moral judgements of misinformation based on the valence of the message (e.g. support or undermine), instead they may be based on the potential consequences for social identity. It was therefore predicted that participants would make moral judgements of sharing misinformation that were favourable towards the ingroup:

> **H2.** Individuals will report the sharing of disinformation that supports their ingroup as more morally acceptable than disinformation that supports the outgroup.
>
> **H3.** Individuals will report the sharing of disinformation that undermines the outgroup as more morally acceptable to share than disinformation that undermines the ingroup.

Finally, while people do appear to care about the accuracy of the information they share online, research suggests they may not actually consider accuracy prior to sharing (Pennycook, Epstein, et al., 2021). As underlying intent plays an important role in how actions and behaviours are perceived from a moral standpoint (Parkinson & Byrne, 2018), whether or not they know information to be untrue is likely to be important. As such, people may judge sharing unverified misinformation against different standards than when knowingly sharing false information (e.g. disinformation). It was therefore expected participants would report the sharing of disinformation to be less morally acceptable than prior to learning it is untrue:

> **H4.** After learning the content is untrue, individuals will report the sharing of the content as more unacceptable than before.

5.2. Method

5.2.1. Development of Stimuli

The stimuli in the present study were developed with the findings and limitations of study one in mind. For example, study one used misinformation that was either favourable or unfavourable towards the UK government. As such it featured only one political party (the Conservative party, who were the majority party at the time of data collection). Furthermore, each item was a unique piece of content found on social media, each with its own message, tone and design. As a result, it was not possible to conclude from the data whether any group differences between participants (who either reported they would vote Conservative or Labour) may indeed exist or whether the choice of stimuli played a key role.

Therefore, for study two, a misleading post was developed that could be adapted in relation to the independent variables in each condition, but otherwise controlled. A common format of misinformation circulating on social media are “screengrabs” of Twitter posts (e.g. “Tweets”) that have in fact been fabricated. Such images are straightforward to create, as many websites quickly and convincingly produce fake tweets for free. These images can then not only be circulated within Twitter but are commonly found on other platforms (e.g. Instagram, Facebook). This therefore presented a realistic format for delivering the experimental manipulation. A small pilot study was carried out to test the effectiveness of the stimuli.

5.2.1.1 Pilot Study

5.2.1.1.1 Materials. Four versions of the stimuli, each featuring a slight adjustment from each other, were tested in the pilot study. For the purpose of this study, the website TweetGen.com was used to create the stimuli and the persona “Simon Evans” was created (Figure 5.1). The profile image used was sourced from the AI image generator “This Person Does Not Exist” and at the time of the study the chosen handle was not claimed on Twitter.

Figure 5.1

Misinformation Stimuli for Study Two by Political Party

The use of attention grabbing but misleading statistics provided an opportunity to deliver the experimental manipulation, while also replicating real-world misinformation. Misinformation about violent crime is often featured on fact-checking websites, due in part to the complexities in reporting crime statistics. Despite conflicting reports, within the UK there has been a long term decrease in violent crime since 1995 (Office for National Statistics, 2021). However, the influence and number of media reports on the subject may have heightened public concern about knife crime and gangs (The Mayor's Office for Policing and Crime, 2019). In reality, changes to borough level violent crime rates vary, with rises and decreases across London over time, adding to the plausibility of this stimuli.

Notably, while councils do not directly manage the police, some local authorities who have cut youth service budgets have seen heightened levels of local knife crime (N.

Smith, 2020). Violent crime rates in London boroughs for 2019 ranged between −6% to +11% in Labour run boroughs and −4% to +14% in Conservative boroughs (Metropolitan Police, 2020). Therefore, while the 20% rates proposed in the stimuli were not factually accurate, violent crime rates do fluctuate either direction, suggesting the stimuli claims could be credible (despite being untrue).

5.2.1.1.2 Participants. 20 participants (2 males) aged 19-59 (M = 37.05, SD = 12.47) were recruited through Prolific to take part in the pilot study. For consistency, the same eligibility requirements were used as for the main study. Participants were required to have an active Facebook account and currently be residing in London. They also had to identify as either a Conservative (N = 10) or Labour supporter (N = 10) and have voted for said party in the 2019 General Elections. Three participants were removed from analysis due to a lack of variance in their responses or for not being social media users. Ethical approval for the pilot and main study was obtained from the University's Psychology Ethics Committee (ETH2021-1792, Appendix G).

5.2.1.1.3 Procedure. The study was hosted online using the survey platform Qualtrics. Participants answered a set of basic demographic questions of gender, age, and location. They were then presented with each of the four images in a random order. Participants were asked to rate how favourable the images were for the named party across a 7-point scale, from "Very unfavourable" to "Very favourable". They were then asked which UK political party they most identify with, before being thanked and debriefed.

5.2.1.1.4 Results. Mean favourability scores for the items are displayed in Table 5.1. Scores below 4 indicate content that was rated as unfavourable while scores over 4, favourable.

Table 5.1

Mean Favourability Ratings of Misinformation Stimuli

Item	N	Minimum	Maximum	Mean	SD
Conservative - Support	17	1	7	5.71	1.61
Conservative - Undermine	17	1	7	2.12	1.80
Labour - Support	17	2	7	5.82	1.38
Labour - Undermine	17	1	6	1.76	1.44

Paired sample *t*-tests showed that favourability scores of "supportive" and "undermining" stimuli were significantly different with large effect sizes for both Conservative ($t(16) = 4.49$, $p < .001$, $d = 1.07$) and Labour ($t(16) = 7.02$, $p < .001$, $d = 1.7$) targeted stimuli. These findings suggest that both within-party ratings are distinctly different in terms of favourability.

5.2.1.1.5 Discussion. Participants judged items claiming violent crime rates had risen as significantly less favourable than those suggesting that they had fallen. This was the case regardless of whether the target was the Conservative or Labour party. Importantly, not only do these findings suggest that the items are different, indeed they imply that participants are able to adjust judgements based on this small change in the content wording.

5.2.2. Main Study

5.2.2.1 Materials and Procedure

Participants were recruited via Prolific for the study which was hosted on Qualtrics. They were first presented with the invitation letter and consent form, followed by a series of basic demographic questions. Participants were then randomly assigned to one of four conditions, where they were presented with Conservative or Labour focused stimuli that either supported or undermined the named party. The image presented to individual participants stayed the same throughout the study.

Participants were first asked how morally acceptable it would be for them to share the image on their social media account, without being made aware that the content was fake (e.g. “misinformation”). Moral acceptability was measured on an 11-point scale, where a score of “0” indicated participants felt sharing was not at all morally acceptable whereas “10” would be completely morally acceptable. Participants were then informed that an independent fact-checker had flagged the post as problematic as it contained false information (e.g. “disinformation”). They were again asked how morally acceptable it would be to share the image using the same 11-point scale. Next, they were asked, now knowing that the post contained false information, whether they would flag or report it if they saw it on their social media newsfeed. An 11-point scale indicated the likelihood of reporting (0 – not at all likely, 10 – extremely likely). Finally, participants were asked which political party they most identified with (e.g. Conservative, Labour, etc). Participants were then thanked and debriefed.

5.2.2.2 Participants

246 participants (120 males) aged 18-71 (M = 35.40, SD = 12.09) were recruited through Prolific to take part in the study. Ethical approval was obtained from the University’s Psychology Ethics Committee (ETH2021-1792, Appendix G). Sample size was determined through a power analysis using G*Power, which indicated that 191 participants were needed to detect $\eta_p^2 = .04$ with 80% power.

For this study, participants were required to have an active social media account (e.g. Facebook, Instagram, Twitter, etc) and must not have taken part in the pilot or Study one. As with the pilot, participants had to identify as either Conservative (N = 121) or Labour (N = 125) supporters and have voted for said party in the 2019 general election. Participants also were required to be living in London at the time of the study. The data collection for the present study took place the day before the May elections in 2021.

A total of 29 participants were removed before analysis for not meeting the recruitment criteria. Of this, 11 participants were not based in London at the time of the study, while 19 did not identify as supporters of the party they had registered on Prolific as identifying with or had previously voted for. As it was not possible to know whether these individuals had changed their political allegiance or incorrectly entered their response, the decision was taken to remove them from analysis. Additionally, 11 participants were removed during the data cleaning process for inauthentic responses judged against a set of criteria. Firstly, certain combinations of moral and reporting scores were deemed implausible. Eight participants were removed for assigning high scores (e.g. 7 and above) for both moral judgements of known disinformation and likelihood of flagging, as this combination implies an inauthentic response. In reality, those who judge sharing a piece of disinformation as acceptable are not then likely to report the post for removal. Another two were removed for a lack of variance in their responses above the scale mid-point (e.g. 5). While a sequence of three low scores (e.g. perceiving sharing to be immoral but not reporting it) is plausible, judging the sharing of the material to be more moral than immoral while also intending to report it suggests an inauthentic response. Finally, one participant was removed for extreme increase between pre and post moral scores (e.g. “1” then “11”) as again this combination suggests an inauthentic response. Notably, the removal of these participants did not affect planned tests in relation to reaching significance. However, in exploratory analysis, removal did lead to the main effect of party to become significant in relation to reporting likelihood. This change is expected, given that eight of the participants removed from analysis for inauthentic responses were identified specifically by their high “reporting” likelihood score accompanying positive moral judgements of sharing known disinformation. As reporting is an action more commonly associated with behaviour that is unethical, removal of these inauthentic responses may therefore be justified for this test. However, a corresponding full set of results including these excluded participants are included in Appendix H.

Participant demographics for the final sample are shown in Table 5.2.

Table 5.2

Participant Demographics for Study Two

	All		Conservative		Labour	
	N	%	N	%	N	%
Total	206	100	99	48.1	107	51.9
Age	35.32		37.92		32.92	
Gender						
Female	108	52.4	46	46.5	62	57.9
Male	97	47.1	53	53.5	44	41.1
Prefer not to say	1	0.5	0	0.0	1	0.9
Education						
GCSEs	13	6.3	6	6.1	7	0.9
A-Levels	34	16.5	16	16.2	18	6.5
Bachelor's	111	53.9	53	53.5	58	16.8
Master's Degree	40	19.4	19	19.2	21	54.2
Doctoral Degree	6	2.9	4	4.0	2	19.6
Other	2	1.0	1	1.0	1	1.9
Social Media						
Facebook	161	78.2	85	85.9	76	71.0
Instagram	161	78.2	75	75.8	86	80.4
Twitter	111	53.9	49	49.5	62	57.9
LinkedIn	112	54.4	60	60.6	52	48.6
Pinterest	39	18.9	21	21.2	18	16.8
YouTube	164	79.6	74	74.7	90	84.1
TikTok	50	24.3	22	22.2	28	26.2
Reddit	67	32.5	24	24.2	43	40.2

5.2.2.3 Data Analysis

The present study used 2x2 factorial analysis of variance (ANOVA) to test H1-H3. "target" ("ingroup" vs "outgroup") and "stance" ("supportive" vs "undermining") were used as between-group factors, with the moral judgements score prior to the stimuli being disclosed as disinformation as the dependent variable. H4 was tested using paired *t*-tests

comparing moral judgement scores before and after learning the stimuli was misleading. The *t*-tests were confirmed using Wilcoxon Signed Rank Tests. All noted tests used α levels of .05.

5.3. Results

Data were first exported from Qualtrics into Excel for data cleaning and then imported to SPSS. As noted in the methods section, a total of 41 participants were removed during the data cleaning process, primarily due to not meeting the recruitment criteria. Participants were each assigned a code based on their relationship to the condition which they were assigned to (e.g. ingroup supporting, outgroup supporting, etc.). This provided the basis for the variables "target" (e.g. ingroup or outgroup) and "stance" (e.g. "supportive" or "undermining").

Descriptive statistics for all variables are shown in Table 5.3, with histograms and QQ plots provided in Appendix I. Any violations of assumptions are discussed throughout the results where relevant.

Table 5.3

Summary of Descriptive Statistics

				Range			
	N	*M*	*SD*	Potential	Actual	Skewness	Kurtosis
Age	206	35.32	12.46		18-71	1.01	0.47
Moral Judgements of Misinformation							
All	206	6.37	3.14	1-11	1-11	−0.20	−1.02
Ingroup - Support	52	7.75	2.57	1-11	1-11	−0.57	−0.24
Outgroup - Support	53	5.55	2.91	1-11	1-11	−0.07	−0.93
Ingroup - Undermine	50	5.32	3.15	1-11	1-11	0.23	−1.11
Outgroup - Undermine	51	6.84	3.34	1-11	1-11	−0.38	−0.97
Moral Judgements of Disinformation							
All	206	2.53	2.27	1-11	1-11	1.72	2.63
Ingroup - Support	52	2.96	2.50	1-11	1-11	1.25	1.05

				Range			
	N	*M*	*SD*	Potential	Actual	Skewness	Kurtosis
Outgroup - Support	53	2.32	2.06	1-11	1-10	1.88	3.55
Ingroup - Undermine	50	1.88	1.48	1-11	1-8	2.22	5.75
Outgroup - Undermine	51	2.96	2.72	1-11	1-11	1.50	1.57
Likelihood of Reporting Disinformation							
All	205	5.12	3.63	1-11	1-11	0.36	−1.30
Ingroup - Support	51	5.10	3.69	1-11	1-11	0.36	−1.36
Outgroup - Support	53	5.36	3.80	1-11	1-11	0.28	−1.42
Ingroup - Undermine	50	5.74	3.85	1-11	1-11	0.05	−1.53
Outgroup - Undermine	51	4.27	3.04	1-11	1-11	0.73	−0.53

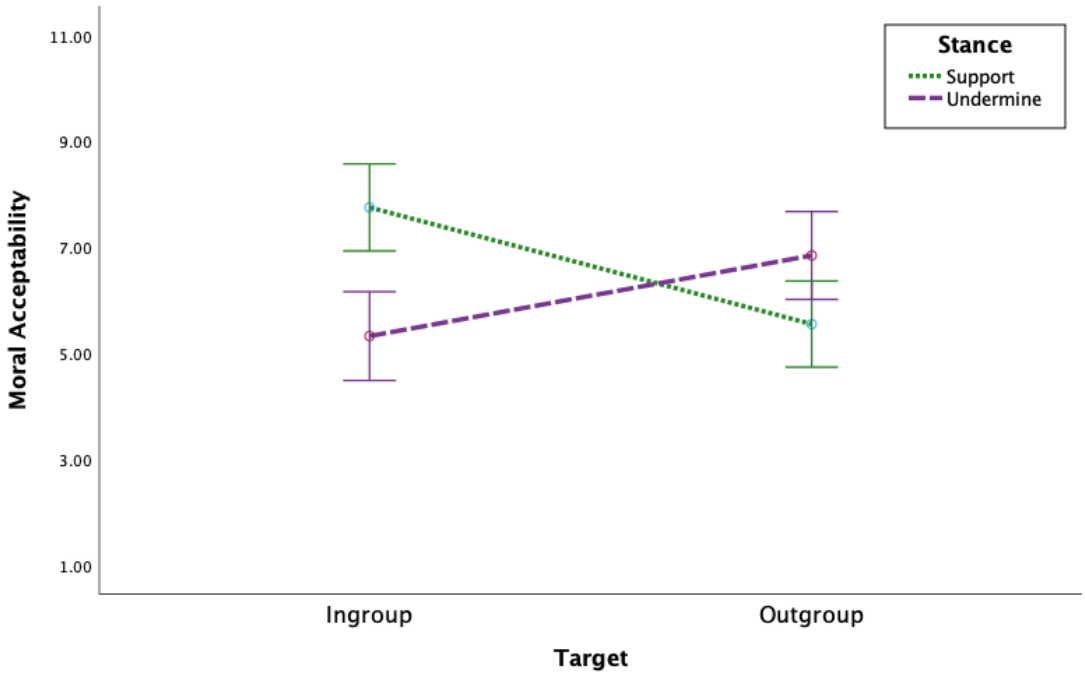
5.3.1. Planned Tests

5.3.1.1 Groups and Moral Judgements of Misinformation

To test the first three hypotheses, the moral judgements of misinformation (e.g. not disclosed as untrue) were analysed using a 2x2 factorial analysis of variance (ANOVA), with "target" ("ingroup" vs "outgroup") and "stance" ("supportive" vs "undermining") as between-group factors. There was homogeneity of variance as assessed by Levene's test (p = .19) and no outliers as assessed by inspection of boxplot (Appendix J). Visual inspection of the histograms revealed that the data were not normally distributed (Appendix J), however ANOVAs are thought to be robust to violations of this assumption (Tabachnick & Fidell, 2013). The ANOVA revealed the interaction between "target" and "stance" was significant with a medium effect size, $F(1, 202) = 19.78$, $p < .001$, $\eta_p^2 = .09$. This is illustrated in Figure 5.2.

Figure 5.2

Estimated Marginal Means of Moral Judgements of Sharing Misinformation



Note: Error bars 95% CI

The interaction was explored further through analysis of simple main effects. All reported *p*-values are Bonferroni-adjusted. Firstly, misinformation that supported the ingroup was judged as significantly more acceptable to share than when it undermined the ingroup with a medium effect size, $F(1, 202) = 16.66$, $p < .001$, $\eta_p^2 = .08$. H1 is therefore accepted. Judgements of misinformation that either supported or undermined the outgroup were also significantly different but with a small effect size, $F(1, 202) = 4.83$, $p = .03$, $\eta_p^2 = .02$. This suggests that moral judgements of sharing misinformation are not simply based on message valence (e.g. whether generally framed positively or negatively).

Indeed, misinformation that supported the ingroup was judged as significantly more acceptable to share than misinformation supporting the outgroup with a medium effect size, $F(1, 202) = 14.10$, $p < .001$, $\eta_p^2 = .07$. Conversely, misinformation undermining the ingroup was judged as significantly less acceptable to share than misinformation undermining the outgroup with a small effect size, $F(1, 202) = 6.49$, $p = .01$, $\eta_p^2 = .03$.

These findings illustrate how moral judgements of misinformation can change in relation to identity in a manner that favours the ingroup. H2 and H3 are therefore accepted.

5.3.1.2 Differences Between Moral Judgements of Misinformation and Disinformation

To test the fourth hypothesis, a series of paired *t*-tests were run to understand whether participants updated their moral judgements upon learning that the content was misleading. Inspection of the histograms suggested that distributions for the differences between moral judgements of known and unknown disinformation were not normally distributed (Appendix K). The decision was taken to continue as Q-Q plots suggested the data were somewhat normally distributed and, in samples this size, paired *t*-tests are thought to be fairly robust to violations of normality (Pek et al., 2018). However, non-parametric tests were also used to confirm the results.

Overall, participants judged disinformation to be significantly less acceptable to spread after learning it was false, $t(205) = -17.21$, $p < .001$, $d = 1.20$. This was confirmed by the Wilcoxon Signed Rank Test, $z = -10.88$, $n = 206$, $p < .001$, $r = .76$. Paired *t*-tests (Table 5.4) and Wilcoxon Signed Rank Tests (Appendix L) for each item were also significant with large effect sizes (Figure 5.3). H4 is therefore also accepted.

Table 5.4

Differences Between Moral Judgements of Misinformation and Disinformation

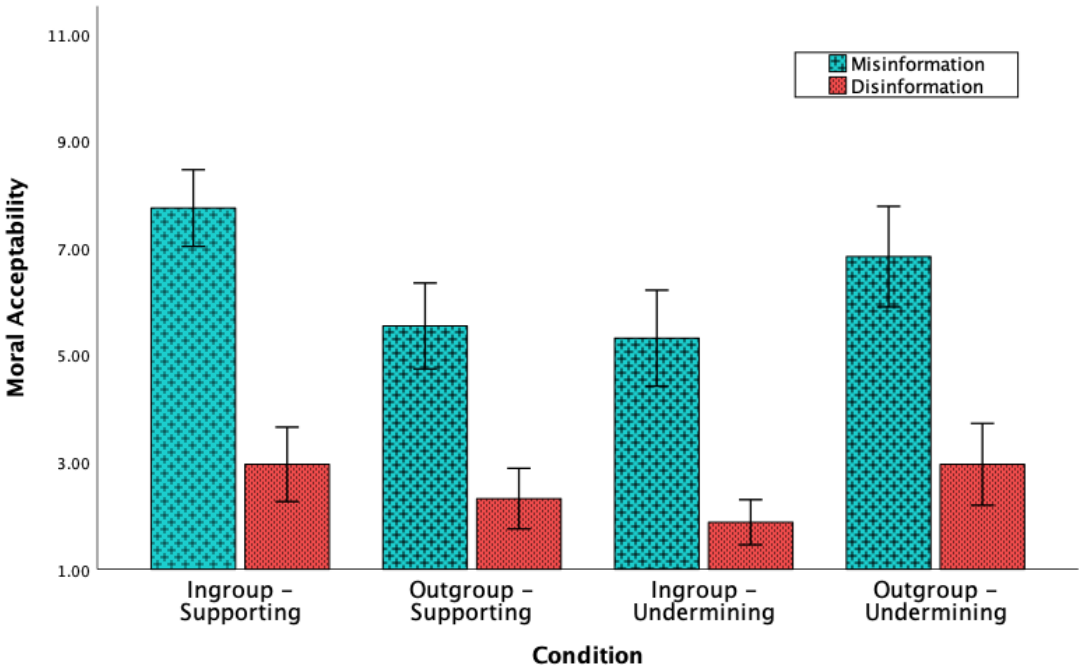
		Misinformation (Unknown)		Disinformation (Known)			
	N	*M*	*SD*	*M*	*SD*	*t*	*d*
Ingroup Supporting	52	7.75	2.57	2.96	2.50	−10.96***	1.52
Ingroup Undermining	50	5.32	3.15	1.88	1.48	−7.61***	1.08
Outgroup Supporting	53	5.55	2.91	2.32	2.06	−7.74***	1.06
Outgroup Undermining	51	6.84	3.34	2.96	2.72	−8.47***	1.19

Note. M = Mean, SD = Standard Deviation

*** $p < .001$.

Figure 5.3

Mean Moral Acceptability Scores for Sharing Misinformation and Disinformation



Note: Means with 95% CI displayed.

5.3.2. Further Exploratory Analyses

5.3.2.1 Moral Judgements of Known Disinformation

To establish whether ingroup bias still occurs upon learning that the post is misleading, another 2x2 factorial ANOVA was run. As before, the ANOVA employed "target" ("ingroup" vs "outgroup") and "stance" ("supportive" vs "undermining") as between-group factors, and moral judgement scores of known disinformation were entered as the dependent variable. Levene's test was significant ($p < .001$) and therefore a significance level of .01 will be applied for interpreting the results. Visual analysis of the boxplots revealed eight outliers; however their removal does not impact the results[3]. Visual

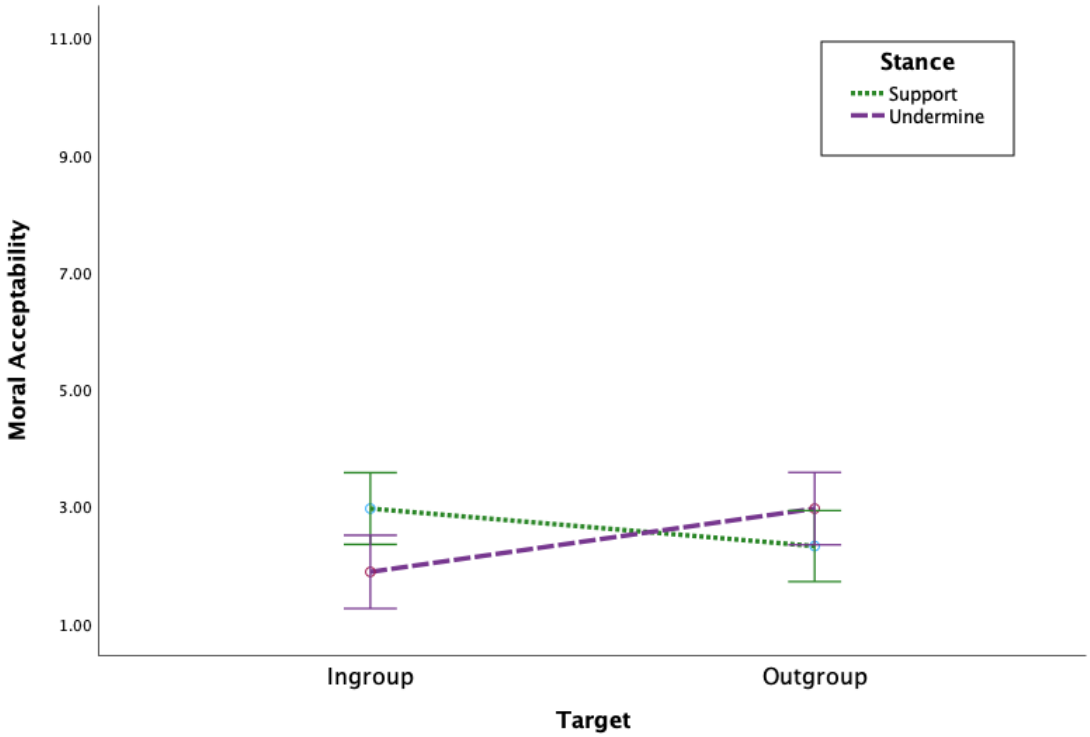
[3] Upon the removal of outliers the interaction effect was still significant, $F(1, 194) = 9.93$, $p < .01$, $\eta_p^2 = .05$. Again, neither of the main effects of "target" ($F(1, 194) = 0.1$, $p = .76$, $\eta_p^2 = .002$) or "stance" ($F(1, 194) = 0.25$, $p = .61$, $\eta_p^2 = .002$) were significant.

analyses of the histograms also showed that the data were not normally distributed (Appendix M), with descriptive statistics demonstrating the presence of both skewness and kurtosis. Floor effects are also present, with a large proportion of participants reporting that sharing would be "not at all" acceptable. This could reflect a legitimate reaction to the sharing of disinformation, but nonetheless the results should be taken with caution.

There was a small but significant interaction between "target" and "stance", $F(1, 202) = 7.59$, $p < .01$, $\eta_p^2 = .04$. Analysis of simple main effects suggests that known disinformation which targeted the ingroup was judged as significantly different across stance (support vs. undermine) with medium effect sizes, $F(1, 202) = 5.94$, $p = .02$, $\eta_p^2 = .03$. For judgements of undermining disinformation, there was also a small but significant difference between judgements of items targeting the ingroup or the outgroup, $F(1, 202) = 5.87$, $p = .02$, $\eta_p^2 = .03$. No other differences were significant (Figure 5.4).

Figure 5.4

Estimated Marginal Means of Moral Judgements of Sharing Disinformation



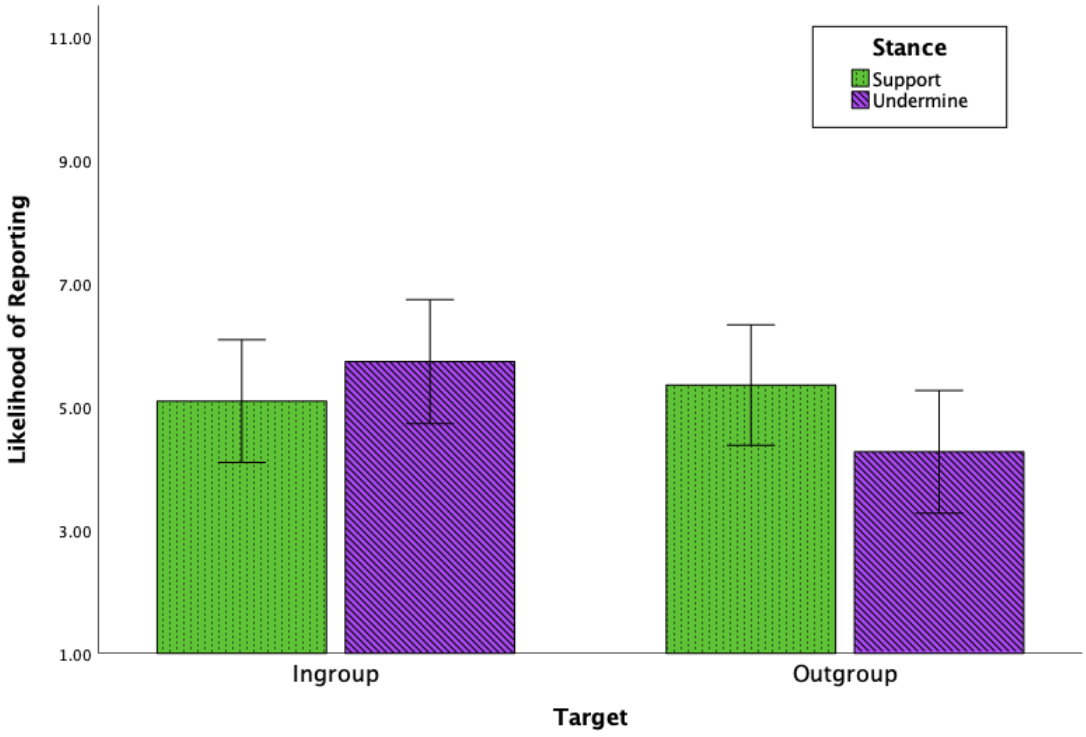
Note: Error bars 95% CI

5.3.2.2 Likelihood of Reporting Disinformation

To understand whether people may make identity-based decisions about reporting content they know to be false to social media platforms, the likelihood of reporting score was entered into the ANOVA. Levene's test was significant ($p < .05$) and so again a significance level of .01 is applied. There were no outliers as assessed by inspection of boxplot but histograms revealed the data were not normally distributed (Appendix N). As illustrated in Figure 5.5, the interaction effect between "target" and "stance" was not significant, $F(1, 201) = 2.92$, $p = .09$, $\eta_p^2 = .01$. Furthermore, neither the main effects of "target" ($F(1, 201) = 1.43$, $p = .23$, $\eta_p^2 = .01$) or "stance" ($F(1, 201) = 0.19$, $p = .66$, $\eta_p^2 = .001$) were significant.

Figure 5.5

Estimated Marginal Means of Likelihood of Reporting Disinformation



Note: Error bars 95% CI

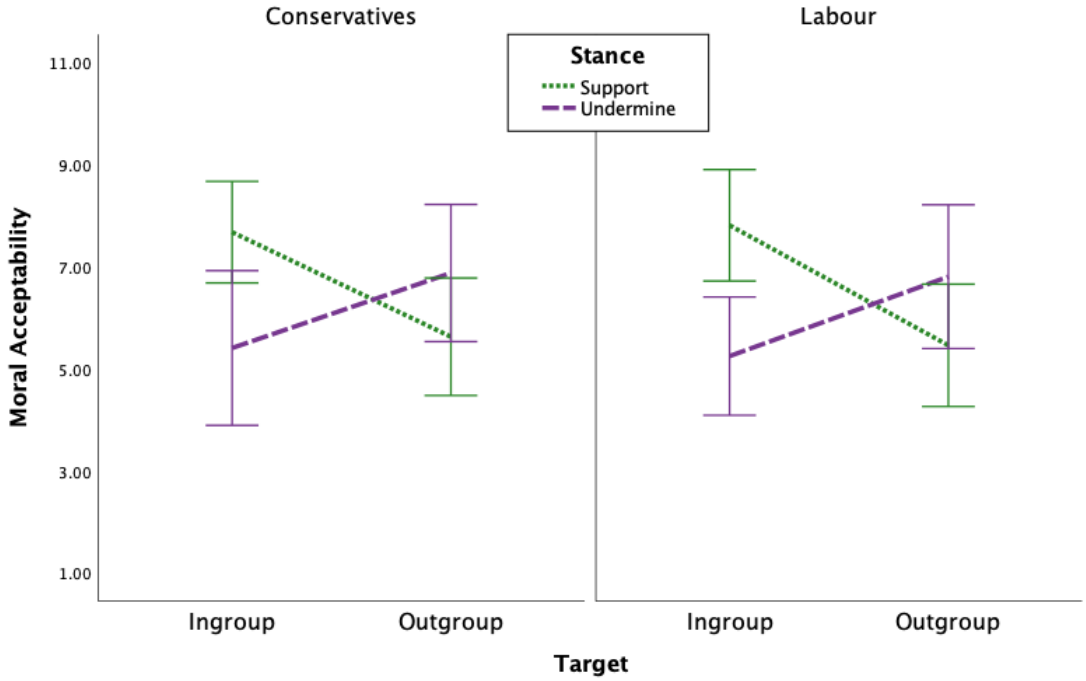
5.3.2.3 Political Differences in Moral Judgements and Reporting Likelihood

5.3.2.3.1 Moral Judgements of Misinformation by Political Party. Expanding on the findings from Study one, a series of ANCOVAs were run to understand whether there were any differences between responses of Conservative and Labour voters. These were 2x2x2 factorial ANCOVAs using "target" ("ingroup" vs "outgroup"), "stance" ("supportive" vs "undermining") and "party" ("Conservative" vs "Labour") as between-group factors. Both age and gender were added as covariates to control for any differences between the two political groups. Inspection of the histograms again showed the data were not normally distributed (Appendix O).

The ANCOVA for moral acceptability ratings of misinformation showed no significant three-way interaction between target, stance and party, $F(1, 195) = 0.74$, $p = .39$, $\eta_p^2 = .004$. However, the two-way interaction between "target" and "stance" was significant ($F(1, 195) = 21.41$, $p < .001$, $\eta_p^2 = .10$). As illustrated in Figure 5.6, the simple two-way interactions were significant for both Conservative voters, $F(1, 95) = 8.43$, $p < .01$, $\eta_p^2 = .08$, and Labour voters, $F(1, 103) = 10.92$, $p = .001$, $\eta_p^2 = .10$. The full ANCOVA results can be found in Table 5.5.

Figure 5.6

Mean Misinformation Moral Judgement Scores Displayed by Political Party



Note: Error bars 95% CI

Table 5.5

Three-Way ANCOVA Statistics for Moral Acceptability of Sharing Misinformation

	$\bar{x}^2$	F (1, 195)	η_p^2
Age	67.81	7.72**	.04
Gender	40.33	4.59*	.02
Stance	13.65	1.55	.01
Target	6.71	0.76	.00
Party	1.96	0.20	.00
Stance x Target	188.18	21.41***	.10
Stance x Party	0.01	0.001	.00
Target x Party	0.22	0.02	.00
Stance x Target x Party	6.54	0.74	.00

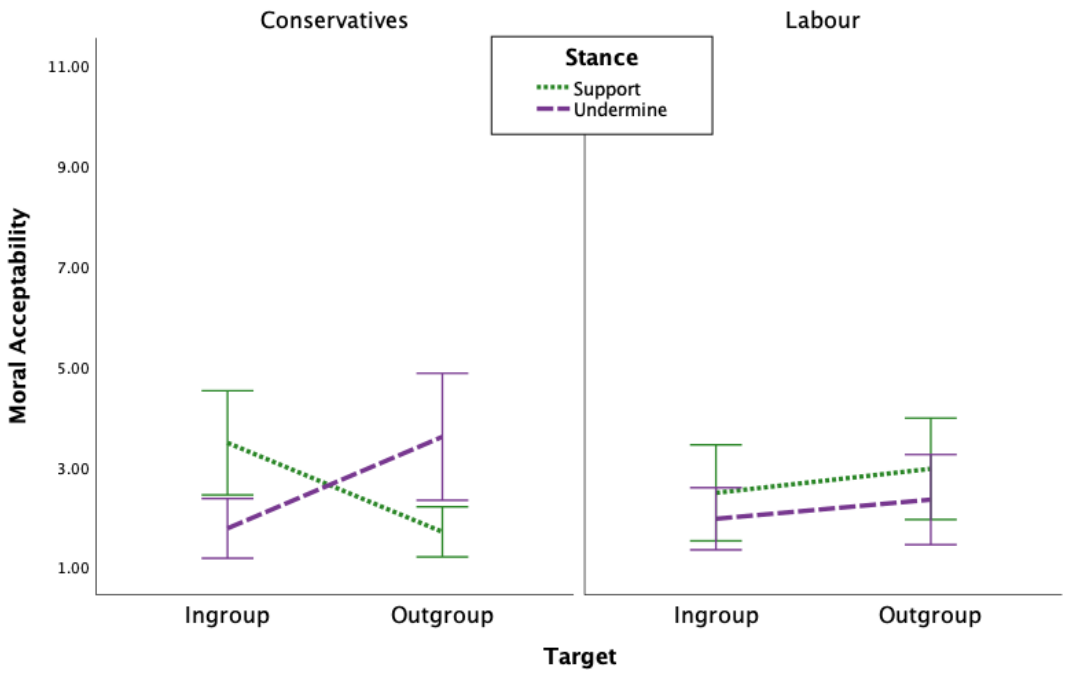
Note. Gender coded as dummy variable, F = 0, M = 1.

*p < .05. ** p < .01. *** p < .001.

5.3.2.3.2 Moral Judgements of Disinformation by Political Party. Next, moral judgements of disinformation (e.g. after learning the information was untrue) were entered into the ANCOVA. Again, inspection of the histograms suggests the data were not normally distributed (Appendix P). The ANCOVA showed a significant three-way interaction between target, stance and party, $F(1, 195) = 7.66$, $p < .01$, $\eta_p^2 = .04$. As illustrated in Figure 5.7, the simple two-way interaction between "target" and "stance" was significant for Conservative voters ($F(1, 95) = 16.59$, $p < .001$, $\eta_p^2 = .15$) but not Labour voters ($F(1, 103) = 0.01$, $p = .91$, $\eta_p^2 = .00$). The full ANCOVA results can be found in Table 5.6.

Figure 5.7

Mean Disinformation Moral Judgement Scores Displayed by Political Party



Note: Error bars 95% CI

Table 5.6

Three-Way ANCOVA Statistics for Moral Acceptability of Sharing Disinformation

	$\bar{x}^2$	F (1, 195)	η_p^2
Age	17.74	3.70	.02
Gender	0.15	0.03	.00
Stance	2.88	0.60	.00
Target	2.20	0.46	.00
Party	5.58	1.16	.00
Stance x Target	40.12	8.37**	.04
Stance x Party	6.64	1.39	.01
Target x Party	1.11	0.23	.00
Stance x Target x Party	36.74	7.66**	.04

Note. Gender coded as dummy variable, F = 0, M = 1.

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

Analysis of simple main effects for Conservative voters was carried out with Bonferroni-adjusted *p*-values. These suggest that, after learning the content was misleading, Conservative voters judged disinformation supportive of the Labour party to be significantly less morally acceptable to share than disinformation supportive of their own party, $F(1, 95) = 8.50$, $p < .01$, $\eta_p^2 = .08$. Conservative voters were also more accepting of sharing disinformation that undermined the Labour party than disinformation that undermined their own party, $F(1, 95) = 8.11$, $p < .01$, $\eta_p^2 = .08$. Conservative voters also judged sharing disinformation that may undermine the Conservative party as significantly less acceptable than sharing disinformation that may support the party, $F(1, 95) = 7.08$, $p < .01$, $\eta_p^2 = .07$. However, when disinformation named the Labour party, Conservative voters felt that sharing disinformation that undermined Labour was significantly more acceptable than sharing disinformation that supported them, $F(1, 95) = 9.69$, $p < .01$, $\eta_p^2 = .09$. These findings, which may reflect an ingroup bias for their own party versus their main political opponent, all have medium effect sizes.

5.3.2.3.3 Likelihood of Reporting by Political Party. Finally, the likelihood of reporting known disinformation was entered into the ANCOVA. Inspection of the histograms suggests the data were not normally distributed (Appendix Q) and therefore the findings should be taken with caution. There was no significant three-way interaction between target, stance and party, $F(1, 194) = 2.39$, $p = .12$, $\eta_p^2 = .01$. However, the main effect of party was significant, $F(1, 194) = 5.16$, $p < .05$ $\eta_p^2 = .03$. The simple two-way interaction between "target" and "stance" was significant for Labour voters ($F(1, 102) = 4.78$, $p = .03$, $\eta_p^2 = .05$) but not Conservative voters ($F(1, 95) = 0.02$, $p = .90$, $\eta_p^2 = .00$). The full ANCOVA results can be found in Table 5.7.

Table 5.7

Three-Way ANCOVA Statistics for Likelihood of Reporting Disinformation

	$\bar{x}^2$	F (1, 195)	η_p^2
Age	7.44	0.57	.02
Gender	0.11	0.01	.00
Stance	5.23	0.40	.00
Target	15.21	1.17	.01
Party	66.93	5.16*	.03
Stance x Target	36.80	2.83	.01
Stance x Party	0.08	0.01	.00
Target x Party	2.67	0.21	.00
Stance x Target x Party	31.02	2.40	.01

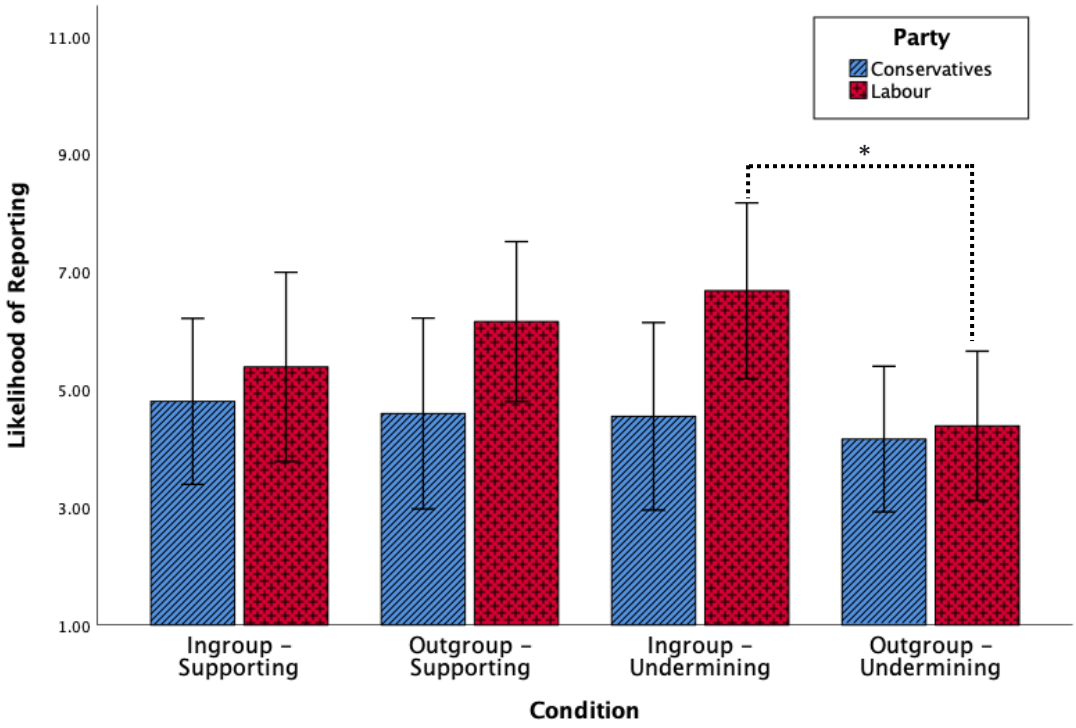
Note. Gender coded as dummy variable, F = 0, M = 1.

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

Analysis of simple main effects for Labour voters was carried out with Bonferroni-adjusted *p*-values. As illustrated in Figure 5.8, Labour voters were significantly more likely to report disinformation that undermined the Labour party compared to disinformation that undermined the Conservative party, $F(1, 102) = 5.46$, $p < .05$, $\eta_p^2 = .05$. No other differences were significant.

Figure 5.8

Mean Likelihood of Reporting Disinformation Displayed by Political Party



Note: Means with 95% CI displayed.

*$p < .05$.

5.4. Discussion

The present study sought to determine whether people are more lenient about sharing misinformation that has the potential to benefit an ingroup (versus an outgroup). Additionally, it also looked at whether people revise their moral evaluations upon learning a post is inaccurate. The study supported H1-3, as participants displayed ingroup favouritism in their moral judgements of misinformation. Specifically, misinformation supporting the ingroup, or undermining the outgroup, was judged as more acceptable to share than misinformation that may undermine the ingroup or supports the outgroup. There was also strong support for H4 as participants judged the post to be much less acceptable after learning it contained false information.

The findings here suggest that users may evaluate relevant social media content in the context of their identity unless it is in some way clear that the content is false or misleading. In the present study, all four groups were presented with posts that were identical except for small adjustments to the target (e.g. Conservative or Labour) or stance (e.g. support or undermine). The responses here indicate that users may make flexible moral evaluations of misinformation in relation to whether a post helps or hinders their ability to have a positive social identity. That is to say participants made such judgements in an identity-protective manner. For instance, whether or not a post threatened the value of the ingroup mattered, as sharing misinformation that expressed the positive achievements of the ingroup (e.g. reduced violent crime rates) was judged to be more morally acceptable than sharing misinformation that undermined the ingroup. Arguably, the latter may threaten the perceived value of the ingroup and therefore sharing may be viewed as anti-normative. As people are also known to react negatively towards ingroup norm violators (e.g. Abrams et al., 2002; Brambilla et al., 2013), it is somewhat logical to expect that ingroup undermining information will be viewed as less morally acceptable to spread.

Moreover, it was not the case that sharing misinformation expressing rising violent crime rates was perceived to be generally unacceptable to share. Indeed, misinformation that undermined a relevant outgroup was judged as more acceptable to spread than misinformation that undermined the ingroup. Arguably, the former may allow users to maximise intergroup differences and make positive comparisons (e.g. achieve positive distinctiveness) and, as such, can help them achieve a positive identity. This supports prior research suggesting that people make motivated moral judgements in the context of their salient identity (e.g. Uhlmann et al., 2009). Outgroup undermining misinformation may therefore in some way prove useful for achieving identity-relevant goals. As such, the valence of the post itself (and indeed whether the post may be in some way detrimental to

other people) may be less important than its value in terms of aiding social identity strategies.

Moreover, participants made much harsher moral judgements upon discovering that the post contained inaccurate information. The large effect sizes seen here may indicate that learning the post was disinformation updated the contextual basis of the judgement and suggests that sharing disinformation is generally seen as "wrong" to do. While that is not to say that participants refrained from considering accuracy when it was not actively disclosed, the knowledge it was "false information" may have presented a potential moral violation in a way that the post previously did not. As affective processes are thought to help draw attention to such threats and, in turn, may guide moral judgements (Haidt & Kesebir, 2010), participants may have relied on moral intuition to alert them to such threats. Indeed, given that intentional violations of moral norms may be judged more harshly than accidental violations (Parkinson & Byrne, 2018), even social media users who do not feel strongly about disinformation themselves may be sensitive to the potential threat arising from violating a social norm. A reliance on affective cues may therefore also explain why participants felt it was relatively more acceptable to spread identity-affirming misinformation compared to misinformation that potentially would undermine achieving a positive identity. It may also help to explain why people report caring about spreading accurate information, but may not necessarily consider accuracy prior to sharing (Pennycook, Epstein, et al., 2021).

Despite judgements of disinformation being comparably "harsher", they also appeared to remain biased. However, exploratory analysis indicated a political asymmetry in moral judgements of disinformation. Prior to learning that the post contained false information, Conservative and Labour supporters made similar moral judgements about sharing. That is to say, they both made preferential judgements in the context of their identity. Moreover, upon learning the posts were inaccurate, both groups judged sharing the content to be significantly less moral and Conservative voters continued to make

identity-based moral judgements. However, Labour voters no longer demonstrated bias in their moral judgements of disinformation. This was despite controlling for other potential factors, and replicates the findings from study one where, unlike Conservative voters, Labour voters made judgements of favourable and unfavourable disinformation in a similar manner. Such findings also support previous work suggesting people who are politically right-leaning may be more tolerant towards politicians who lie (De Keersmaecker & Roets, 2019). Moreover, as moral evaluations of disinformation can help to predict whether people go on to spread it further (Effron & Raj, 2020), the political asymmetry observed here may provide important context for research suggesting political conservatives may be more likely to spread disinformation (Baptista et al., 2021; Garrett & Bond, 2021; A. Guess et al., 2019). Any such moral leniency towards disinformation could potentially have real world implications in the context of social media spread.

There are several proposed explanations for this political asymmetry. Firstly, strength of affective responses to a moral dilemma may influence whether judgements relate to consequential or deontological principles (J. D. Greene et al., 2001). Judgements made by Conservative voters may have been influenced by perceived consequences. In contrast, given the lack of significant differences between the moral judgements of disinformation that Labour voters made, this suggests the posts may have been viewed as "wrong" to spread regardless of any potentially beneficial consequences (e.g. deontological). Research suggests such judgements may be the result of more emotionally-driven processing (J. D. Greene et al., 2001). Moreover, the level of emotional outrage can also predict the severity of assigned punishments (Kahneman et al., 1998). Given that affective processes also help alert people to potential norm violations (Ellemers et al., 2002), the differences between the two groups seen here may have arisen based on how strongly they perceive spreading identity-beneficial disinformation to be a norm violation. As illustrated in study one, not all disinformation is judged by each group equally.

Therefore, variations in group norms about spreading disinformation may help explain the political asymmetry observed here.

Alternatively, it could also be the case that there are underlying differences in how Conservative and Labour voters make moral judgements. Indeed, research on Moral Foundations Theory has suggested that political liberals are much more likely to prioritise "fairness" and "harm" than political conservatives (Graham et al., 2009). In this context, after learning the information was false, Labour voters may have been more attuned to the potential unfairness or harm of spreading such information (e.g. moral violations), even if spreading the information could potentially benefit them in other ways. Conversely, political conservatives have been shown to prioritise "binding" foundations such as ingroup loyalty more than political liberals (Graham et al., 2009; Voelkel & Brandt, 2019). While the findings here do indicate that Conservative voters did still care about the fact the content was inaccurate (as illustrated by the reduced moral acceptability scores), the leniency towards spreading identity-affirming disinformation and disinformation that would help make the ingroup appear positively distinct (e.g. disinformation that undermined the Labour party) suggests that identity-related goals (e.g. ingroup loyalty) were still being prioritised. Therefore, whether the political asymmetry observed here related to ingroup norms or underlying cognitive processes will be explored further in study four.

However, the social context surrounding the present study is also important to acknowledge when discussing the political asymmetry observed here. At the time of data collection, Labour was thought to have a substantial lead in the election. The loss of an election arguably presents a threat to the value of an ingroup, and can have a strong emotional impact on those who identify with a party (Pierce et al., 2016). As such, strong group identifiers are likely to feel motivated to behave in ways to tackle said threat (Ellemers et al., 2002). People may also more readily prioritise ingroup loyalty in moral judgements when said group is under threat (Leidner & Castano, 2012). Given that people

who are politically right-leaning may more readily prioritise ingroup loyalty anyway (Graham et al., 2009), false information which benefits their ingroup may arguably feel more acceptable to share if it helps them uphold said values in the face of threat (e.g. a consequential judgement). Put simply, it should not be ruled out that the timing of data collection in the present study may have potentially produced or amplified the political asymmetry found here. Future studies may therefore wish to explore the influence of external identity-threats on moral judgements and spread of disinformation.

Furthermore, as exploratory analysis here demonstrated, the unbiased evaluations of disinformation shown by Labour voters did not necessarily translate into their reporting behaviour. Overall they were slightly more likely to report disinformation when compared to Conservative voters and this appeared to be driven by Labour voters' intentions to report disinformation that undermined their own party. When it came to disinformation that undermined the outgroup, Labour voters were no more likely than Conservative voters to report. While the effect sizes are small, this may suggest that people selectively intervene in the spread of false information, even if they feel it is generally "wrong" to spread. Indeed, research suggests that through expressing moral credentials and intentions people may be able to protect their moral self (Monin & Miller, 2001) even if they go on to act in an "immoral" way (Cascio & Plant, 2015). As such, judging the sharing of disinformation to be "wrong" generally may allow people to feel they are "moral" even if they selectively report disinformation in a biased manner and highlights the difference between expressed morality and actual "intervention" behaviour. While selectively choosing to report known disinformation is of course preferable to taking no action, selectively engaging in such actions may arguably privilege the safety of certain groups over others. As such, the question of how people make moral evaluations of identity-affirming and identity-threatening disinformation is explored further in study four of the present thesis.

There are of course limitations with the present study. Firstly, the study used only a single item to test the hypothesis. As political conservatives are thought to be more likely

than political liberals to view police as authority figures (Frimer et al., 2014) it could be argued that this in some way influenced the political asymmetry observed here. Subsequent studies may therefore wish to introduce a greater variety of disinformation narratives balanced across the ideological spectrum. Moreover, the design of the present study required participants to evaluate the posts prior to and after learning it was untrue. It is therefore not possible to ascertain here how users judge freshly encountered disinformation as they may in real-world contexts. However, this question will be addressed further in study four. Finally, the findings are based on a very limited sample of social media users in London who vote for specific parties. Whether these findings apply to other groups or the wider population will need to be established.

Importantly, the present findings suggest that the saliency of disinformation being "disinformation" may be an important factor in whether individuals perceive it to be morally acceptable to share or not. Without this cue, users who would otherwise condemn the sharing of disinformation may readily spread misinformation unknowingly, without realising they may potentially be violating this moral standard themselves. Creating a greater awareness of the many forms that disinformation can take and the ways in which it may cause harm may help shift standards from simply "sharing accurate content" to "sharing content one has established to be accurate". If individuals are more accepting of being seen as "incompetent" than "immoral", then asking users to be critical of information presented to them may not always be effective. Ultimately, spreading false information may only be identified as a potential issue if could lead to negative moral consequences for themselves.

5.4.1. Conclusion

The present study sought to explore the influence of group membership on the moral judgements of identity-related disinformation. The day before an election, supporters of two opposing political parties were asked to rate the moral acceptability of

sharing one of four posts that either supported or undermined their ingroup or the outgroup. Participants were then informed that the contents of the post were inaccurate and asked again how morally acceptable it would be to share and also whether they would report the content to the platform. Prior to finding out that the information was false, participants made judgements that were more lenient towards sharing misinformation that could help their own party, regardless of their own political affiliation. Upon learning the post was untrue, participants judged sharing to be much less morally acceptable. However, across the political parties, supporters of the Conservative party retained an ingroup-bias in their judgements, whereas Labour voters judged all disinformation similarly. Conversely, Conservative voters were unlikely to report disinformation generally, while Labour voters were instead more likely to report disinformation that undermined their party.

Chapter 6. Study Three

6.1. Introduction

The main purpose of this chapter was to develop a scale to represent potential contributions to the onward spread of disinformation. While in study one the concept of "spread" was approached somewhat literally (e.g. only considering actions that would amplify reach), the present study also considers actions which may help to reduce spread. Additionally, following the findings regarding adjustments in moral judgements in studies 1 and 2, the present study provided an opportunity to better understand how this moral leniency may influence misinformation spread.

First, the impact of user actions on social media platforms (SMPs) will be discussed in greater detail with a specific focus on algorithmic contributions to spread. Next, users' willingness to report problematic content online and issue social corrections within SMPs will be addressed, followed by the influence of affective moral cues in guiding behaviour. A replication of study one is carried out to test the sensitivity of the "Social Media Spread" scale as a measure of overall likelihood of spread contribution. Additionally, a series of mediation analyses are utilised to understand how important moral judgements are in the relationship between beliefs and disinformation spread. Finally, group comparisons are again made to establish if there continue to be differences between how Labour and Conservative voters make these decisions.

6.1.1. The Algorithmic Amplification of Disinformation

Even seemingly small interactions have the ability to influence the onward spread of disinformation within SMPs. While the exact details surrounding the functionality of social media algorithms are mostly confidential, leaked internal documents from Meta (The Facebook Papers) confirmed that each type of "action" and relationship within Facebook may be weighted differently in regards to algorithmic calculations (Hagey & Horwitz, 2021). While it is difficult to know the unique contribution that each individual

action has on the onward spread of a post, most platforms do openly acknowledge that such interactions contribute to ranking in some way (LinkedIn, 2021; Mosseri, 2021; TikTok, 2020). The algorithmic ordering of SMP feeds is, however, central to understanding how content spreads within a platform.

In 2017 Facebook introduced "Meaningful Social Interaction" (MSI) rankings in attempts to personalise feeds and encourage users to spend more time on the platform. For instance, when calculating a post's position on a newsfeed, the number of likes it had received may contribute one point, "reactions" and reposts (with no text) five points, while comments, messages and reposts with text provided 30 points (Hagey & Horwitz, 2021). For example, a post generating 20 "angry" reactions would gain 100 points compared to a post with 20 "likes" (e.g. 20 points). However, data scientists at Meta raised concerns about this approach in 2019, finding that problematic posts (including those classed as "misinformation") were more likely to receive angry reactions (Merrill & Oremus, 2021). Today, it is thought that angry reactions hold no weight and therefore may no longer contribute to the algorithmic model. It is believed that this change led Facebook users to be shown less misinformation on their feeds (Merill & Oremus, 2021). Therefore, not only do the specific formulas develop over time but, notably, individual weights can be (and have previously been) adjusted. This suggests that the specific contribution of any action within the platform is unlikely to be stable. However, this real-life example reinforces the need to consider the user-spread of disinformation beyond direct "sharing" actions by incorporating AI-driven consequences into our concepts of user-contributions to "spread".

6.1.2. Intervening in the Spread of a Social Media Post

Exercising moral agency not only involves people avoiding acting in an inhumane manner (inhibitive) but also engagement in humane acts (proactive) (Bandura, 1999). These proactive acts are often driven by factors such as social norms and strongly held convictions. Therefore, in addition to simply refraining from interacting with

misinformation (inhibitive action), users may also choose to take proactive steps which may potentially help reduce said spread.

6.1.2.1 Reporting Content

One way in which users may help to reduce the spread of content is to anonymously report it to the platform. Although not all reported content will be taken down, SMPs such as Facebook claim to reduce the reach of reported content that is determined to be "fake" by third-party factcheckers, through a process known as "downranking" (Silverman, 2017). Additionally. users who repeatedly share misinformation may also have the reach of all their posts limited, even when it is "true" (Facebook, 2021). Some researchers have also found users are less likely to spread content which has been assigned a warning label (Lanius et al., 2021; Mena, 2020). Even if the content is not necessarily taken off an SMP after being reported, by limiting its reach the spread may ultimately be reduced.

However, only a small minority of users may engage with reporting functions when faced with content they know to be false (Tandoc et al., 2020). For instance, certain individual differences may influence reporting likelihood, as more readily prioritising moral values of harm and fairness have been associated with reporting problematic content online (Wilhelm et al., 2020). However, the uncertainty of the outcomes post-reporting may also be important, as users are more willing to report problematic content if they perceive reporting may lead to an effective outcome (Wong et al., 2021). Therefore, people may need to not only know how to report, but also feel that it will be taken seriously before they make the effort.

The content itself, however, is likely to play an important role in whether someone reports. Indeed, the perceived severity of any potential norm violation may also influence whether users report misinformation to a platform. Indeed, people are more likely to be willing to report content when they perceive it to be an emergency situation (Wong et al.,

2021). This may be why people are more likely to report violence than rumours and conspiracies (Wilhelm et al., 2020), and hate speech compared to disparaging speech targeting the same group (Kunst et al., 2021). In the context of misinformation, this perceived "severity" may be literal such as the difference between vaccine disinformation and political disinformation, which may ultimately influence what type of disinformation content users report. Finally, as demonstrated in study one, levels of belief-consistency may also influence how acceptable a user perceives the spread of misinformation to be.

6.1.2.2 Social Corrections: Willingness to Intervene and Self-Censorship

While reporting content to a platform affords a user a level of anonymity, users may also directly intervene (either privately or publicly) when encountering disinformation on social media in a way that has the potential to help reduce spread in a number of ways. Firstly, social corrections in the form of comments from other users may provide feedback to the original poster (OP) regarding how their post is being perceived by others (Y. Wang et al., 2011). As such, comments that criticise a post (for instance, highlighting that it is disinformation) may signal to the OP that they have (inadvertently or otherwise) violated social norms. If the OP then regrets the post, then research suggests they may delete it in an attempt to resolve the situation (Y. Wang et al., 2011). They may even be willing to delete the post when it potentially benefits them in other ways. For instance, (Mun & Kim, 2021) found people who are more likely to use self-presentational lies within social media posts were also more likely to delete their posts afterwards, potentially due to perceived psychological risks to others. Therefore, even if an OP is not initially aware that their post potentially violates social norms, they may go on to remove it if they learn said post presents a risk to themselves or others.

Yet, even if a post is not removed by an OP or the SMP, then comments criticising the post may still help reduce spread albeit indirectly. For instance, when users point out the factual inaccuracies of a post within the comments these social corrections may help

other users identify it as disinformation (Bode et al., 2020; Bode & Vraga, 2018). Similarly, the presence of negative comments on a post may reduce the likelihood that others will share it, even when other positive comments are present (Boot et al., 2021). Indeed, research by Colliander (2019) suggests social corrections may be more effective at changing users' attitudes and intentions to interact with misinformation than official warning labels from SMPs. Therefore, users may still be able to play a part in helping reduce the wider spread of a post through critical commenting.

Certain factors within the content may increase the likelihood that a user intervenes in this way. For instance, people report more willingness to leave a critical comment on disinformation that threatens their ingroup, potentially due to concern that others may believe it (E. L. Cohen et al., 2020). Users are also more likely to publicly or privately intervene in cyberbullying on Facebook when the victim is perceived to be similar to themselves (S. Wang, 2021). Therefore, if people feel they are negatively impacted by misinformation in some way they may be more willing to speak up, not necessarily because they know it to be untrue. Moreover, emotional responses of anger and depression can increase a person's willingness to speak out, even within hostile opinion climates (Masullo et al., 2021). Therefore, perceived identity threats and affective responses may increase the likelihood that users are willing to speak out against content, potentially even publicly. Disinformation that is perceived as a clear threat may therefore generate more public criticism than disinformation where the threat is ambiguous or abstract.

Posting critical comments is, however, not without risks. As such, users may be conscious of the visibility of their comments, particularly if a correction carries social risks within the immediate environment. This may be an especially important consideration in the context of SMPs, where users have to be conscious of both the somewhat permanent record of any act as well as considerate of multiple social identities. Indeed, the perceived presence of a virtual audience consisting of ingroup members may influence identifiable users to express themselves in a normative manner (Douglas & McGarty, 2001, 2002).

Therefore, users may potentially be careful about how they critique posts in the context of their social identities and personal identifiability. As research also suggests increased fear of social isolation may lead people to refrain from publicly expressing opposing opinions within SMPs (H. T. Chen, 2018), inaction may be a more appealing prospect under certain conditions. As such, any underlying threats presented by misinformation may need to be weighed up against the perceived threats associated with speaking out.

This may be why uncertainty over who ultimately views the post may also affect decisions to intervene. Research suggests audience size may influence whether a user chooses to intervene within online environments (Obermaier et al., 2016). Arguably, in some digital environments larger audiences may increase the likelihood of ingroup members seeing a comment. This may also explain why users may be more likely to experience social corrections within intimate digital environments (e.g. WhatsApp) compared to Facebook (Rossini et al., 2021). Rather than risk other ingroup members perceiving a critical comment as being a norm violation, users may instead choose to anonymously report, or alternatively contact the poster directly.

6.1.3. Affect, Morality and Engaging in "Immoral" Behaviour

Users may not be consciously aware of the many evaluations they will be making while scrolling through their SMP feeds. They may, however, at times experience a sense that a piece of content is "right" or "wrong", without any conscious effort to make such a judgement. Social intuitionism proposes that moral intuitions are automatic, affective processes promoting evaluations of "good" or "bad" (or indeed "right" or "wrong') and may influence any subsequent moral reasoning that occurs (Haidt, 2001). From this perspective, if an individual were to view a piece of content on an SMP that perhaps violated a moral norm they may automatically sense that it is "wrong" to spread it further, but they may have to work, and even struggle, to articulate exactly "why".

Additionally, affective processes are thought to help guide (rather than dictate) behaviour (see Baumeister et al., 2007), and so may unconsciously influence whether individuals choose to engage with misinformation or not. From a social cognitive perspective, when a person violates their own personal moral standards they may experience strong, negative affect as a consequence (Bandura, 1991b). People may therefore self-regulate their behaviour to avoid this happening. Indeed, previous work suggests that the more that people morally condemn spreading a piece of disinformation, the less likely they are to share it with other people (Effron & Raj, 2020; Helgason & Effron, 2022). However, repeatedly encountering disinformation was found to reduce levels of moral condemnation and, in turn, was associated with an increased likelihood of sharing the content (Effron & Raj, 2020). One explanation for this is that repeated exposure may reduce the strength of each subsequent affective response. This somewhat suggests people may not make objective moral evaluations of disinformation, and instead may be guided by feelings of "wrongness".

There are, however, a multitude of reasons why people could experience negative affect upon viewing misinformation. Firstly, they may notice a potential moral violation, for instance, that the content is potentially harmful to others in some way. While sharing the content may not have a negative impact on them personally, affective forecasting may allow them to anticipate emotions that could be experienced if they were to press "share" (see Wilson & Gilbert, 2003 for an overview). For instance, if a person remembers a negative experience after sharing something similar in the past, they may also experience emotions associated with that memory. These emotions may ultimately encourage the user to self-regulate to avoid potentially encountering the experience again.

However, words, images, or concepts within the content may induce affect in a manner that can shape evaluations. For instance, previous work suggests that the emotions that people experience when thinking about "global warming" may influence judgements made about climate change policy (N. Smith & Leiserowitz, 2014). Furthermore, people

may experience cognitive dissonance when they view information that conflicts with their beliefs, a process that is thought to be accompanied by negative affect (Harmon-Jones, 2000). As such, the influence of affect is likely to differ across situations and individual users. Yet, these emotional cues may also increase the likelihood that someone believes misinformation (Martel et al., 2020). Moreover, in addition to guiding the self-regulation of behaviour, certain emotions are more likely than others to encourage action (see Brader & Marcus, 2013). For instance, Tweets containing moral emotional language are thought to be more readily spread within ingroup networks on Twitter (Brady et al., 2017). As such, the relationship between affect and people's evaluations of social media content is unlikely to be straightforward.

With this in mind, users may need to have a realistic understanding of the potential impacts of disinformation to effectively self-regulate their behaviour. However, the first challenge here is that the impact of disinformation is often difficult to comprehend let alone quantify. For instance, research suggests that lies that benefit the receiver (e.g. prosocial and altruistic lies) may be viewed as more morally acceptable than the truth and lies that harm the receiver (Levine & Schweitzer, 2014). People may therefore view disinformation that they perceive to be prosocial as more morally acceptable if they are otherwise unaware of the true reason the content was initially disseminated. For instance, the Internet Research Agency have previously spread disinformation depicting the suffering of children in the Syrian war, but in an attempt to drive support for Russia's operations in Syria and President Bashar Al-Assad (DiResta et al., 2019). As was found in study one, people may also be more morally lenient towards belief-consistent disinformation, and therefore being factually "inaccurate" is not necessarily enough to make disinformation feel "wrong" to spread if it "feels true" in other ways.

Another challenge is the reliance on people to consider the possibility that the content they encounter within SMPs may be disinformation (and as such "wrong" to spread). As the findings from study two suggest, once individuals are made aware that

content is false or misleading, they may judge it to be less acceptable to spread. However, prior to this their judgements may be more readily influenced by other factors, such as pre-existing beliefs and identity. If an individual is presented with belief-consistent misinformation (which therefore may align with what they perceive to be accurate (Huber, 2009)), they may arguably be less likely to sense potential moral violations relating to spreading inaccurate content. In contrast, content that potentially undermines such beliefs may be more readily identified as potentially morally "wrong" to spread on the sole basis that it conflicts with what is perceived as being true (but notably may not actually have to be factually untrue). As such, levels of belief-consistency may play an important role in influencing the moral evaluations people make about spreading misinformation.

6.1.4. The Present Study

A key focus of this study is to test a "Social Media Spread" scale that may better represent the contributions people may make to the spread of disinformation, either through actions which directly contribute to said spread or inaction. By incorporating actions that may also reduce the spread of disinformation on social media this scale may help to in some way differentiate between users who choose not to amplify the spread of content because of disagreement from those who are simply disinterested.

Furthermore, as in study one, it is proposed that individuals will be more likely to spread disinformation when it is consistent with or supports their beliefs. Again, the study measured participants' beliefs surrounding the UK Government's handling of the COVID-19 pandemic (measured by the Citizen Trust in Government Organisation scale (Grimmelikhuijsen & Knies, 2015)). The related disinformation was the same as used in study one, and framed as either "favourable" or "unfavourable" towards the UK Government and their performance. In other words, it was predicted that an individual whose beliefs would result in a low "trust" score would be more likely to spread "unfavourable" disinformation as it is more consistent with said belief.

The following hypotheses are therefore slight adjustments to study one hypotheses. The amendments replace the previous focus on interactions with disinformation and consider a combination of factors that reflect actions which amplify or reduce social media "spread'. It is predicted that individuals will be more likely to contribute to the spread of disinformation when it supports an issue-related belief. Therefore, Hypotheses 1 and 2 are:

H1: Individuals who have lower trust in Government handling of COVID-19 will report a greater likelihood of contributing to the spread of misinformation that undermines the Government.

H2: Individuals who have higher trust in Government handling of COVID-19 will report a greater likelihood of contributing to the spread of misinformation that supports the Government

The present study also provides an opportunity to better understand how moral reasoning influences intended interactions and spread, specifically how our moral judgements about sharing content play a role in whether we choose to engage with said content. Moral condemnation has previously been found to mediate the role between fluency (based on repeated exposure) and interactions with disinformation (Effron, 2020). Additionally, Social Cognitive Theory supposes that individuals self-regulate their conduct against personal standards (Bandura, 1991b). If users perceive spreading the content to be morally problematic, they may avoid interacting with it, even if it may benefit them or is something they may usually engage with.

Similarly to study one, it is also predicted that individuals will be more likely to judge spreading misinformation that supports an issue-related belief as more morally acceptable. Therefore, hypotheses 3 and 4 are:

H3: Individuals with lower trust in the Government will report the sharing of misinformation that undermines Government as more morally acceptable than those with higher trust in the Government.

H4: Individuals with higher trust in the Government will report the sharing of misinformation that supports Government as more morally acceptable than those with lower trust in the Government.

Finally, it is predicted that when beliefs have a greater consistency with misinformation participants will view it as more acceptable to spread and, in turn, be more likely to spread it themselves. Therefore, Hypotheses 5 and 6 are:

H5: Moral judgements of sharing "Government undermining" misinformation will mediate the relationship between low trust and increased likelihood of spreading "undermining" misinformation.

H6: Moral judgements of sharing "Government supporting" misinformation will mediate the relationship between high trust and increased likelihood of spreading "supporting" misinformation.

6.2. Method

6.2.1. Materials and Measures

The procedure and materials for this study were replicated from study one, with any changes noted below. Ethical approval for this study had previously been obtained from the University's Psychology Ethics Committee (ETH2021−0777) for study one (Appendix B).

6.2.1.1 Social Media Spread Scale

Six items from the original study were presented at random to participants. This was misinformation that was either "Favourable" or "Unfavourable" towards the UK Government. As before, participants were not explicitly informed that the content was false or misleading to avoid social desirability effects.

For this study, participants were instead asked "If this image came up on your social media feed, how likely is it you would engage with the following actions?". They were then presented with a list of eight actions: "Like or upvote the content", "Comment in agreement / support", "Repost the content on a personal social media account (e.g. "retweet")", "Send the content directly to one other person", "Share the content with a group of other people (e.g. WhatsApp group)", "Report the message to the platform" (R), "Post comment asking for content to be taken down" (R) and "Directly contact the poster to ask them to remove" (R). For each of these actions, responses were given using an 11-point scale from "Not at all likely" to "Extremely likely".

As indicated above, prior to analysis three items within the scale were reverse scored as they represented actions that may help reduce the wider spread of disinformation through either removal or algorithmic de-ranking. Therefore, proactive attempts to help prevent the spread of disinformation may be differentiated from inaction. The scale items for each item of stimuli were summed and a mean score created. The scores from the three stimuli items were then combined and an overall mean "spread" score produced for each disinformation category.

6.2.1.2 Moral Judgements of Sharing Disinformation

Participants were also asked to rate how morally acceptable it would be to share each item of misinformation. Responses were given using an 11-point scale from "Not at all acceptable" to "Completely morally acceptable'. However, unlike in study one, participants were not informed that the items were misleading until the debrief.

6.2.2. Participants

To ensure enough power for mediation analysis, sample size planning was conducted using MedPower (Kenny, 2017). A minimum effect size of β = .2 is thought to be the minimum effect size that is practically significant in social science research (Ferguson, 2009). To detect β = .2 at 80% power, 250 participants would be required. Allowing for data screening exclusions, the target sample size was 280 participants.

A total of 302 participants were initially recruited through Prolific. Of this, eighteen participants were unable to progress in the study due to having accessed it using a mobile or tablet device and so were removed from the dataset[4]. One participant was removed for not consenting, another for not meeting the recruitment criteria regarding current location, while a further three did not complete the study. Additionally, Qualtrics flagged four participants as suspicious, and as such were also removed. Finally, 24 participants were removed for a lack of variance in their "Trust" or "Risk" scores[5], suggesting inauthentic responding. The remaining 251 participants (83 males) aged 18-71 (M = 35.47, SD = 13.29) were those included in the analysis.

As with study one, participants were required to be based in England to ensure a consistent understanding of "Government'. Recruitment was expanded to include users of other social media platforms and was balanced across political ideology only. Again, participants on the whole reported they would vote for one of the two major political parties if an election were held tomorrow, Conservatives (N = 90), Labour (N = 85), Liberal Democrats (N = 6), Other (N = 34), Unsure / Would not vote (N = 32), Prefer not to say (N = 4). When political parties are compared in analysis, participants who indicated

[4] Accessing the study from a computer was a condition of signing up to the study, as the mobile optimised version on Qualtrics was less user-friendly and therefore the option was switched-off. The Qualtrics survey was then set up to detect any participant attempting to inauthentically access the study on a mobile device and automatically prevent them from progressing in any way (however, were still picked up by the software as having started the study).

[5] Removal of participants for this reason had also been specified at pre-registration.

anything other than Conservatives or Labour were excluded from analysis due to small samples.

Participant demographics are shown in Table 6.1.

Table 6.1

Participant Demographics for Study Three

	N	%
Total	251	100.0
Gender		
Female	163	64.9
Male	83	33.1
Non-Binary	3	1.2
Prefer not to say	2	0.8
Education completed		
Less than GCSEs	4	1.6
GCSEs	25	10.0
A-Levels	78	31.1
Bachelor's Degree	96	38.2
Master's Degree	42	16.7
Doctoral Degree	2	0.8
Other	4	1.6
Political Party		
Conservatives	90	35.9
Labour	85	33.9
Liberal Democrats	6	2.4
Other	34	13.5
Unsure / Would not vote	32	12.7
Prefer not to say	4	1.6

6.2.3. Data Analysis

Data analysis for planned tests were pre-registered through AsPredicted.org (#78270, Appendix R). All tests applied an α level of .05. As with study one, four multiple regressions were run predicting the spread and moral judgements of "Favourable" and "Unfavourable" misinformation. These used predictors of "Trust in Government",

"Perceived Risk", age and gender as predictors. Additionally, Spearman's correlations were run following each regression to confirm the findings.

H5 & H6 were tested using mediation models using a bootstrapping approach (n = 5000), where "Trust" was the predictor variable for both models. Significant predictors of both spread and moral judgements were also included as covariates. The first model focused on Unfavourable misinformation, while the second model addressed Favourable misinformation. "Trust" acted as the predictor variable, while the corresponding moral judgement acted as the mediator and social media spread score as the dependent variable.

6.3. Results

Qualtrics data were exported into Excel for data cleaning before importing into SPSS for analysis. As per Study one, responses for "Trust in Government", "COVID-19 Perceived Risk" and the morally acceptability ratings for each of the two disinformation stimuli sets were summed and mean scores calculated. To create the social media spread scores, reverse scores were calculated for the relevant items (e.g. "reporting", etc.) and pooled mean scores were created for each disinformation item as well as the full stimuli sets.

Descriptive statistics for all variables are shown in Table 6.2, with histograms and Q-Q plots provided in Appendix S.

Table 6.2

Summary of Descriptive Statistics by Misinformation Category

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skewness	Kurtosis
Age	251	35.47	13.29			18-71	0.61	3.37
Trust in	251	3.37	1.46	.96	1-7	1.11-6.67	0.25	−0.96
COVID-19 Perceived Risk	251	2.76	0.69	.84	1-5	1.13-4.63	−0.06	−0.18
Spread Likelihood								
Favourable	251	5.20	1.19	.60	1-11	1.33-9.38	0.10	2.12
Unfavourable	251	5.80	1.46	.82	1-11	2.25-11.00	1.38	2.20

	N	*M*	*SD*	*α*	Range		Skewness	Kurtosis
					Potential	Actual		
Moral Judgement								
Favourable	251	6.93	2.73		1-11	1.00-11.00	−0.31	−0.59
Unfavourable	251	5.45	2.91		1-11	1.00-11.00	0.29	−0.80

6.3.1. Planned tests

6.3.1.1 Social Media Spread Scale

A series of Cronbach Alpha tests were run to check the reliability of the Social Media Spread Scale. These were run across the responses for each individual misinformation item. All items individually had an $\alpha > .69$ suggesting an adequate level of reliability (Table 6.3).

Table 6.3

Cronbach Alpha scores for Individual Disinformation Items

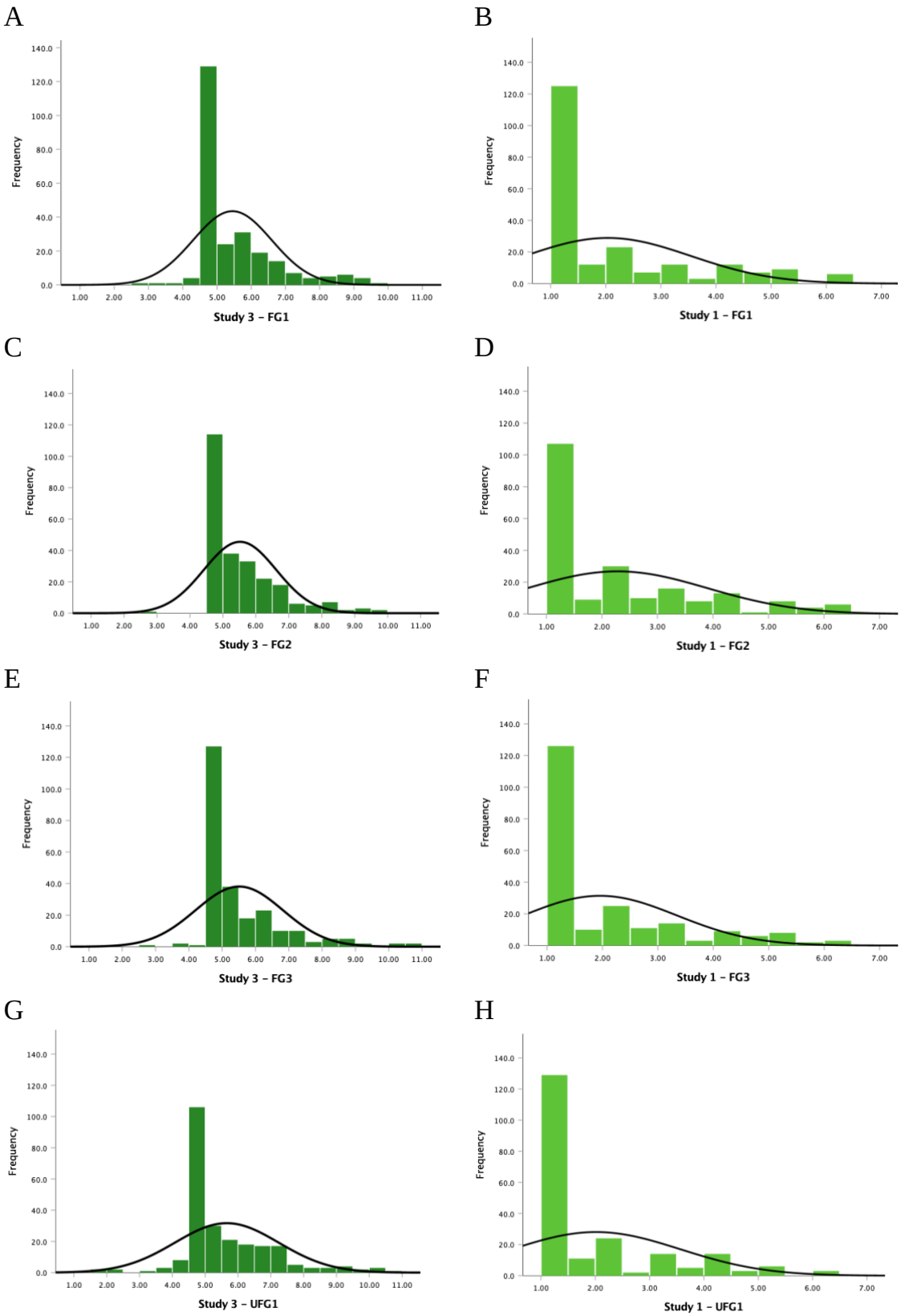
Item	α
FG1	.74
FG2	.69
FG3	.79
UFG1	.80
UFG2	.79
UFG3	.81

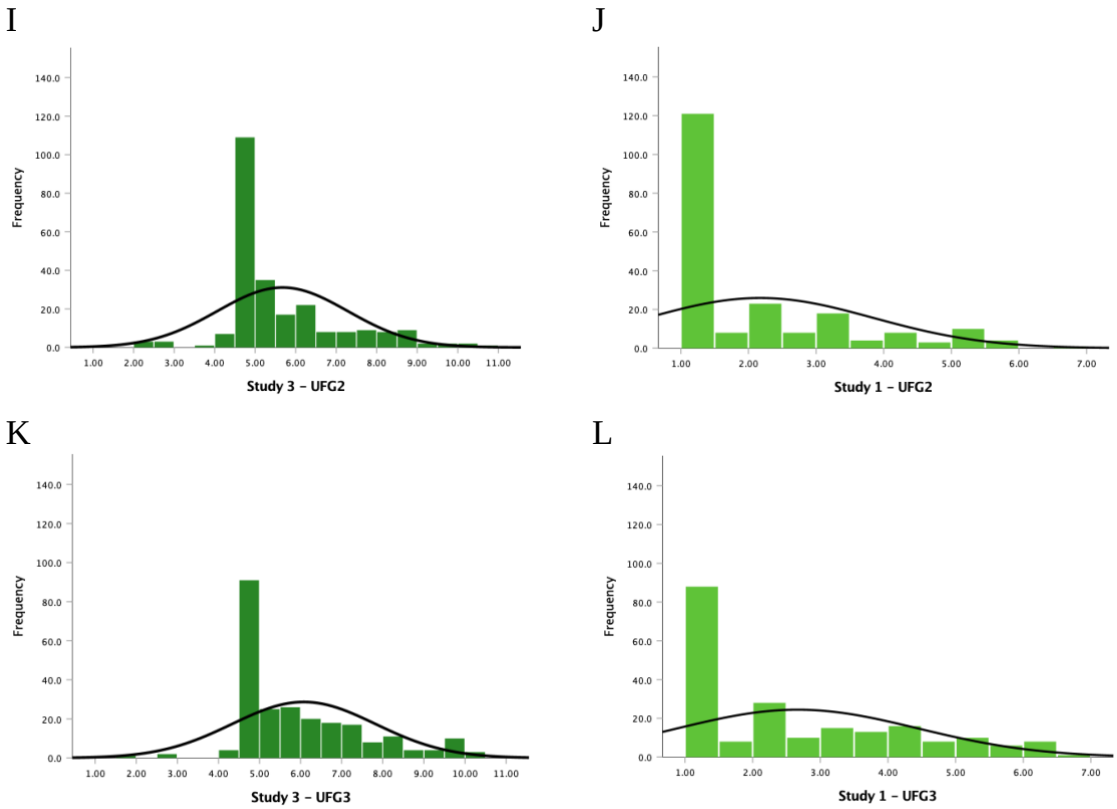
Note. "FG" = "Favourable" to UK Government; UFG = "Unfavourable" to UK Government.

Figure 6.1 also displays the distribution of the new social media spread scale parallel to the findings from study one (where only "like", "share with a friend" and "share publicly" were included in the scale) for each individual item. These show that the distribution of the scores has slightly improved, as those who engage with opportunities to reduce spread may be differentiated from "true inaction". While fewer participants received lower scores, this is to be expected with content that does not clearly fall into the category of "problematic content". A sharp peak will always be expected in this context, however, as individuals do not interact with every piece of content they see on social media (regardless of the motive of the interaction).

Figure 6.1

Histograms of "Spread" (Study Three) vs "Interaction" (Study One) Scores for Each Item





Note. Panels A, C & E show distributions of intentions to spread individual "Favourable" misinformation items. Panels B, D & F show the corresponding distributions for intentions to interact from study one. Panels G, I & K show distributions of intentions to spread individual "Unfavourable" misinformation items from the present study. Panels H, J & L show the corresponding distributions for intentions to interact from study one.

6.3.1.2 Effects of Belief Consistency on Spread of Misinformation

To ensure no violation of the assumptions for multiple regression, preliminary analyses were conducted to assess normality, linearity, multicollinearity, and homoscedasticity. Any violations are noted within the results.

As with study one, two multiple regressions were carried out, both with "Trust in Government", "COVID-19 Risk Perception", age and gender added as predictor variables to the models. The first model predicted the likelihood of spreading "Unfavourable" misinformation and the second model predicted the same for "Favourable" misinformation. The P-P plots for both models suggest that the residuals for both regressions may not be normally distributed (Appendix T) and therefore the results should be taken with caution.

The first model predicting the likelihood of spreading Unfavourable misinformation was significant, $F(4, 241) = 10.82$, $p < .001$, adj. $R^2 = .14$. While Trust in Government and "COVID-19 Risk Perception" were both significant predictors, Trust was the strongest ($\beta = -.31$, $t(241) = -4.76$, $p < .001$). The second model which predicted the likelihood of spreading "Favourable" misinformation was also significant, $F(4, 241) = 4.15$, $p < .01$, adj. $R^2 = .05$. While the second model accounted for only 5% of variance, this is above the minimum value for a practically significant effect for data in social science (Ferguson, 2009). Trust and age were both significant predictors of spread, with Trust again being the most important predictor ($\beta = .25$, $t(241) = 3.67$, $p < .001$). Regression coefficients for both models can be found in Table 6.4.

Table 6.4

Summary of Regressions Predicting Intentions to Spread Political Misinformation

	Unfavourable			Favourable		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	6.22***	.45		5.14	.38	
Age	−0.01	.01	−.09	−0.01*	.01	−.16
Gender	0.18	.19	.06	−0.16	.16	−.06
Trust	−0.31***	.06	−.31	0.20***	.06	.25
Risk	0.30*	.13	.14	−0.02	.11	−.01
R^2		.15			.06	
Adj. R^2		.14***			.05***	
F		10.82			4.15	

Note. Gender coded as dummy variable, M = 0, F = 1.

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

As the assumptions for a regression had not been fully met, Spearman's correlations were also run to lend further support for a significant relationship between the predictor of interest (e.g. Trust) and the dependent variables. These confirmed that Trust had a significant relationship with likelihood of spreading Unfavourable misinformation (*r*

= −.35) and Favourable misinformation (r = .2) with medium and small effect sizes respectively (Cohen, 1992).

6.3.1.3 Effect of Beliefs on the Moral Judgements of Misinformation

To assess whether levels of Trust in Government would predict moral judgements of sharing misinformation prior to learning the content is false or misleading, two further multiple regressions were carried out. The first model significantly predicted moral judgements of spreading Unfavourable misinformation, $F(4, 241) = 14.92$, $p < .001$, adj. $R^2 = .19$. Trust and age both added significantly to the model, with Trust being the most important predictor, $\beta = -.37$, $t(241) = -5.91$, $p < .001$. For Favourable misinformation, the model also significantly predicted moral acceptability ratings, $F(4, 241) = 8.34$, $p < .001$, adj. $R^2 = .11$. Trust, age and gender all added significantly to the model, but again Trust was the most important predictor, $\beta = .27$, $t(241) = 4.13$, $p < .001$. Regression coefficients for both models can be found in Table 6.5.

Table 6.5

Summary of Regressions Predicting Moral Judgements of Political Misinformation

	Unfavourable			Favourable		
	B	*SE B*	β	*B*	*SE B*	β
Model						
Constant	8.79***	.87		8.32***	.85	
Age	−0.03*	.01	−.15	−0.04**	.01	−.19
Gender	−0.07	.36	−.01	−0.89*	.35	−.16
Trust	−0.73***	.12	−.37	0.50***	.12	.27
Risk	0.13	.25	.03	−0.40	.24	−.10
R^2		.20			.12	
Adj. R^2		.19***			.11***	
F		14.92			8.34	

Note. Gender coded as dummy variable, M = 0, F = 1.

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

As before, Spearman's correlations demonstrated that moral judgements of sharing Unfavourable misinformation significantly correlated with "Trust" ($r = -.42$) with a medium effect size. Furthermore, Trust positively correlated with the moral acceptability ratings of Favourable misinformation ($r = .24$) with a small effect size.

6.3.1.4 Moral Judgements as a Mediator Between Beliefs and Misinformation Spread

Moral judgements of spreading misinformation were predicted to mediate the relationship between belief consistency and intentions to spread misinformation. First, moral judgements and intentions to spread Unfavourable misinformation were put into PROCESS macro model 4. Standardized results are shown unless noted otherwise.

Lower levels of trust were related to higher moral acceptance of spreading Unfavourable misinformation ($a = -.37$, $t(247) = -5.94$, $p < .001$). This, in turn, was related to a higher likelihood of spreading said Unfavourable misinformation ($b = .52$, $t(246) = 8.26$, $p < .001$). A 95% bias-corrected confidence interval based on 5,000 bootstrap samples indicated that the indirect effect ($ab = -.19$) was entirely below zero ($-.27$ to $-.12$), suggesting a significant result. However, when the indirect effect of the moral judgment was taken into consideration, lower levels of trust still predicted an increased likelihood of spreading Unfavourable misinformation on social media ($c" = -.12$, $t(247) = -2.18$, $p < .05$). Model coefficients can be found in Table 6.6 and Figure 3.6.

Table 6.6

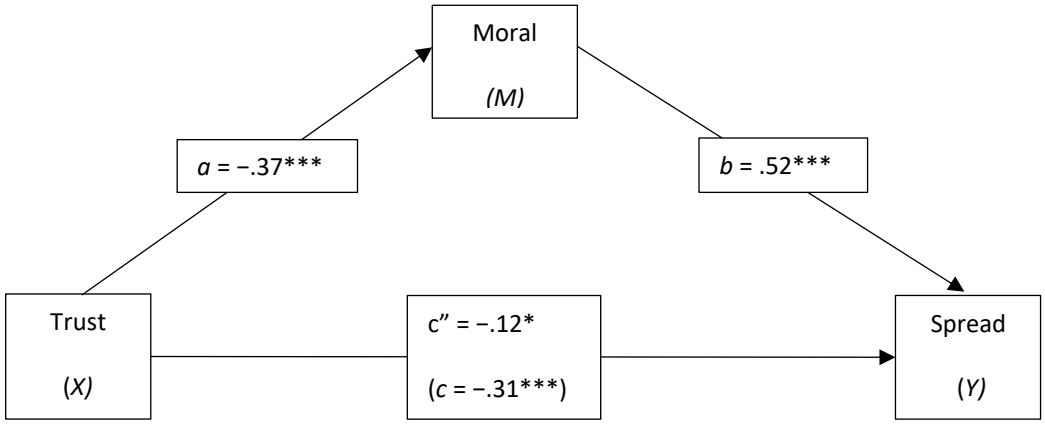
Model Coefficients for Mediation Model for Unfavourable Misinformation

		Consequent						
		M (MORAL				*Y* (SPREAD)		
Antecedent		Coeff.	*SE*	*Beta*		Coeff.	*SE*	*Beta*
X (TRUST)	*a*	−0.73***	.12	−.37	*c'*	−0.12*	.06	−.12
M (MORAL					*b*	0.26***	.03	.52
C_1 (RISK)	f_1	0.13	.23	.03	g_1	0.29**	.10	.14
C_2 (AGE)	f_2	−0.03*	.01	−.15	g_2	−0.001	.01	−.01
Constant	i_M	8.75***	.76		i_Y	4.03***	.48	
		R^2 = 0.20 F (3, 247) = 26.79, p < .001				R^2 = 0.36 F (4, 246) = 27.18, p < .001		

*p < .05. ** p < .01. *** p < .001.

Figure 6.2

Standardised Coefficients for the Relationship Between Trust and Likelihood of Spreading Unfavourable Misinformation Mediated by Moral Judgements



Note. Controlled for Risk and age. Presented effects are standardised

*p < .05. ** p < .01. *** p < .001.

The second model predicted intentions to spread "Favourable" misinformation. This found that levels of trust were related to higher moral acceptance of spreading

"Favourable" misinformation (a = .26, t(242) = 3.62, p < .001). Higher moral acceptance was again linked to higher likelihood of spreading the misinformation (b = .37, t(241) = 5.89, p < .001). A 95% bias-corrected confidence interval based on 5,000 bootstrap samples indicated that the indirect effect (ab = .08) was entirely above zero (.03 to .13). When the indirect effect of the moral judgment was again taken into consideration, higher levels of trust still predicted an increased likelihood of spreading "Favourable" misinformation on social media (c'' = .13, t(241) = 2.54, p < .05). Model coefficients can be found in Table 6.7 and Figure 6.3.

Table 6.7

Model Coefficients for Mediation Model for Favourable Misinformation

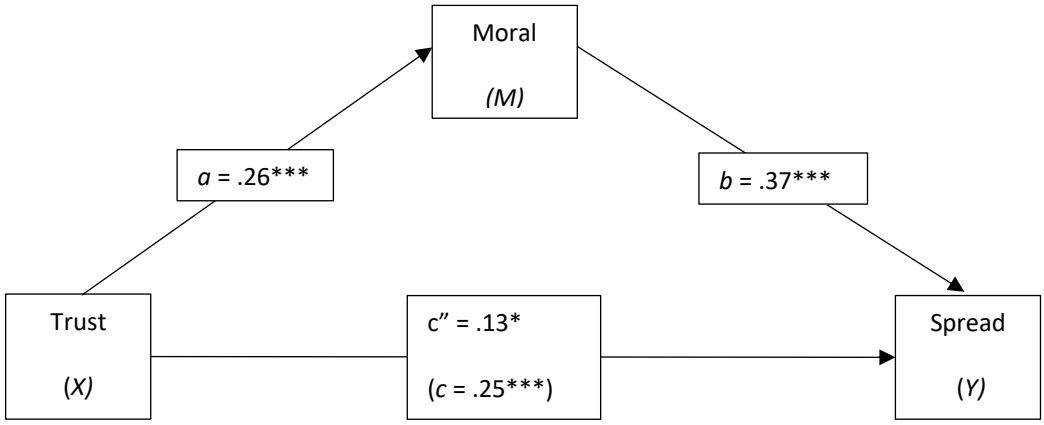
		Consequent						
		M (MORAL				*Y* (SPREAD)		
Antecedent		Coeff.	*SE*	*Beta*		Coeff.	*SE*	*Beta*
X (TRUST)	a	0.49***	.13	.26	c'	0.13*	.05	−.15
M (MORAL					b	0.16***	.03	.37
C_1 (AGE)	f_1	−0.04**	.02	−.21	g_1	−0.01	.01	−.08
C_2 (GENDER)	f_2	−0.98**	.35	−.17	g_2	−0.003	.15	−.001
Constant	i_M	7.48***	.62		i_Y	3.90***	.31	
		R^2 = 0.11				R^2 = 0.18		
		F (3, 242) = 8.33, p < .001				F (4, 241) = 15.47, p < .001		

Note. Gender coded as dummy variable, M = 0, F = 1.

*p < .05. ** p < .01. *** p < .001.

Figure 6.3

Standardised Coefficients for the Relationship Between Trust and Likelihood of Spreading Favourable Misinformation Mediated by Moral Judgements



Note. Controlled for gender and age. Presented effects are standardised.

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

6.3.2. Exploratory Analysis

6.3.2.1 Political Differences in Spread Intentions and Moral Judgements

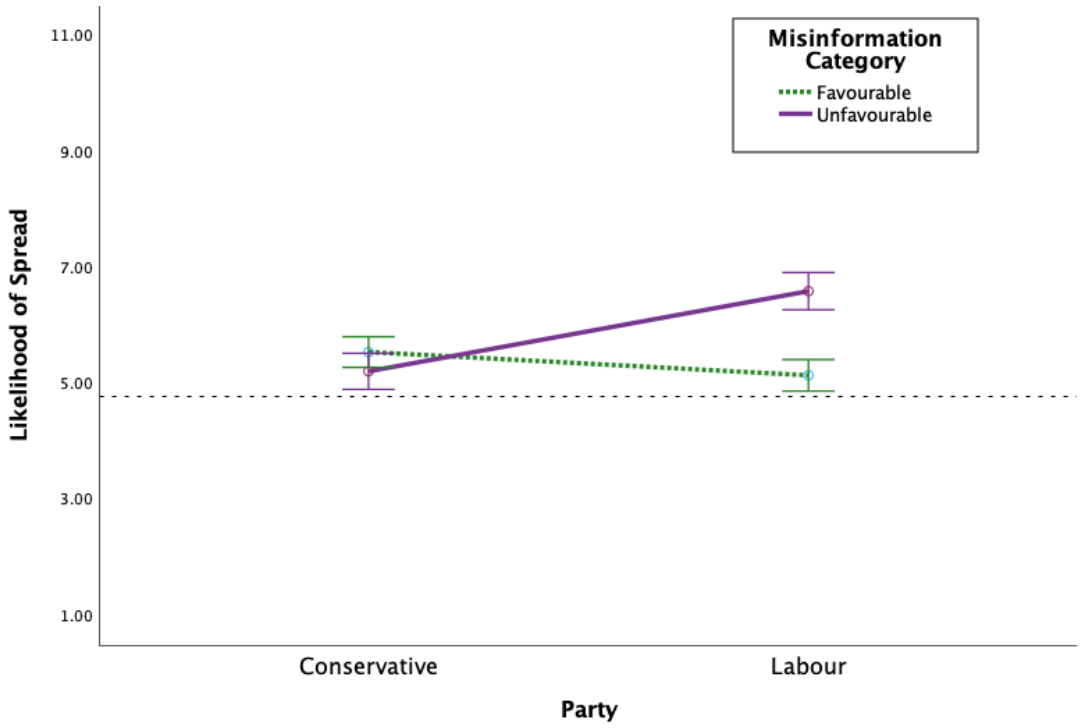
As in study one, there appeared to be differences in how both Conservative and Labour voters responded in relation to spread and moral judgements. As the variance in ages across the two groups were found to be significantly different ($t(153.59) = 7.30$, $p < .001$, $d = 1.09$), age was controlled for in these analyses. Two 2x2 mixed analysis of variances (ANCOVA) were run, both with between-group factors of "partisanship" (Conservative vs. Labour) and within-group factors of "misinformation type" ("Favourable" vs "Unfavourable') and controlled for age.

The ANCOVA for "Spread" revealed that the interaction effect between "partisanship" and "misinformation type" was significant with a large effect size, $F(1, 172) = 45.82$, $p < .001$, $\eta_p^2 = .21$. Individually, the main effect of "partisanship" was significant

($F(1, 172) = 7.42$, $p < .01$, $\eta_p^2 = .04$), however the main effect of "misinformation type" was not significant ($F(1, 172) = 0.03$, $p = .87$, $\eta_p^2 = .00$). Analysis of simple main effects suggest Labour voters were more likely to spread Unfavourable than Favourable misinformation with a large effect size, ($F(1, 172) = 66.40$, $p < .001$, $\eta_p^2 = .28$). This is illustrated in Figure 6.4.

Figure 6.4

Estimated Marginal Means of Likelihood of Spreading Misinformation



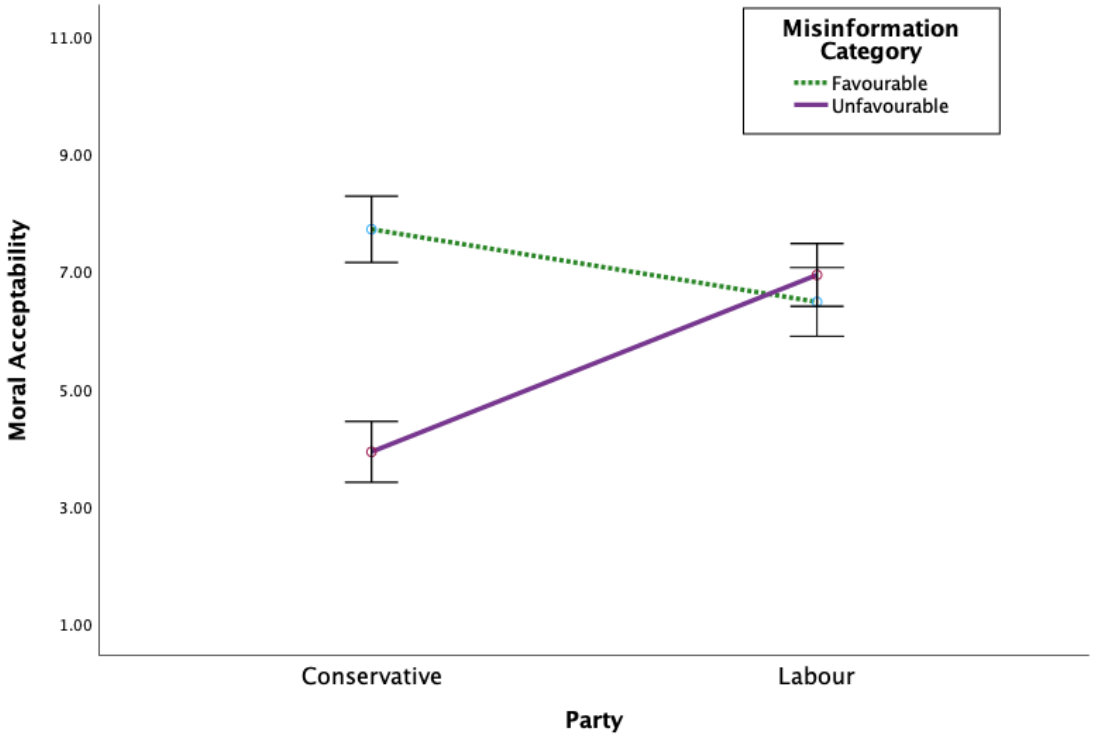
Note. Controlled for age. Error bars: 95% CI. Dashed line indicates point at which participants with no intentions to interact in any manner (e.g. amplify or intervene) would fall.

For the second ANCOVA testing the moral judgements of spreading misinformation, both the main effects of "partisanship" ($F(1, 172) = 6.77$, $p < .01$, $\eta_p^2 = .04$) and "misinformation type" were significant ($F(1, 172) = 11.11$, $p < .001$, $\eta_p^2 = .06$). Furthermore, the interaction effect between "partisanship" and "misinformation type" was also significant with a large effect size, $F(1, 172) = 75.80$, $p < .001$, $\eta_p^2 = .31$. Analysis of simple main effects found that Conservative voters were significantly more likely to feel it

was morally acceptable to spread Favourable compared to Unfavourable misinformation with a large effect size, ($F(1, 172) = 140.89$, $p < .001$, $\eta_p^2 = .45$). This is illustrated in Figure 6.5.

Figure 6.5

Estimated Marginal Means of Moral Acceptability of Spreading Misinformation



Note. Controlled for age. Error bars: 95% CI.

6.4. Discussion

A key aim of this study was to understand whether a "Social Media Spread" scale could help to capture users' intentions to contribute to the spread of misinformation. It was also an opportunity to better understand whether moral judgements of sharing misinformation help to explain the relationship between belief-consistency and intentions to spread. The new scale had acceptable reliability across all six misinformation items. H1 and H2 were supported using the spread scale, in that participants reported being more likely to spread misinformation that was consistent with their views. Furthermore, H3 and H4 were supported, in that moral judgements of misinformation spread (prior to learning

the content was false or misleading) could be predicted by greater levels of belief-consistency. Finally, H5 and H6 were supported, as moral judgements of spreading misinformation accounted for a proportion of the relationship between belief-consistency and misinformation spread.

The findings suggest that the social media spread scale may be a suitable measure for capturing intentions to contribute to the digital spread of a SMP post. Not only did the scale have an acceptable level of reliability, the regression models for both types of misinformation were significant and accounted for acceptable amounts of variance. On the whole, it is expected that when presented with misinformation (or any type of social media content) that the majority of users will simply do nothing, which the distribution of these scores certainly support. Given that users are more likely to intervene in problematic behaviour online when the situation induces negative responses in said user (E. L. Cohen et al., 2020; Masullo et al., 2021), it may also be the case that a different selection of items would produce better distributions and potentially lower likelihood of spread overall. The social media spread scale therefore provides an opportunity to distinguish proactive attempts to reduce spread from simple inaction.

As in study one, the present findings indicate people may make moral evaluations of misinformation in relation to how closely it aligns with their beliefs (e.g. levels of belief-consistency). The difference here is that participants were not informed the content was false or misleading which, as study two suggests, may potentially reduce the overall levels of moral acceptability. Given how beliefs represent what a person perceives to be true (Huber, 2009) people may arguably be less likely to perceive potential moral violations in belief-consistent misinformation compared to misinformation that conflicts with said beliefs. As the latter may undermine what they perceive to be true, users may experience cognitive dissonance and, therefore, negative affect (see Harmon-Jones, 2000). Therefore, users may sense that content that undermines their beliefs is “wrong” to spread further, regardless to whether it is factually inaccurate or not.

Conversely, if people spread false information that they genuinely perceive to be "true" they are arguably unlikely to feel they are being dishonest or deceitful (see Barber, 2020 for an overview) or experience cognitive dissonance. Therefore they may potentially be less likely to experience a sense that is it "wrong" to spread the content further, and therefore may not perceive the act as a potential moral violation. Additionally, people also rely on affective cues associated with potential moral violations to inform when to regulate their behaviour (Bandura, 1991b). As such, users may feel relatively free to spread belief-consistent misinformation from a moral perspective and therefore other factors may determine whether they go on to amplify the content further.

Notably, while the weight of the predictors themselves were similar in both study one and the present study, the amount of variance the model accounted for in judgements of favourable disinformation almost doubled when the inaccurate "status" was not disclosed (e.g. misinformation). It may be that learning that information is inaccurate influences the basis against which these judgements are made. As in, beliefs may have played a greater role in determining moral judgements of misinformation, suggesting that learning disinformation is untrue provides additional, but potentially important, context from a moral perspective. Interestingly, however, this was not the case for unfavourable misinformation, where regression results in both studies were similar despite the differences in knowledge of disinformation "status". Whether any changes in variance are due to differences in "disinformation disclosure" cannot, however, be established from the present findings, particularly given the time frame between the two studies. The effect of fact-check tags and content valence are tested within study four of the present thesis.

Moreover, the present study supports recent work suggesting that levels of moral acceptability can act as mediator in predicting intentions to spread misinformation (Effron & Raj, 2020; Helgason & Effron, 2022). Previously, Effron & Raj (2020) found that repeated exposure to a piece of disinformation reduced levels of moral condemnation, which in turn increased intentions to share. Here, moral acceptability related to levels of

belief consistency. Together, these findings highlight the value of exploring why people may be more morally lenient towards misinformation to better understand why they may go on to spread it further.

Given the findings from studies one and two, it was somewhat expected that Conservative and Labour voters would judge the misinformation differently from each other, despite neither group being made aware that it was misleading. On the whole, Labour voters again were more likely to spread misinformation that undermined the UK Government (e.g. outgroup) and appeared to judge the spread of both types of content similarly, despite having much lower trust overall. However, Conservatives judged misinformation that supported the UK Government (which at the time of data collection was their own party) to be more acceptable to spread, and misinformation that negatively framed the Government as much less acceptable to spread than any other rating. From the perspective of Social Identity Theory (Tajfel & Turner, 2004), this may be because undermining misinformation negatively frames their ingroup and therefore threatens the group's value. Behaviour which may have a negative impact on others or the self can violate a subtype of social norm known as a moral norm (FeldmanHall et al., 2018; Schein & Gray, 2018). Notably, violations of moral norms may lead to more severe punishment (Schmidt et al., 2012), which may explain the findings here.

In contrast, favourable misinformation may allow users to express the positive distinctiveness of their ingroup, potentially providing them with a means with which to achieve or maintain a positive social identity. Similarly, unfavourable misinformation may provide Labour voters with a way of positively differentiating their ingroup from the outgroup (e.g. Conservatives) and may therefore be perceived as relatively more acceptable to spread. However, notably, the patterns of moral judgements did not mirror differences in intentions to spread here. Indeed, while Labour voters were more likely to spread unfavourable misinformation, Conservative voters appeared no more likely to spread either type of misinformation (despite differences in moral evaluations).

It may therefore be the case that higher levels of moral acceptability help provide conditions within which people feel it is acceptable to amplify the spread of misinformation, however, this alone will not mean people will spread it. Indeed, previous work suggests that under conditions where there are no threats to identity, committed group members may engage in identity expression (Ellemers et al., 2002), however, they may only be motivated to do this when the group identity itself lacks distinctiveness (Spears et al., 2002). As such, Conservative voters may not have the same motivations to spread identity-affirming content as Labour voters may have had here. Indeed, from a social identity perspective, the fact that the Conservative party formed the UK Government at the time of data collection may have meant that for Labour voters the content depicted a "higher status" outgroup (at least in terms of political power). Those who strongly identified as Labour voters may have then been motivated to challenge any perceived hierarchy through social creativity (Tajfel & Turner, 2004). As such, misinformation that was unfavourable about the UK Government's handling of the pandemic may have provided Labour voters with a means with which to make expressions of positive distinctiveness. Study four in the present thesis explores the influence of identity threats and political orientation on evaluations of misinformation in more detail.

The present study does of course include several limitations. Firstly, self-reports were relied on to measure moral judgements which means that individuals must use a level of moral reasoning to make said judgement. However, it has been suggested that moral intuition lays an initial path for moral reasoning and therefore individuals' conclusions may be shaped by intuition (Haidt, 2001). It may also be the case that by inducing moral reasoning, participants' responses could have been shaped by their "spread" responses. However, the present study did not disclose any further information between the spread and the moral judgement stages (for example, disinformation status) and therefore it is less likely that any underlying basis of said judgement would change substantially. However, future studies exploring belief-consistency may wish to ask other moral questions about

misinformation items without asking participants how they themselves would "interact" to better understand if perceiving the act itself influences moral evaluations about an item content.

Where beliefs can be quantified as the probability of a certain likelihood being true (Huber, 2009) then they may shape moral evaluations of misinformation in a way that may lead to biased judgements. Where misinformation that conflicts with said beliefs may be readily detected as "wrong" to spread, there is a chance that people are simply less likely to detect potential moral violations in spreading belief-consistent misinformation. As such, people's moral evaluations of misinformation may be dynamic: influenced by the content itself as well as changes in external factors. As the present study illustrates, where people are more lenient towards misinformation, they may be more willing to spread it further. Indeed, as social cognitive theory proposes (Bandura, 1991b), individuals will self-regulate their behaviour but only up to a point. Developing interventions that may help to raise "moral thresholds" for spreading unverified content may therefore be a consideration for helping reduce the spread of disinformation within SMPs. This is explored in more detail in study five.

6.4.1. Conclusion

The present study sought to test a new scale to better represent users' contributions to the digital spread of misinformation. It also looked at how moral evaluations mediate the relationship between levels of belief-consistency and intentions to spread misinformation. A replication of study one was carried out, where participants were asked again about their levels of trust in the UK Government's handling of the COVID-19 pandemic, how likely it is they would interact with misinformation "favourable" and "unfavourable" towards the UK Government and how morally acceptable they felt it was to share said misinformation. The social media spread scale had acceptable reliability and, as in study one, "Trust" had a positive relationship with intentions to spread favourable misinformation, but a negative

relationship with spreading unfavourable misinformation. Moral judgements of misinformation also partially mediated the relationship between “Trust” and misinformation spread. The moral processes associated with intentions to spread misinformation are explored in more detail in study four, described in the next chapter.

Chapter 7. Study Four

7.1. Introduction

This chapter expands on the group-based findings from study two, applying the contribution to spread scale developed in the previous study. As the previous chapters indicate, moral judgements of spreading misinformation and disinformation may not be based solely on the act itself (e.g. the amplification of false information) but are influenced by factors such as belief consistency and how it may impact social identity. A major focus of this chapter is therefore to explain how social media content (in this case misinformation and disinformation) can present identity-relevant cues which influence digital behaviour via adjustments in moral judgements. Football fans were presented with identity-relevant content, allowing any ideological effects to be better distinguished from identity-level effects. Overall, the study presented here is intended to help identify the factors which influence moral evaluations of this content. Firstly, 'content' will be contextualised as potential identity-threats (either individual or group directed) which may motivate behaviour. Next, the impact of these threats on moral judgements is discussed, from both a modular (e.g. Moral Foundation Theory) and constructionist (e.g. Theory of Dyadic Morality) stance. The potential influence of ideology in these threat-induced moral judgements is then explored. The impact of message valence (i.e. positive or negative towards the ingroup) and inclusion of fact-check information (i.e. no information or a fact-check "tag" indicating the content is false) on contributions to spread and moral judgements are tested using ANCOVA. A moderated mediation demonstrates how these factors influence digital interactions with content via moral judgements. Exploratory analysis first looks at the role identity strength plays in spread intentions. Next, linguistic analysis using the extended Moral Foundations Dictionary helps to identify considerations underpinning moral judgements. Using conditional process analysis, it is then

demonstrated how ideology may impact these considerations. Finally, the influence of engaging in foundation-specific thinking is considered.

7.1.1. Identity Threats Within Social Media Content

Social media platforms (SMPs) are environments where users may express and experiment with their identity through content creation and digital interactions. However, self-expression, as well as a desire to present a positive image online, have also been linked to the sharing of misinformation (Apuke & Omar, 2021). Yet such acts of digital self-expression have an audience which, in turn, is likely to shape user-behaviour. Indeed, going against the group consensus within online environments can attract criticism and even exclusion (Bradshaw et al., 2021; Ditrich & Sassenberg, 2017). As such, SMPs can not only present situations that threaten identity at a group-level; there may also be situations that threaten their position within said group (e.g. through norm violations) or compromise personal moral standards in a manner that negatively impacts the self. Here it is proposed that contextualising the digital spread of harmful content (such as misinformation and disinformation) through the taxonomy of Identity Concerns and Self Motives (Ellemers et al., 2002) may help develop a stronger understanding of the way in which people interact (or refrain from interacting) with identity-related misinformation and disinformation.

Drawing from social identity theory (Tajfel & Turner, 2004) and self-categorisation theory (Turner & Reynolds, 2012), the taxonomy outlines three types of situations (no threat, individual-directed threat, group-directed threat) and proposes that identity concerns and motives within each situation will be determined by level of group commitment (Ellemers et al., 2002). When a person is presented with a post on social media that presents an ingroup in a positive light (e.g. suggests the ingroup's actions led to positive change) it may not present any clear threat to the value of the ingroup (e.g. in terms of status or morality). However, as the social context can also be a source of threat, aspects of

said post (and any engagement with it) may be seen as unacceptable to other ingroup members, who may be part of the user's potential audience. Therefore, both the post content and the SMP audience are factors that may induce threats to social identity.

When a post does not present a clear threat to identity then it is the level of group commitment that may determine whether a person engages in digital identity expression or not (Ellemers et al., 2002). Indeed, Ellemers et al. (2002) suggests that while those who have low group commitment may respond to non-threatening social stimuli with non-involvement. In contrast, high commitment to the ingroup may motivate people to express group identity in affirmative ways. This may include expressions of positive distinctiveness, endorsing group norms, and prosocial collective behaviour. Indeed, research suggests that strong group identifiers may be more likely to spread identity-relevant misinformation (Osmundsen et al., 2021). It may therefore be the case that such posts do not present a clear threat to a person's identity.

However, while some content may allow users to engage in identity expression generally (and may even benefit the group in other ways) if aspects of the post violate social norms (e.g. it is discriminatory or deceptive) it could be viewed as problematic by the user or their audience. As such, spreading the post further may constitute a norm-violation and, as such, may attract criticism or exclusion. For a strongly committed group member, an individual-directed threat such as potential exclusion may promote behavioural conformity (Ellemers et al., 2002). Indeed, research suggests people refrain from spreading disinformation due to reputational concerns (Altay et al., 2020), particularly if they use SMPs to engage in social comparison (Talwar et al., 2019). When spreading disinformation violates social norms (e.g. deceiving others) then social media users may refrain from openly interacting with it to protect their own identity.

However, drawing attention to the inaccuracy of identity-benefitting disinformation can also produce an individual-directed threat. Indeed, users may wish to intervene if they are aware that information shared by a fellow ingroup member is inaccurate. However,

situational factors may make the poster defensive towards ingroup criticism, such as when the group is under threat (Adelman & Dasgupta, 2019). They may also be more sensitive to ingroup criticism and feel more negatively about the source when their criticism is visible to outgroup members (Elder et al., 2005), which may include public social media pages. Awareness of a potential outgroup audience may then reduce the likelihood of strong-identifiers to refrain from speaking out (Packer, 2014). SMP users may therefore resort to less-visible actions to highlight the inaccuracy (such as anonymously reporting the post to the platform) or simply disengage from their moral standards to protect their group membership.

The final level of threat is group-directed, which includes threats to the value (e.g. status or morality) of a group (Ellemers et al., 2002). This can include information about the actions and behaviour of an ingroup that may have negative reputational impact. Such a threat may produce defensive reactions, such as discrediting the information (whether disinformation or not). Indeed, research suggests people are less likely to believe news that negatively frames an ingroup (Pereira et al., 2023). “Fake news” which threatens the ingroup has also been found to motivate users to react in a defensive manner to restore their identity (E. L. Cohen et al., 2020). Such responses have also been found to be more common in strong identifiers (Nauroth et al., 2015). Notably, what is illustrated by the taxonomy of identity concerns and self motives (Ellemers et al., 2002) is that the underlying concerns and motivations for not spreading group-threatening content and individual-threatening content differ. As such, it may be important to make these distinctions to better understand why users spread (or refrain from spreading) misinformation.

7.1.2. Identity Threats and Adjustments in Moral Judgement

The sharing of content which threatens identity (either at an individual or group level) will arguably be viewed less favourably than non-threatening content. As

demonstrated in study two, social media users may judge misinformation which undermines the ingroup as less acceptable to spread than when it supports an ingroup (or indeed, undermines an outgroup). These adjustments suggest that the moral judgements were not made regarding the act itself (e.g. is it right to spread misleading information generally, a deontological view[6]) but instead influenced by context, which may alter perceived consequences. However, there is ongoing theoretical debate surrounding the psychological processes underpinning moral judgements. Two theories will be discussed below. The first, Moral Foundations Theory proposes a modular approach to moral judgements (Graham et al., 2013) and has recently begun to be utilised in the context of disinformation research. However, despite its popularity, there is limited neurobiological evidence to support modularity. In contrast, constructionist theories of moral reasoning such as the Theory of Dyadic Morality (Schein & Gray, 2018) view moral reasoning as a more generalised and dynamic process. The impact of identity-directed threats on moral judgements will be discussed in the context of both of these theories.

As previously discussed in chapter 5, Moral Foundations Theory (MFT) proposes that moral judgements occur through distinct cognitive modules or structures (Graham et al., 2013). This includes modules relating to the following foundations: Care/ Harm, Fairness/ Cheating, Ingroup Loyalty/ Betrayal, Authority/ Subversion, Purity/ Disgust, Liberty/ Oppression[7] (Haidt & Joseph, 2008; R. Iyer et al., 2012). It is supposed that people tend to engage with certain foundations over others at a chronic level (measured by the Moral Foundations Questionnaire, MFQ), with research suggesting that such patterns of prioritisation may align with ideology (Graham et al., 2011). Notably, recent work has

---

[6] Deontology focuses on potential violation of proscriptive and prescriptive norms, while disregarding consequences that the particular situation may afford (Gawronski & Beer, 2017). Those taking a deontological stance may not only refrain from interacting with misinformation that amplifies spread, they may actively engage in attempting to reduce spread (regardless of level of agreement with its stance).

[7] The original moral foundations theory contained five "universal" foundations (e.g. care, fairness, ingroup loyalty, authority and purity), however, it was never supposed that there would only be five (Haidt & Joseph, 2008). A sixth foundation of liberty was subsequently introduced, along with additions to the original questionnaire (Iyer et al., 2012).

also linked engagement with "binding" foundations[8] with the "embracing" of disinformation (Ansani et al., 2021), lower acceptance of misinformation corrections (Trevors & Duffy, 2020), susceptibility to pseudoscience (Piejka & Okruszek, 2020) and increased bullshit receptivity (Nilsson et al., 2019). Additionally, people who prioritise loyalty foundations over fairness may also be less likely to report social media posts that violate community guidelines (Wilhelm et al., 2020). The modular approach proposed by MFT may therefore help explain the partisan asymmetries in moral judgements found within previous studies in this thesis (discussed in more detail in the next section).

Yet MFT does not suppose that individuals only engage with a fixed set of foundations. Rather, the baseline tendency to engage with certain foundations may be temporarily overridden by external factors (Tamborini, Prabhu, et al., 2018). For instance, the consumption of news about terrorist attacks has previously been shown to increase accessibility of binding foundations (Tamborini et al., 2017, 2020). Threats to the moral image of an ingroup may also increase accessibility of binding foundations, yet when outgroups carry out the same "immoral" act, accessibility of individualising foundations may increase (Leidner & Castano, 2012). This distinction may allow the behaviour of ingroup members to be rationalised as a "loyal" act (and therefore more moral), as opposed to an "unfair" or indeed "harmful" act. Such shifts in evaluations may then help to explain the ingroup biases in judgements seen in study two. When viewed through the lens of "fairness" or "harm" disinformation may appear unacceptable to spread. However, when evaluated in the context of "loyalty", identity-affirming disinformation may be perceived as more reasonable to spread when the status of the group is at stake.

[8] The main five moral foundations are often grouped into two higher-order categories, particularly when discussed in the context of ideology. These are 'binding' foundations (e.g. loyalty, authority, purity) and 'individualising' foundations (e.g. harm, fairness), with conservatives supposedly prioritising binding foundations more than liberals (Graham et al., 2011)

Indeed, shifts in foundation engagement can also influence subsequent decision making (Tamborini, Bowman, et al., 2018; Tamborini et al., 2017). While Waytz et al. (2013) found the level to which people engage with fairness over loyalty foundations influenced their likelihood of whistleblowing in a variety of scenarios, they also found priming fairness could encourage people to report unethical behaviour (e.g. uphold fairness). People may be more likely to behave in the context of the foundation (e.g. module) that is most accessible. Notably, research has also identified patterns between the use of foundation related language and behaviour within SMPs. Indeed, analysis of #HongKongPoliceBrutality tweets found that care/harm framed tweets spread further (through retweets and favourites), while tweets framed in relation to fairness or authority were less likely to be spread (R. Wang & Liu, 2021). Moreover, differences in moral framing have also been observed within SMP posts by anti-vaccination users compared to pro-vaccination users (Shi et al., 2021; Weinzierl & Harabagiu, 2022). As such, people's responses to misinformation may be influenced by the moral foundation they are presently engaging with (which in turn may be influenced by the misinformation itself) and the relevant evaluations made in the context of said foundation.

However, rather than identity-related threats activating judgements within a particular location (e.g. a modular perspective), it may be that threats simply amplify judgements via induced affect. Indeed, the Theory of Dyadic Morality (TDM), supposes moral evaluations are based on a combination of norm-violations, negative affect, and perceived levels of harm (Schein & Gray, 2018). Rather than discrete cognitive modules, TDM instead views moral "foundations" as representing categories of values against which perceived harm can be evaluated. Schein & Gray (2018) also argue that non-MFT values may also be moralised. These categories of values then help guide individual-level interpretations of norm violations; however it is ultimately the perceived harm that leads the norm-violation to become moralised (Schein & Gray, 2015). This helps to explain why,

for instance, judgements about incest or bestiality may be moralised, but judgements about people being boring or forgetful are not.

As such, both value-driven norm violations and perceptions of harm help guide the strength of the moral evaluation, in combination with negative affect (which can be either integral or incidental) (Schein & Gray, 2018). The act of spreading identity-threatening content such as disinformation may therefore be evaluated against these factors. Firstly, what kind of norms may intentionally be violated (if any). For instance, as disinformation involves the intentional sharing of false information it may be judged against a different set of norms than misinformation. Secondly, as TDM proposes that harm perceptions are also influenced by who receives the potential harm (e.g. "the patient"), threats to the self or ingroup may amplify perceptions of harm compared to threats to others. As such, any contextual changes within a post could produce moral judgements adjustments across a continuum.

7.1.3. Political Ideology and Moral Judgements

Studies one and two within this thesis have provided evidence of political asymmetry in moral evaluations of identity-related disinformation. Specifically, upon learning that content is false or misleading, Conservative voters (i.e. right-leaning) reported finding disinformation which may assist in achieving positive distinctiveness more morally acceptable to spread than Labour voters (i.e. left-leaning). As previously discussed, MFT research has found differences in how moral foundations are prioritised across the political spectrum (Graham et al., 2011). Although there is cultural and subcultural variation (Kivikangas et al., 2021), ideological differences in foundation prioritisation have been observed across a variety of countries including the United Kingdom (Graham et al., 2011),. Moreover, it has been argued that such differences are important for understanding political ideology (Haidt & Kesebir, 2010). Political liberals are thought to prioritise the individual in their moral considerations, with concerns

revolving around fair treatment and care of others (i.e. individualising foundations of care/harm and fairness). Conversely, political conservatives may be more likely to prioritise tight communities, focusing their moral concerns on factors such as protecting order, duties, and family (i.e. binding foundations of loyalty, authority, and purity[9]).

Given the associations that recent research has found between increased prioritisation with binding foundations and susceptibility to misinformation (Ansani et al., 2021; Lunz Trujillo et al., 2021; Piejka & Okruszek, 2020), it may be that the political asymmetry observed in the previous studies within this thesis can be explained by MFT. However, the proposed psychological basis of this relationship has attracted criticism. While ideology is heritable and generally stable, evidence suggests that moral foundations do not appear to be (K. B. Smith et al., 2017). Indeed, longitudinal research suggests incidents such as a terrorist attack or exposure to anti-immigration political campaigns can change chronic accessibility of foundations (Van de Vyver et al., 2016; Voelkel & Brandt, 2019). Arguably, the MFQ may not be measuring the moral processes that MFT claims to be. Smith et al. (2017) does, however, suggest that the MFQ may be useful for interpreting consequences of ideology (rather than ideology itself). For instance, the MFQ may be a better predictor of moral approval regarding issues including abortion, animal testing and the death penalty than ideology (Koleva et al., 2012). The MFQ may therefore at the very least indicate categories of (potentially ideology-related) values associated with susceptibility to spreading disinformation.

In contrast, TDM suggests any political asymmetry in moral judgements would be based on underlying differences in perceptions of what constitutes as "harm" (Schein & Gray, 2015, 2018). Rather than a discrete module, the influence of "foundations" instead act as a lens to evaluate harm and are context dependent. For instance, political liberals

[9] However, there is some suggestion that conservatives may prioritise these original five foundations equally (Graham et al., 2009).

have been shown to make judgements about same-sex marriage in relation to “fairness”, whereas for political conservatives it may be “sanctity” (Frimer, Skitka, et al., 2017). Yet, when making judgements about oil pipelines the opposite was true. Moreover, the connections between “foundations” and “harm” are also thought to differ between political conservatives and liberals. Indeed, Turner-Zwinkels et al. (2021) found that political liberals tend to mostly associate harm with fairness. However, political conservatives were found to associate any of the foundations with harm. Rather than ideological differences in how discrete cognitive modules are engaged with (as proposed by MFT), differences in moral evaluations may instead be influenced by how harm is understood.

7.1.4. The Present Study

As previously discussed, the way in which a post frames an ingroup (e.g. valence) will likely influence user-interactions in different ways. Posts that threaten the value of the ingroup image may generate fewer interactions which may amplify the spread of the content but may attract interactions intended to reduce its onward spread (e.g. reporting, downvoting, etc). As such, it is predicted that people are less likely to contribute to the wider spread of misinformation on social media when it negatively frames the ingroup. Therefore, Hypothesis 1 is:

> **H1.** Individuals will be more likely to spread content that is positive about their ingroup than content that is negative about their ingroup.

As demonstrated in study two, people may also make harsher moral judgements of posts after learning they are false or misleading. The present study seeks to assess whether real-world interventions produce similar effects. One strategy utilised by SMPs to tackle disinformation spread is the inclusion of fact-check “tags” which state that the information is false or misleading. Previous research has found these to reduce intentions to share by a

varying degree (Nekmat, 2020). One reason may be because the content is labelled for others to see and as such could present an individual-directed threat. It is therefore predicted that people will be less likely to spread content that has been deemed inaccurate by a fact-checker and perceive this content to be less acceptable to spread:

H2. Individuals will be less likely to spread content that displays a fact-check tag compared to content with no tag.

H3. Individuals will judge it to be less morally acceptable to spread content that displays a fact-check tag compared to content with no tag.

Study three also found that moral judgements helped to explain the relationship between the degree of belief consistency and intentions to spread misinformation. It is therefore also predicted that the relationship between the valence of a post (e.g. positive or negative about the ingroup) and the likelihood of spread will be partially explained by moral evaluations. Both the relationships between valence and moral evaluation, and valence and spread will also be weakened by the inclusion of a fact-check tag:

H4. The relationship between content valence and spread will be mediated by moral acceptability and moderated by the inclusion of a fact-check tag.

While the previous studies in the present thesis featured disinformation related to politics and therefore focused on group membership in the context of political parties, the present study seeks to understand if these relationships apply to group membership more broadly. By removing political cues (e.g. party or narratives related to political ideology) it may also be possible to understand whether the previously observed political asymmetry is context specific or an indication of potential differences in moral cognition. Specifically, this study recruited supporters of various football teams and presented them with identity-

relevant misinformation. Disinformation by definition can of course be utilised for financial gain (Digital Culture Media and Sport Committee, 2019) and may have either beneficial or detrimental implications for businesses. Football is no exception (Rojas Torrijos & Mello, 2021).

Exploratory analysis will also seek to further investigate the potential processes underpinning moral evaluations of identity-relevant misinformation. This notably includes the use of computational text analysis on participants' free text responses using the extended moral foundations dictionary. Moreover, individual factors such as strength of identity, political affiliation and scores on the moral foundation questionnaire may help to identify, or rule out, the influence of these factors on moral evaluations.

7.2. Method

7.2.1. Development of Stimuli and Pilot Study

Supporters of five football teams from the English Premier League were recruited for this study. To control for extraneous variables, original stimuli were developed as per Study two.

7.2.1.1 Materials

For consistency, a series of stimuli was developed to ensure participants were presented with stimuli personalised to their team and assigned condition. Therefore, twenty versions of the stimuli were created and tested in the pilot study. These had an overarching narrative of charity fundraising, which was felt to be plausible (both as a potentially real story and as disinformation). Football has long been associated with charity in England, where teams may be viewed as serving their local communities (Rosca, 2011). Yet this positive aspect has previously attracted disinformation regarding clubs or individuals making large donations (e.g. Dupuv, 2019). There are also instances of individuals associated with clubs committing theft and / or fraud (e.g. Kilpatrick, 2014). Therefore, the

stimuli for this study focused on the raising (i.e. “positive”) or stealing of charitable funds (i.e. “negative”) as manipulations for the “valence” condition.

As in study two, the text was developed for consistency across the four conditions, with small adjustments made for the valence manipulation (Figure 7.1). Draft statuses were created using Facebook’s custom background tool, a relatively common misinformation format within SMPs (e.g. Allen-Kinross, 2020) that also allowed for colour manipulation based on official team colours (e.g. red for Liverpool, blue for Chelsea, etc). Colours are a core part of a team’s identity that are utilised in fans’ own identity-expressions (Derbaix & Decrop, 2011), while incongruent colour presentations can also negatively influence judgements (Galli & Gorn, 2011). Each generated status was added to Photoshop templates replicating a standard Facebook post, with or without a fact-check “tag” identical to those used by Facebook at the time of the study. This was the basis for the second manipulation of “tag”.

Figure 7.1

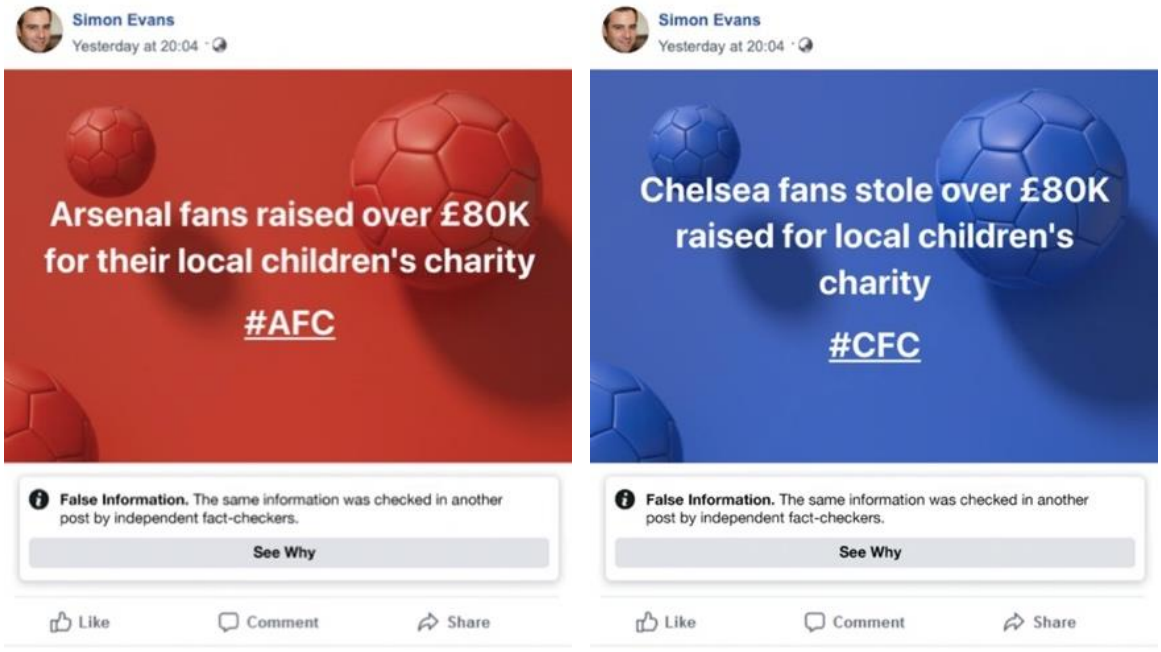
Examples of Study Four Stimuli by Condition

No Fact-check Tag (e.g. Misinformation)

Positive Valence *Negative Valence*

Fact-check Tag (e.g. Disinformation)

Positive Valence *Negative Valence*



Note. A total of 20 unique stimuli were created targeting five different English Premier League teams. Participants only viewed stimulus relating to own team.

7.2.1.2 Participants

20 participants (10 males) aged 21-66 (M = 37.45, SD = 10.26) were recruited through Prolific to take part in the pilot study. Ethical approval was obtained from the University's Psychology Ethics Committee (ETH2122-2442, Appendix U). Similar eligibility requirements were used to the main study for consistency. Participants were required to have a social media account, speak fluent English and currently be residing in the United Kingdom. They also had to identify as either an Arsenal (n = 4), Chelsea (n = 4), Liverpool (n = 4), Manchester United (n = 4) or Tottenham Hotspur (n = 4) fan.

7.2.1.3 Procedure

The study was hosted online using the survey platform Qualtrics. Participants answered basic demographic questions (gender, age, and location), as well as confirming their affiliated team. Next, participants were presented with team-consistent stimuli randomised across the four conditions. They were asked to rate how favourable the images were across a 7-point scale, from "Very unfavourable" to "Very favourable". They were then thanked and debriefed.

7.2.1.4 Results

Mean favourability scores for the items are displayed in Table 7.2. Scores below 4 indicate content that was rated as "unfavourable" while scores over 4, "favourable".

Table 7.1

Favourability Ratings of Disinformation

Item	N	Minimum	Maximum	Mean	*SD*
Positive – No Tag	20	5	7	6.60	0.60
Negative – No Tag	20	1	7	1.90	0.44
Positive – Fact-check Tag	20	2	7	5.55	0.37
Negative – Fact-check Tag	20	1	7	1.95	0.37

To establish whether items were perceived differently to their related pair, a series of paired *t*-tests were carried out. These showed that favourability scores of positively and

negatively valenced stimuli were significantly different when a fact-check tag was ($t(19) = 7.13$, $p < .001$, $d = 2.26$) and was not included ($t(19) = 9.87$, $p < .001$, $d = 2.13$). This suggests that the first manipulation effectively directed the valence of the content.

There were also differences between the favourability scores of fact-checked and non-fact-checked content for positive content ($t(19) = 2.58$, $p < .001$, $d = 1.82$) suggesting that the "tag" manipulation was effective[10].

7.2.1.5 Discussion

Participants judged social media posts framed positively about their ingroup as more favourable than negative posts. Fact-checked, positively-framed posts were also rated as less favourable than positive posts with no fact-check.

7.2.2. Main Study

7.2.2.1 Materials

7.2.2.1.1 Strength of Identity. (Postmes et al., 2013) single item measure of identity strength was used (Postmes et al., 2013). Participants were asked to state their level of agreement (1-*strongly disagree* to 7-*strongly agree*) with the statement "*I identify with being a …. Supporter"*, updated to reflect their reported team allegiance.

7.2.2.1.2 Political Orientation. A single item question was used to identify participants political orientation (PO), *"In politics people sometimes talk of 'left and 'right'. Where would you place yourself on this scale, where 0 means the left and 10 means the right?".*

7.2.2.1.3 Moral Foundations Questionnaire, Appendix V. The 30-item Moral Foundations Questionnaire (MFQ) was used to identify the moral foundations which participants most readily prioritise (Graham et al., 2011). Participants first rated relevance (*0-not at all relevant, 5-extremely relevant*) of 16 statements in deciding right or wrong

---

[10] The difference for negative content was not significant ($t(19) = -0.14$, $p = .45$). However, this is likely due to floor effects in both conditions.

(e.g. "*Whether or not someone showed a lack of respect for authority"*). They then rated levels of agreement (*0-strongly disagree, 5-strongly agree)* with 16 new statements (e.g. *"Respect for authority is something all children need to learn"*). Two catch questions are also included as attention checks (e.g. rating that maths ability as highly relevant in judging right and wrong).

The MFQ is thought to be generalisable to British samples (Graham et al., 2011). In the current data, alpha scores varied from $\alpha = 0.52$ (Fairness foundation) to $\alpha = 0.73$ for sanctity foundations. Lower alpha scores are a known limitation of using the MFQ and previous work has reported lower levels than these (Graham et al., 2009). Additionally, "binding" foundations of loyalty ($r = .41$***), authority ($r = .43$***) and sanctity ($r = .29$***) positively correlated with PO, while the 'individualising' foundation of fairness was negatively correlated with PO[11] ($r = -.35$***). This corresponds with previously reported ideological differences across the MFQ (Graham et al., 2011).

7.2.2.1.4 MFQ Liberty Items, Appendix W. Libertarians are thought to have moral concerns that are different to conservatives and liberals (Iyer et al., 2012). Nine additional items developed by Iyer et al. (2012) relating to either economic / government or lifestyle aspects of libertarianism were presented alongside the MFQ-30 items. The economic / government liberty subscale had an alpha score of $\alpha = 0.53$. However, alpha is influenced in this calculation by the number of items per scale, where fewer items may lead to lower scores. As the subscale had only six items, and $\alpha < .50$ is still felt to indicate moderately reliability (Hinton et al., 2014), the decision was taken to retain the subscale. However, the decision was taken to drop the subscale of lifestyle liberty due to its very low reliability, $\alpha = 0.40$.

[11] The other individualising foundation 'care / harm' did not significantly correlate with PO ($r = -.11$)

7.2.2.1.5 Social Media Spread Scale. The scale developed in study three was used again, but with the addition of a "downvote / not interested" question to reflect changes made by social media platforms. The scale had acceptable reliability across the four conditions, where the lowest alpha was $\alpha = 0.69$.

7.2.2.1.6 Moral Judgements of Spreading. As with previous studies in this thesis, participants were asked to rate how morally acceptable it would be for them to share the content, rated on an 11-point scale *(0 –not at all morally acceptable, 10 - completely morally acceptable).*

7.2.2.1.7 Extended Moral Foundations Dictionary. Participants were asked for free-text responses to explain their spread and moral related answers to be analysed using the extended Moral Foundation Dictionary (eMFD) (F. R. Hopp et al., 2021). This is one of at least three dictionaries that can be used to identify the presence of moral domains (e.g. foundations) in language: the original "Moral Foundations Dictionary" (MFD) (Graham et al., 2009), the "MFD2" (Frimer, Haidt, et al., 2017), as well as the eMFD. The eMFD was selected due to its much larger pool of words (3,270) and supposed improved external validity. Notably, the eMFD was developed through crowdsourcing, where probability scores were created based on domain allocations by 557 annotators (whereas other dictionaries use discrete scores, based on allocations made by small groups of students and / or experts).

Responses were first pre-processed by (1) removing punctuation, (2) merging contractions, (3) spell checked and (4) changed to lower-case. (5) Stop words[12] were removed based on Python's Natural Language Toolkit. (6) Stem words were condensed before any (7) numbers and (8) domain specific phrases were removed. The text was then compared against the eMFD using the *eMFDscore* Python package to produce individual

[12] Stop words are commonly used words that lack relevance in natural language processing (e.g. 'a', 'the', 'is') and are removed so that scores can be calculated in proportion to the remaining words.

scores for each domain. As words in the eMFD may cross domains, scores were allocated only to the domain with the highest weighting (e.g. where a word scores highest on "fairness", but has low probability of inclusion in other domains, the score is allocated to "fairness" only) as advised by Hopp et al. (2021). Probability scores for each moral domain (e.g. the sum of the probability scores associated with each word divided by the response word count) were then added into the main dataset.

The eMFD has previously been used to identify moral cues in text that predict sharing behaviour online (F. R. Hopp et al., 2021) and therefore is relevant for the present study. Notably, Hopp et al. (2021) have previously advised against this approach for smaller passages of text (such as tweets) which are less likely to use moral language. However, as participants were asked a question that specifically related to morality, this should not be a concern.

7.2.2.2 Procedure

Recruitment took place on Prolific and the study was hosted on Qualtrics. Participants were first presented with the invitation letter and consent form, and then asked to confirm their device[13] and the football team they support[14]. Demographic information was collected (including political orientation), as well as strength of identity with aforementioned team. Participants then completed both parts of the MFQ-30 and additional Liberty questions. They were then randomly assigned to one of four conditions, where they were presented with a positive or negatively framed post about their team, and either contained a fact-check tag or did not. The same image was presented to participants throughout the study.

Participants rated their likelihood of contributing to the image's spread using the social media spread scale (study three) and asked to explain their response in a free-text

---

[13] Participants were required to access the study using a computer due to the free-text task.
[14] To ensure that current fans were recruited, participants had to select which of the English Premier League teams they supported. This had to align with their participant information provided to Prolific.

box. Next, they were asked to rate how morally acceptable it would be for them to share the image on social media and again asked to explain their response in a free-text box. Moral acceptability was measured on an 11-point scale, where a score of '0' indicated that it was not at all morally acceptable whereas '10' would be completely morally acceptable. Participants were then thanked and debriefed.

7.2.2.3 Participants

262 participants (141 males) aged 18-77 (M = 40.06, SD = 13.65) were recruited through Prolific to take part in the study. Ethical approval was obtained from the University's Psychology Ethics Committee (ETH2122-2442). To ensure enough power for the moderated mediation analyses (H4), sample size planning was first conducted using MedPower (Kenny, 2017). A minimum effect size of β = .2 is thought to be the smallest that would be practically significant in social science research (Ferguson, 2009). To detect β = .2 at 80% power, 250 participants would be required.

However, this tool is designed for mediation analysis planning specifically. Therefore, to accommodate the inclusion of moderator variables, this proposed sample size was confirmed using G*Power by planning for a linear multiple regression. To reach a minimum of r^2 =.04 with five predictors, 191 participants would be needed for 80% power, suggesting a sample of 250 would be acceptable for a moderated mediation analysis. For H1-H3, to detect η_p^2 = .04, 191 participants would also be needed to achieve 80% power. Allowing for data screening exclusions, the target sample size was 280 participants.

For this study, participants were required to have an active social media account (e.g. Facebook, Instagram, etc) and must not have taken part in the pilot or studies one-three. Participants had to identify as a supporter of one of five Premier League teams located across England (Table 3) which had the highest pool of participants available on Prolific. They also had to be located within the United Kingdom and speak fluent English.

Gender and political orientation were balanced where possible during recruitment using Prolific's pre-screening tools.

A total of 15 participants attempted to use an incompatible device (either identified by Qualtrics or self-report), 22 participants chose a different team to their Prolific account, and one participant did not consent. These participants were automatically prevented from proceeding with the study. Qualtrics's proprietary software flagged 10 participants as potentially fraudulent and therefore their responses were removed. Another four dropped out of the study before answering dependent variable questions, and so were also removed.

Two participants assigned exclusively high scores (e.g. 7 and above) for the "spread" responses and so these participants were deemed to be inauthentic (e.g. reported high likelihood of both spread contributing and reducing actions) and so were removed[15]. Five participants failed the MFQ catch questions and therefore their MFQ scores were removed. Participant demographics for the final sample are shown in Table 7.2.

[15] The removal of these participants did not affect planned tests in relation to reaching significance. A corresponding full set of results including these excluded participants are included in Appendix X.

Table 7.2

Participant Demographics for Study Four

	All		Arsenal F.C.		Chelsea F.C.		Liverpool F.C.		Manchester United F.C.		Tottenham Hotspur F.C.	
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Total	262	100	51	19.50	47	17.90	55	21.00	56	21.40	53	20.20
Age	40.06 (13.65)		37.33 (12.64)		37.38 (13.36)		42.35 (12.96)		41.45 (15.17)		41.23 (13.54)	
Gender												
Female	120	45.80	27	52.90	22	46.80	27	49.10	24	42.90	20	37.70
Male	141	53.80	24	47.10	25	53.20	28	50.90	32	57.10	32	60.40
Non-binary	1	0.40	0	0.00	0	0.00	0	0.00	0	0.00	1	1.90
Education completed												
Less than GCSE's	2	0.80	1	2.00	1	2.10	0	0.00	0	0.00	0	0.00
GCSE's	33	12.60	4	7.80	6	12.80	9	16.40	7	12.50	7	13.20
A-Level's	64	24.40	15	29.40	10	21.30	13	23.60	17	30.40	9	17.00
Bachelor's Degree	126	48.10	26	51.00	23	48.90	19	34.50	26	46.40	32	60.40
Master's Degree	30	11.50	3	5.90	6	12.80	11	20.00	6	10.70	4	7.50
Doctoral Degree	6	2.30	2	3.90	1	2.10	2	3.60	0	0.00	1	1.90
Other	1	0.40	0	0.00	0	0.00	1	1.80	0	0.00	0	0.00

7.2.2.4 Data Analysis

Data analysis for planned tests was pre-registered through AsPredicted.org (#96907, Appendix Y). The planned tests in the present study used 2x2 factorial analysis of covariance (ANCOVA) to test H1-H3. "Valence" ("positive" vs "negative") and "fact-check tag" ("tag" vs "no tag") were used as between-group factors, with age and gender entered as controls. The first ANCOVA had a DV of likelihood of contributing to the "spread" of disinformation (H1 & H2), while the second had a DV of "moral acceptability" (H3). H4 was tested using a moderated mediation analysis. "Moral acceptability" was included as the mediator (*M*) between "valence" (*X*) and "spread" (*Y*). "Fact-check tag" was included as the moderator (*W*). All noted tests used α levels of .05 unless otherwise specified. This was followed by exploratory analysis, utilising MANCOVA, mediation and conditional processes analyses.

7.3. Results

Descriptive statistics are included in Table 7.3. There was some skewness and kurtosis in the strength of identity variable. There is also some negative skewness in the moral acceptability scores in the positive (no tag) condition. It is, however, thought that any risks associated with skewness and kurtosis are reduced with a large sample (Tabachnick & Fidell, 2013). Histograms for all variables can be found in Appendix Z.

Table 7.3

Summary of Descriptive Statistics

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skewness	Kurtosis
Age	262	40.06	13.65			18.00 - 77.00	0.48	−0.47
Strength of Identity	262	5.99	1.16		1 - 7	1.00 - 7.00	−1.67	3.78
Political Alignment	262	5.62	2.27		1 - 11	1.00 - 11.00	−0.05	−0.39
Positive (No Tag)								
Spread	66	7.77	1.39	.74	1 - 11	4.67 - 11.00	0.08	−0.51
Moral Acceptability	66	10.21	1.28		1 - 11	6.00 - 11.00	−1.72	2.38

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skewness	Kurtosis
Negative (No Tag)								
Spread	66	5.52	1.51	.69	1 - 11	1.67 - 9.56	0.24	0.85
Moral Acceptability	66	4.14	3.38		1 - 11	1.00 - 11.00	0.80	0.56
Positive (FC Tag)								
Spread	66	6.44	1.91	.80	1 - 11	1.00 - 10.78	−0.10	0.28
Moral Acceptability	66	5.91	3.52		1 - 11	1.00 - 11.00	0.06	−1.35
Negative (FC Tag)								
Spread	63	4.74	1.63	.71	1 - 11	1.00 - 9.67	−0.17	0.66
Moral Acceptability	63	3.03	2.49		1 - 11	1.00 - 11.00	1.38	1.58
MFQ - 5-Factor (30-item scale)								
Harm	262	4.64	0.69	.60	1 - 6	2.17 - 6.00	−0.66	0.55
Fairness	262	4.56	0.61	.52	1 - 6	2.33 - 6.00	−0.34	0.41
Loyalty	262	3.34	0.81	.68	1 - 6	1.00 - 5.83	−0.03	0.03
Authority	262	3.77	0.81	.70	1 - 6	1.33 - 5.83	−0.37	−0.32
Sanctity	262	3.39	0.87	.73	1 - 6	1.00 - 5.83	0.20	−0.02
MFQ – Liberty items								
Economic / Government	262	3.98	0.68	.53	1 - 6	2.00 - 5.83	−0.18	0.26

7.3.1. Planned Tests

7.3.1.1 Likelihood of Contributing to "Spread"

To test H1 and H2, a 2x2 between-groups analysis of covariance (ANCOVA) was carried out. This looked at the impact of message "valence" (i.e. "positive" or "negative" towards the ingroup) and the inclusion of a "tag" stating the information was false or misleading (i.e. "no tag" or "fact-check tag") on the likelihood of spread. Covariates of age and gender were included as controls. There was homogeneity of variance as assessed by Levene's test (p = .16). Visual inspections of histograms suggest the data are normally distributed (Appendix AA). Inspection of boxplots revealed eight outliers, however,

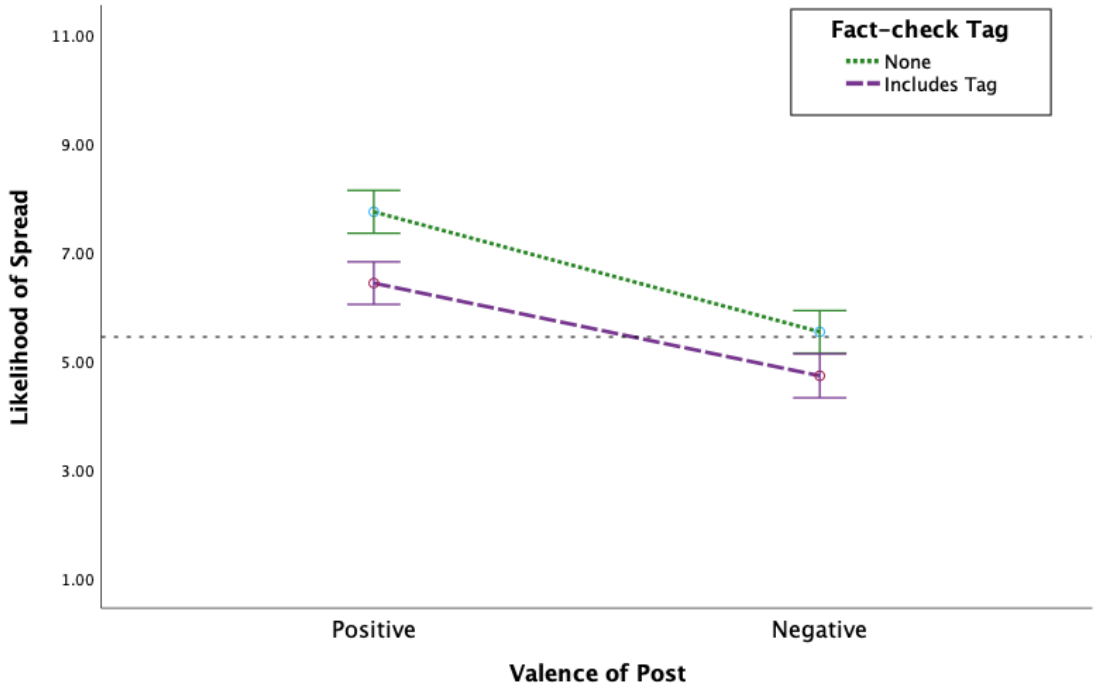
removal of these outliers does not change the significance of effects[16]. Neither age (F(1, 253) = 1.64, p = .20, η^2_p = .01) nor gender (F(1, 253) = 1.27, p = .26, η^2_p = .01) were significant covariates.

The ANCOVA showed a significant main effect of "valence" with a large effect size (J. Cohen, 1992), F(1, 253) = 94.61, p < .001, η^2_p = .27. Overall, positive posts about the ingroup were more likely to be spread than negative posts about the ingroup. H1 is therefore accepted. Furthermore, the main effect of "tag" was significant with a medium effect size, F(1, 253) = 27.88, p < .001, η^2_p = .11. Therefore H2 was also accepted. As there was no significant interaction effect between "valence" and "tag" (F(1, 253) = 1.57, p = .21, η^2_p = .01) this suggests that viewing a fact-check tag reduced intentions to contribute to spread, but did not reduce ingroup bias (Figure 7.2). Pairwise comparisons confirmed that the differences between all pairs of conditions were significant (Appendix BB). Notably, participants were still more willing to amplify the spread of positive disinformation (6.43 ± 0.20, M ± SE) compared to negative disinformation (4.73 ± 0.21) despite both being marked as false, 1.71 (95% CI, 1.14 to 2.27), p < .001.

[16] Upon the removal of outliers the main effects of "valence" (F(1, 245) = 114.30, p < .001, η^2_p = .32) and "tag" (F(1, 245) = 34.99, p < .001, η^2_p = .13) on spread were still significant. The interaction effect was again not significant, F(1, 245) = 0.44, p = .51, η_p^2 = .002.

Figure 7.2

Estimated Marginal Means of Spread by Post Valence and Tag Inclusion

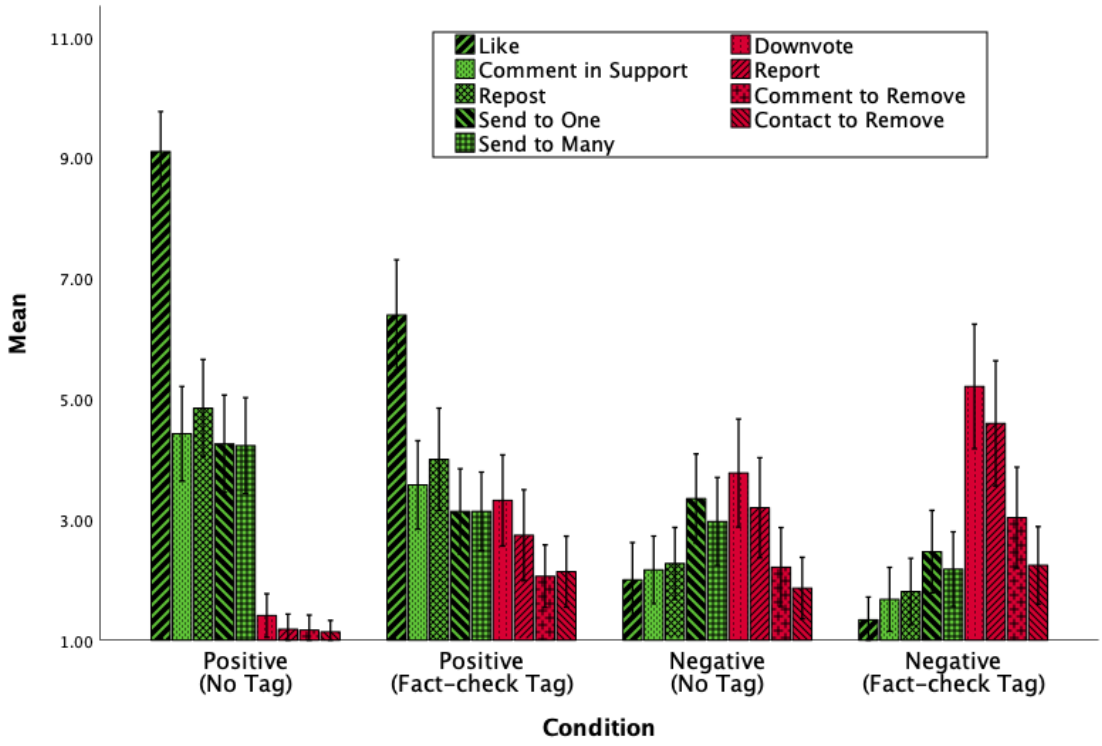


Note. Controlled for age and gender. Error bars indicate 95% confidence interval. Dashed line indicates point at which participants with no intentions to interact in any manner (e.g. amplify or intervene) would fall.

As illustrated in Figure 7.3, greater intentions to spread reflected increased engagement with "amplifying" actions and were more commonly associated with positive post conditions. However, it appears that threats (either group or individual-directed) may influence intentions to engage in spread-reducing actions.

Figure 7.3

Likelihood of Engaging in Specific Digital Interaction by Valence and Tag Inclusion



Note. Green bars (1-5) indicate behaviours intended to add to spread. Red bars (6-9) indicate behaviours intended to reduce spread. Error bars indicate 95% confidence interval.

7.3.1.2 Moral Judgements of Spreading the Post

Next, another 2x2 ANCOVA was run to ascertain if moral judgements are influenced by inclusion of a fact-check tag (H3). The factors were again "valence" and "tag", with covariates of age and gender, but the dependent variable was "moral acceptability". Levene's test was significant ($p < .001$) and therefore a significance level of .01 will be applied here. Visual inspections of histograms suggest that the data was not normally distributed (Appendix CC), however, ANOVAs are thought to be robust to violations of this assumption (Tabachnick & Fidell, 2013). Box-plots revealed nine

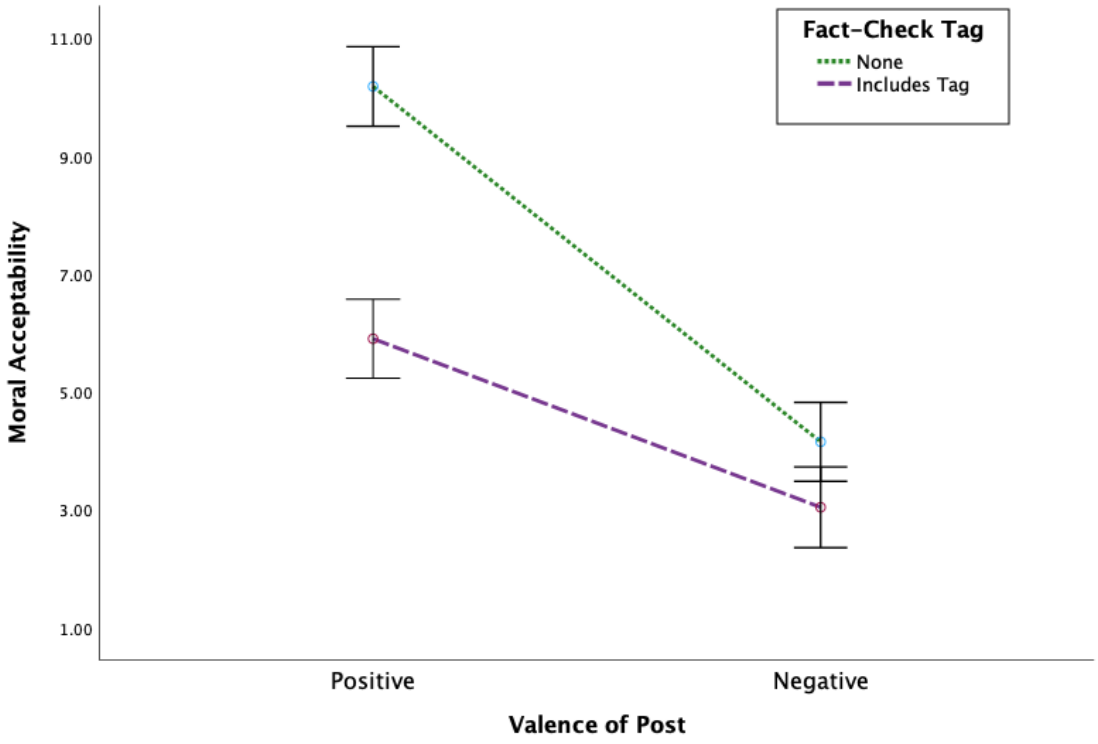
outliers, but removal again made no difference to the significance of effects[17]. Age significantly contributed to the model ($F(1, 254) = 13.10$, $p < .001$, $\eta^2_p = .05$), but removing this covariate did not impact results (Appendix DD).

Main effects suggested that fact-check "tags" led content to be judged as less acceptable to spread further, $F(1, 254) = 61.95$, $p < .001$, $\eta^2_p = .20$. H3 is therefore accepted. Furthermore, the main effect of "valence" was also significant ($F(1, 254) = 167.93$, $p < .001$, $\eta^2_p = .40$), suggesting greater acceptance of spreading content that supports the ingroup. The interaction effect between "valence" and "tag" was also significant, $F(1, 254) = 21.42$, $p < .001$, $\eta^2_p = .08$. As the data was not normally distributed, non-parametric tests were used to support these findings. Mann-Whitney U tests confirmed that an ingroup bias occurred whether fact-check "tags" were included ($U = 1091.50$, $z = -4.72$, $p < .001$) or not ($U = 366.00$, $z = -8.50$, $p < .001$). Yet, while a fact-check "tag" does begin to reduce this bias, people were notably more accepting of spreading disinformation (e.g. includes "tag") that is positive about the ingroup than spread negative misinformation (e.g. no "tag"), $U = 1525$, $z = -3.01$, $p = .003$ (Figure 7.4).

[17] Upon the removal of outliers the main effects of "valence" ($F(1, 245) = 198.62$, $p < .001$, $\eta^2_p = .45$) and "tag" ($F(1, 245) = 79.59$, $p < .001$, $\eta^2_p = .25$) on moral judgements were still significant. The interaction effect was also still significant, $F(1, 245) = 23.62$, $p < .001$, $\eta_p^2 = .09$.

Figure 7.4

Estimated Marginal Means of Moral Judgements by Valence and Tag Inclusion



Note. Controlled for age and gender. Error bars indicate 95% confidence interval

7.3.1.3 Conditional Effects on Moral Judgements and Intentions to Spread

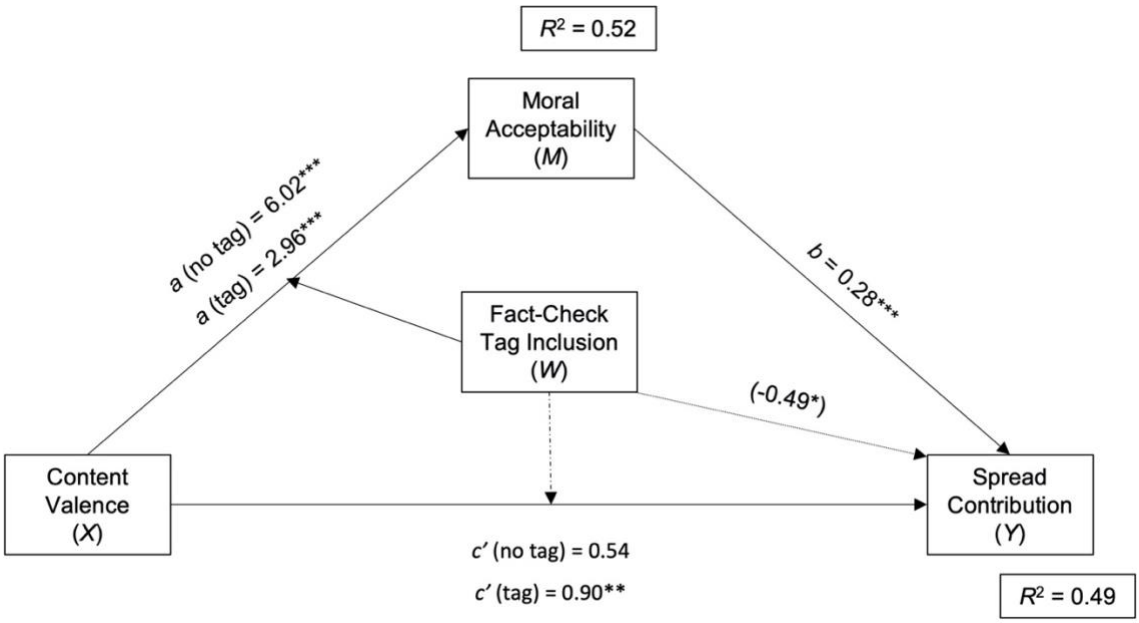
A moderated mediation (PROCESS model 8) was carried out to better understand how post valence and fact-check tags influence spread through adjustments in moral judgements (H4). This model tests for moderated mediation, which is a form of conditional process analysis where the sizes of the indirect and / or direct effects are dependent on the level of the moderator (*W)* (A. F. Hayes, 2017). Inclusion of a fact-check tag moderated the paths between content valence and moral judgements (e.g. *X→M),* and valence and spread (e.g. *X→Y*). Assumptions for regression were checked, with no violations observed (Appendix EE). As the regressions showed age was a significant predictor of moral judgements it was included in the model as a control. Bootstrapping was set to 5,000 and

heteroscedasticity-consistent inference was set to HC4. Effects are unstandardised unless stated otherwise.

Valence (*X*) had a weaker effect on moral judgments (*M*) when a fact-check was present compared to having no fact-check (*W*), $B = -3.06$, $B_{SE} = .69$, $t(254) = -4.47$, $p < .001$. Higher moral acceptability was then associated with increased likelihood of spreading content (*Y*), $B = 0.28$, $B_{SE} = .03$, $t(253) = 8.84$, $p < .001$. The index of moderated mediation was significant = −0.86 (95% CI = −1.30, −0.47). This suggests that the fact-check tags (*W*) did influence the strength of effects. Specifically, the conditional indirect effect (*ab*) through moral judgement was strongest when there was no fact-check ($B = 1.68$, $B_{SE} = 0.23$, 95% CI = 1.25, 2.17), which led to a non-significant direct effect (*c'*) ($B = 0.55$, $B_{SE} = .28$, $t(253) = 1.96$, $p < .05$). Therefore, differences in the spread of positive and negative misinformation may be explained by moral judgements of said spread (Figure 7.5).

Figure 7.5

Conditional Indirect Effects of Content Valence and Intentions to Spread via Moral Judgement, With and Without a Fact-Check Tag



Note. Unstandardised values shown. Controlled for age.

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

However, when a fact-check was present, moral judgements only partially explained the relationship between valence and spread (B = 0.83, B_{SE} = 0.18, 95% CI = 0.49, 1.19) as the direct effect (*c*') remained significant at this level (B = 0.90, B_{SE} = .28, $t(253) = 3.22$, $p = .002$). There was no evidence to suggest this conditional effect was due to an interaction ("tag" x "valence") in the *c'* path (B = 0.35, B_{SE} = .37, $t(253) = 0.95$, p = .34). "Tag" was, however, a direct predictor of spread in the *c'* path ($B = -0.49$, B_{SE} = .25, $t(253) = -1.97$, $p < .05$), with a slightly lower likelihood of spread for fact-checked content after accounting for indirect effects (Appendix FF). Therefore, ingroup biases in spread contribution of known disinformation are mostly (but not entirely) explained by moral judgements (Table 7.4).

Table 7.4

Ordinary Least Squares Regression Coefficients (With Standard Errors) From a First Stage Moderated Mediation Model Predicting Likelihood of Contributing to Spread

		Outcome		
		M: Moral Judgement		*Y:* Contribution to Spread
Constant		5.82*** (0.64)		4.37***(0.38)
X: Valence	$a_1 \rightarrow$	6.02*** (0.44)	$c' \rightarrow$	0.55 (0.28)
W: Tag	$a_2 \rightarrow$	−1.15* (0.52)	$b_2 \rightarrow$	−0.49* (0.25)
XW: Valence x Tag	$a_3 \rightarrow$	−3.06*** (0.68)	$b_3 \rightarrow$	0.35 (0.37)
Age		−0.04*** (0.01)		−0.0002 (0.01)
M: Moral Judgement			$b_1 \rightarrow$	0.28*** (0.03)
	R	0.72		0.70
	R^2	0.52		0.49
			Index	95% bootstrap CI[a]
Moderated mediation			−0.86	−1.30, −0.47

Note. Valence (0 = negative, 1 = positive) and Tag (0 = no tag, 1 = fact-check tag) coded as dummy variables.

[a] Percentile bootstrap CI based on 5,000 bootstrap samples.

7.3.2. Exploratory Analysis

7.3.2.1 Strength of Identity and the spread of misinformation

To clarify whether strength of identity (SOI) was related to contributions of spread, a series of Spearman's Rho correlations were run. Due to the level of measurement of SOI (i.e. ordinal) a non-parametric test was used. Stronger SOI was associated with an increased likelihood of spreading positive misinformation (i.e. no fact-check), $r_s(64) = .35$, $p < .01$. No other relationships were significant (Table 7.5).

Table 7.5

Spearman's Correlations of Moral Foundations, Spread and Moral Judgements by Condition

						Positive				Negative			
				All Conditions		(No tag)		(FC Tag)		(No tag)		(FC Tag)	
Variable	*M*	*SD*	*α*	Spr	MJ	Spr	MJ	Spr	MJ	Spr	MJ	Spr	MJ
Intentions to Spread (Spr)					.70***		.14		.66***		.51***		.43***
Moral Judgement (MJ)				.70***		.14		.66***		.51***		.43***	
Age	40.06	13.65		-.12*	-.18**	-.08	-.15	-.09	-.31*	-.21	-.30*	-.06	-.10
Strength of Identity (SOI)	6.02	1.08		.04	.02	.35**	.15	.13	-.13	-.23	-.08	-.19	.04
Political Orientation (PO)	5.66	2.26		.03	.05	-.18	-.06	.19	.27*	.08	-.08	-.13	.01
MFQ													
Care / Harm	4.66	0.69	.60	-.04	-.10	.11	.15	.01	-.24	-.07	-.19	-.11	.06
Fairness	4.54	0.60	.52	.03	-.02	.04	.15	-.04	-.28[a]	-.12	-.08	.26[a]	.09
Loyalty	3.34	0.80	.68	.03	-.05	.10	.02	.26[a]	.21	.002	-.30[a]	-.13	-.03
Authority	3.78	0.78	.70	.003	-.05	.004	-.13	.10	.07	-.004	-.27[a]	-.06	-.01
Sanctity	3.38	0.87	.73	.12	.01	.26[a]	-.13	.05	.05	-.10	-.30[a]	.18	.16
Liberty (Ec / Gov)	4.00	0.68	.53	.13[a]	.02	.30[a]	.22[a]	.07	-.02	.17	-.13	-.01	.06
eMFD domain scores													
Care / Harm	0.06	0.05		.33***	.32***	.10	.16	.44***	.31[b]	.17	.26[b]	.32*	.15
Fairness	0.13	0.09		-.38***	-.45***	-.14	-.11	-.43***	-.50***	.08	-.16	-.12	-.21
Loyalty	0.06	0.05		.09	.16*	-.05	.11	.18	.20	-.06	.04	.003	.06
Authority	0.04	0.05		.10	.04	.16	.01	.21	.26[b]	-.14	-.35**	.03	.05
Sanctity	0.02	0.03		-.04	-.02	-.08	-.06	-.30[b]	-.15	.17	.13	-.09	-.28[b]

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

[a] *p*-value less than .05 but not within the threshold set out by holms-bonferroni calculation for MFQ score comparisons (Appendix GG)

[b] *p*-value less than .05 but not within the threshold set out by holms-bonferroni calculation for eMFD score comparisons

7.3.2.2 Computational Text Analysis – Changes to Moral Domain

The eMFD was used to analyse participants' written responses regarding spread and moral judgements. Moral domain scores (MDP) were computed to estimate the probability that participants were engaging with a particular moral foundation within each condition. A two-way MANCOVA was run using the five moral domain scores (Care, Fairness, Loyalty, Authority, Sanctity) as dependent variables (DV) and between group factors of "tag" and "valence". Age and gender were included as control variables. Box's test indicated the assumption of homogeneity of covariance matrices was violated ($p <$ .001). Box's test is known to be highly sensitive for large samples and therefore the more robust Pillai's Trace will be reported (Tabachnick & Fidell, 2013). Three dependent variables had homogeneity of variance as assessed by Levene's test, however, "Fairness" and "Authority" did not. Adjusted alpha scores ($p < .025$) are therefore applied for these variables (Tabachnick & Fidell, 2013).

Overall, the effects of "valence" ($V = .06$, $F(5, 230) = 3.14$, $p < .01$) and "tag" ($V = .11$, $F(5, 230) = 5.88$, $p < .001$) were significant (Table 7.6), but the interaction effect ("valence" x "tag") was not ($V = .01$, $F(5, 230) = 0.63$, $p = 0.68$). Individual DVs were analysed using Bonferroni adjusted alpha levels ($p < .005$). Both "valence" ($F(1, 234) = 13.86$, $p < .001$, $\eta^2_p = .06$) and "tag" ($F(1, 234) = 26.62$, $p < .001$, $\eta^2_p = .10$) were significant for "Fairness" scores only (Figure 7.6).

Table 7.6

MANCOVA Between-Subjects Effects of Moral Domain Scores by Valence and Tag

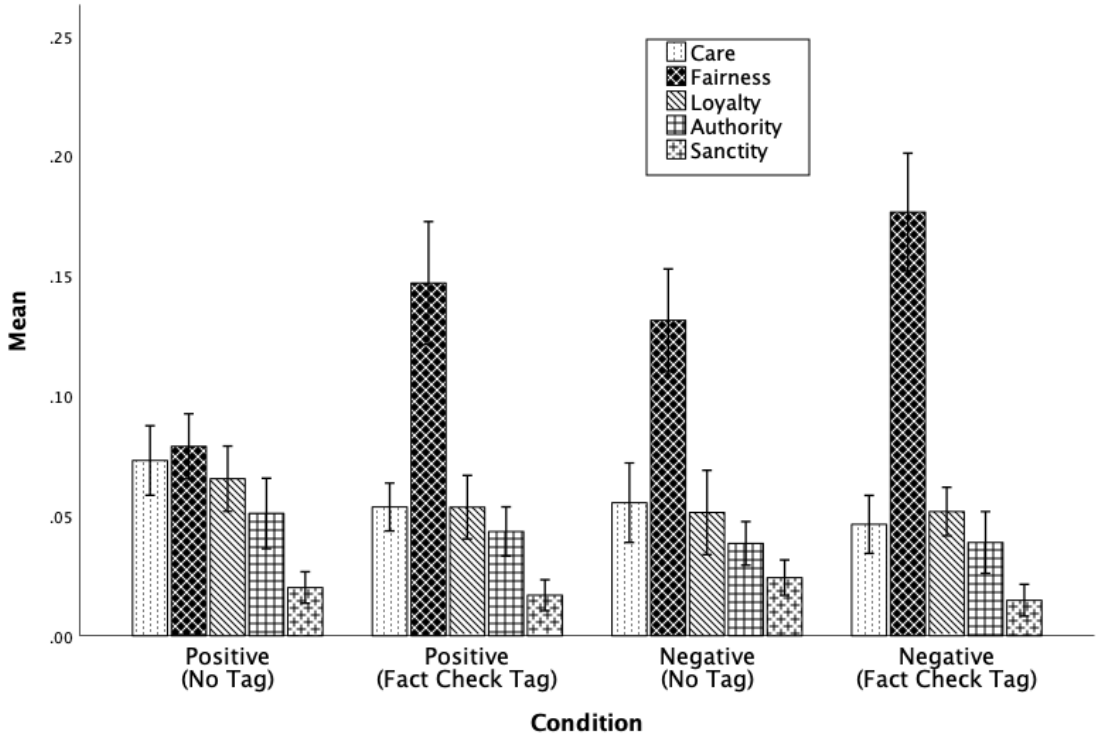
	Valence	Tag	Valence x Tag
Care	2.94	4.07[a]	0.49
Fairness	13.86***	26.62***	1.08
Loyalty	1.25	0.60	0.84
Authority	2.05	0.37	0.43
Sanctity	0.06	3.48	0.87

Note. [a] p-value less than .05 ($p = .049$) but not within the Bonferroni adjustment ($p < .005$).

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

Figure 7.6

Mean Probability of Engaging with a Specific Moral Domain by Valence and Tag



Note. Error bars indicate 95% confidence interval

Pairwise comparisons suggest that participants who saw a positive post without a fact-check were less likely to engage with "fairness" (0.08 ± 0.01, M ± SE) than those saw a fact-checked, positive post (0.15 ± 0.01), and this difference was statistically significant, 0.07 (95% CI, 0.04 to 0.10), $p < .001$. Therefore, fact-check tags may help prompt considerations of "fairness". Interestingly, participants were also more likely to consider fairness when evaluating a negative post without a fact-check (0.13 ± 0.01) compared to a positive, untagged post (0.08 ± 0.01, M ± SE), 0.05 (95% CI, 0.02 to 0.08), $p < .001$. Social media users may therefore be less likely to consider fairness when evaluating misinformation that neither threatens the individual or the group.

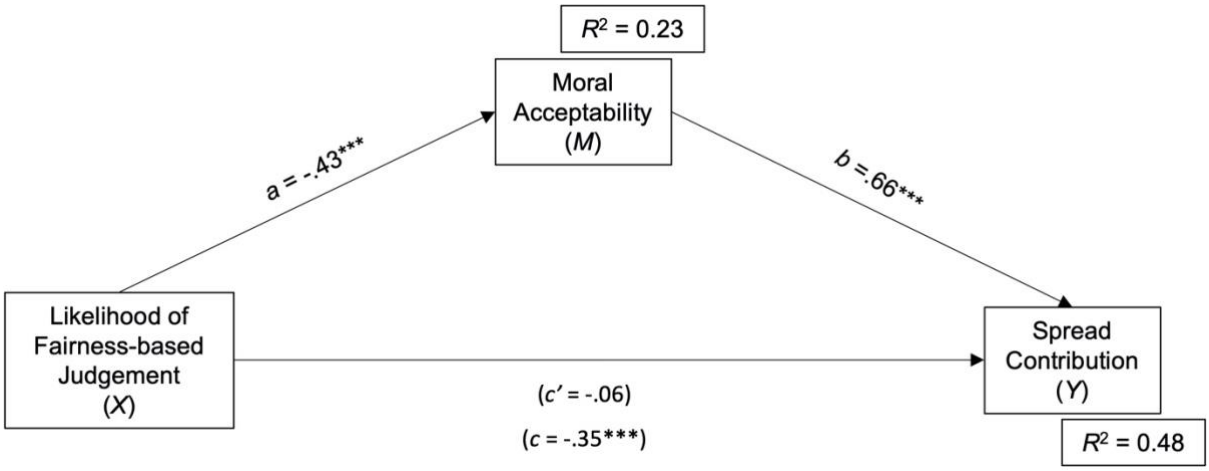
7.3.2.3 Fairness as a predictor of spread

Next, a mediation analysis was run to explore the role of fairness considerations in contributions to spread. PROCESS model 4 was used, where moral judgements (*M*)

mediated the relationship between fairness domain scores (*X*) and likelihood of spread (*Y*). Assumptions for a regression were met, with age and gender included as control variables. As illustrated in Figure 7.7, those who had an increased probability of having considered fairness[18] also judged the content as less acceptable to spread ($B = -19.01$, $B_{SE} = 2.11$, $t(234) = -9.01$, $p < .001$). As before, moral acceptability was positively related to higher contributions to spread ($B = 0.33$, $B_{SE} = 0.03$, $t(233) = 12.31$, $p < .001$). Based on 5000 bootstrapped samples, this indirect effect ($ab = -6.26$, $B_{SE} = 0.86$) was significantly different from zero (95% $CI = [-10.10, -4.95]$) and mediated the relationship between fairness and spread ($c' = -1.27$, $B_{SE} = 1.20$, $t(233) = -1.05$, $p = .29$). If (as demonstrated above) engagement in fairness considerations are content specific, then this may help explain the differences in moral judgements.

Figure 7.7

Standardised Coefficients for the Relationship Between "Fairness" MDP and Contribution to Spread Mediated by Moral Judgements



Note. Standardised values shown. Controlled for age and gender.

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

[18] The fairness probability score was based on likelihood of engaging with fairness-based judgements rather than across a vice-virtue continuum. Therefore, negative framing (e.g. 'unfair') would still contribute to a higher score.

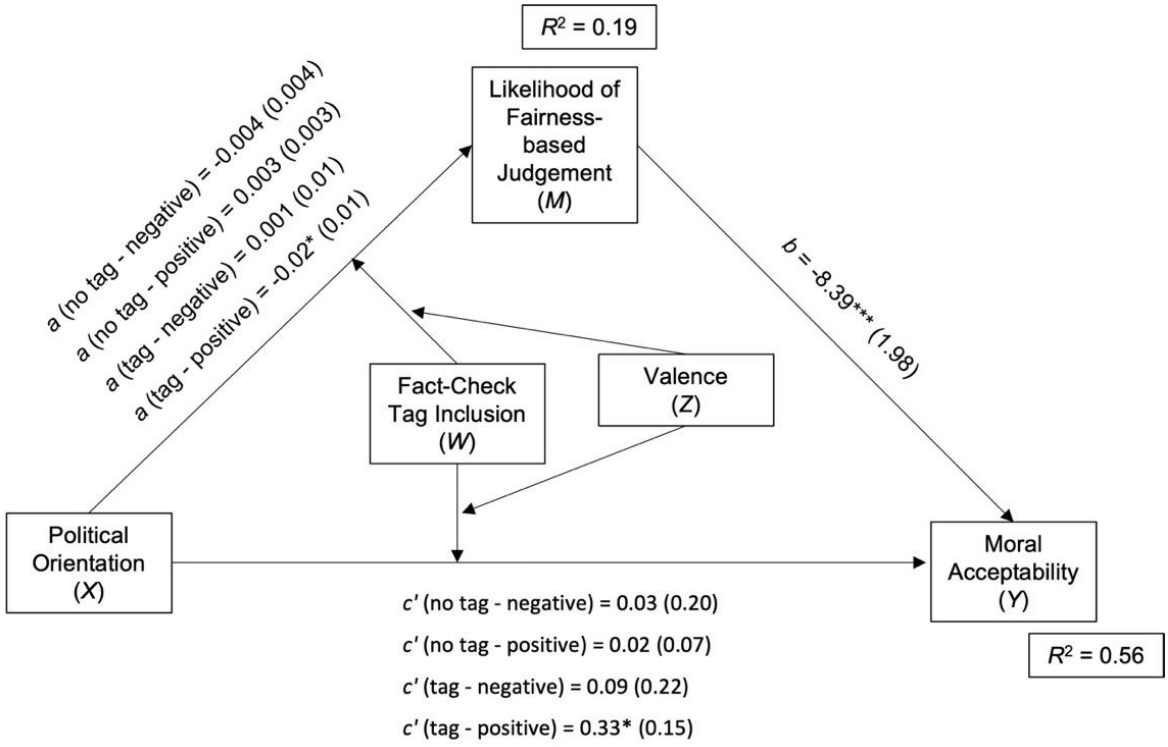
7.3.2.4 Ideology and Moral Judgements of Disinformation

A series of Spearman's Rho correlations were run to compare political orientation (PO, where an increase in values signifies the shift from left to right) with the moral judgements scores for each condition. These found that PO only significantly correlated with moral judgements of spreading positive misinformation when a fact-check tag was included, $r(64) = .27$, $p < .05$.

7.3.2.4.1 Ideological Moral Differences and the Moral Domain. To better understand the relationship between political orientation *(X)*, moral judgements *(Y)* and evaluations made within the 'fairness' domain (*M)*, PROCESS Model 12 was run (Figure 7.8). Assumptions for regression were again checked, with no violations observed. As the regressions found age and gender were significant predictors of moral judgements they were included in the model as a control. Model 12 is a moderated moderated mediation model, within which conditional effects produced by two moderators can be tested. Specifically, whether the impact of one moderator (e.g. *W*) on the strength of one relationship (e.g. $X \rightarrow M$) is dependent on another moderator (e.g. *Z*). By entering fact-check tag as a first moderator (*W*) and valence as the second (*Z*), it is possible to detect any ideological effects within each context (e.g. defined by levels of *W* and *Z*).

Figure 7.8

Conditional Indirect Effect of Political Orientation on Moral Judgements via Fairness Evaluations



Note. PO – value increase reflects shift from political left to right. Controlled for age and gender. *p < .05. ** p < .01. *** p < .001.

The index of moderated moderated mediation did not cross zero (B = 0.19, B_{SE} = 0.09 (95% CI = 0.01, 0.37)), suggesting that any indirect differences in effect size between misinformation and disinformation were dependent on valence (Table 7.7). The indices of conditional moderated mediation indicated that only labelling positive content with a fact-check "tag" produced a moderation effect (compared to no tag), B = 0.15, B_{SE} = 0.06 (95% CI = 0.04, 0.29).

Table 7.7

Ordinary Least Squares Regression Coefficients (With Standard Errors) From a First Stage Moderated Moderated Mediation Model Predicting Moral Judgements

		Outcome		
		M: Fairness		*Y:* Moral Judgement
Constant		0.14*** (0.03)		7.12***(1.12)
X: Political Orientation	$a_1 \rightarrow$	−0.004 (0.004)	$c' \rightarrow$	0.03 (0.20)
W: Tag	$a_2 \rightarrow$	0.02 (0.06)	$b_2 \rightarrow$	−1.19 (1.79)
Z: Valence	$a_3 \rightarrow$	−0.09* (0.03)	$b_3 \rightarrow$	5.50*** (1.19)
XW: PO x Tag	$a_4 \rightarrow$	0.01 (0.01)	$b_4 \rightarrow$	0.06 (0.29)
XZ: PO x Valence	$a_5 \rightarrow$	0.01 (0.01)	$b_5 \rightarrow$	−0.01 (0.21)
WZ: Tag x Valence	$a_6 \rightarrow$	0.15* (0.07)	$b_6 \rightarrow$	−4.37* (1.96)
XWZ: PO x Tag x Valence	$a_7 \rightarrow$	−0.02* (0.01)	$b_7 \rightarrow$	0.26 (0.33)
Age		0.00 (0.00)		−0.05*** (0.01)
Gender		0.01 (0.01)		0.43 (0.36)
M: Fairness			$b_1 \rightarrow$	−8.39*** (2.13)
	R	0.44		0.75
	R^2	0.19		0.56
			Index	95% bootstrap CI[a]
Moderated moderated mediation			0.19	0.01, 0.37
Conditional moderated mediation				
by Tag (*W*) where	Negative (*Z* = 0)		−0.04	−0.17, 0.13
	Positive (*Z* = 1)		0.15	0.04, 0.29

Note. Gender coded as a dummy variable (M = 0, F = 1).

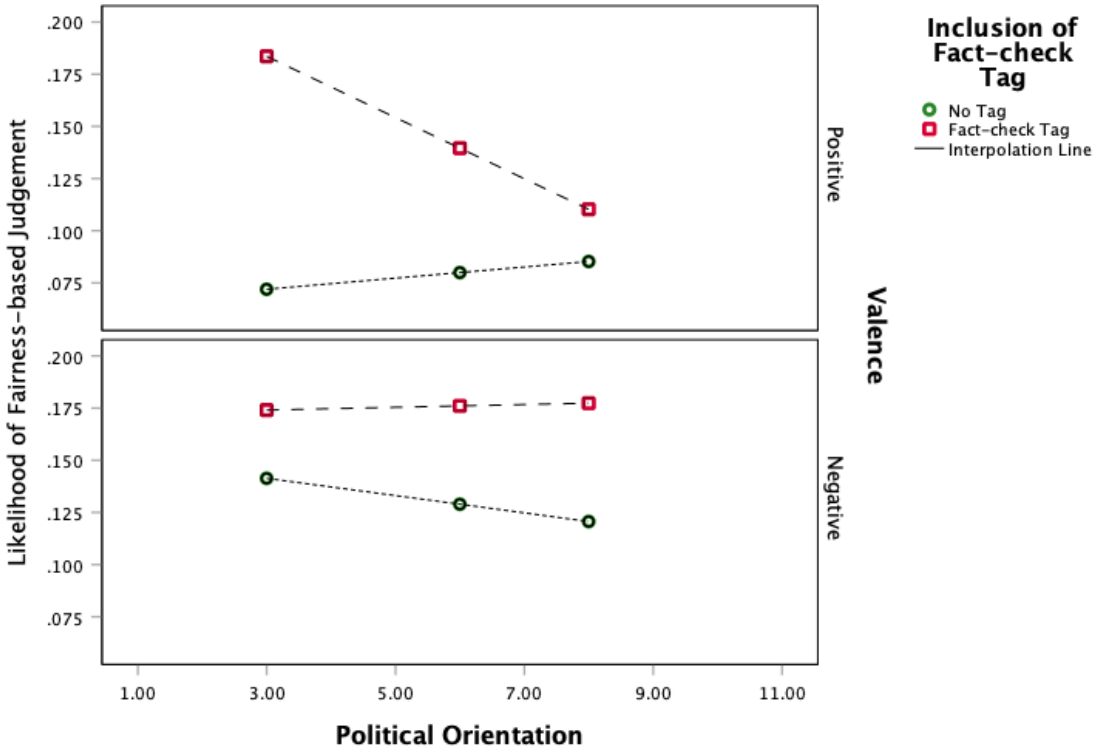
[a] Percentile bootstrap CI based on 5,000 bootstrap samples.

Notably, political orientation did not appear to affect the likelihood of engaging with the fairness domain when evaluating positive misinformation = −0.02, B_{SE} = 0.03, 95% CI = −0.08, 0.02. However, when participants were aware the content was disinformation (i.e. when a "tag" was included), political orientation influenced the likelihood of making fairness-based evaluations = 0.12, B_{SE} = 0.05, 95% CI = 0.03, 0.24.

This helped explain political asymmetry in moral judgements within this condition. As illustrated in Figure 7.9, left-leaning individuals may be as likely to engage with the fairness domain when presented with disinformation that supported their ingroup than when it undermined their ingroup. However, levels of fairness-based evaluations of positive disinformation more closely resembled evaluations of positive misinformation (i.e. "no tag") in right-leaning participants.

Figure 7.9

Conditional Effects of Valence and Tag on Engagement with Fairness Domain



7.3.2.5 Moral Foundations Questionnaire

Spearman's Rho correlations were carried out to determine whether MFQ scores were related to disinformation susceptibility (see Table 7.5, p.208). After applying Holms-Bonferroni adjustments, any significant direct relationships with either spread or moral

judgements were rejected[19]. Alternatively, MFT also supposes that high scores in a foundation corresponds with engagement with said "module", in this instance corresponding language use. However, a series of Spearman's Rho correlations suggest there is no relationship between a foundation's MFQ score and corresponding MDP scores overall (Table 7.8).

Table 7.8

Spearman's Correlations Between Moral Foundation (MFQ) and Domain (MDP) Scores

		Positive		Negative	
	All Conditions	No Tag	FC Tag	No Tag	FC Tag
n	237	61	58	61	57
Care / Harm	.04	.12	.02	-.04	.08
Fairness	.01	-.19	.36**	.07	-.08
Loyalty	-.02	.02	-.01	-.32*	.02
Authority	-.002	-.18	-.08	-.001	-.003
Sanctity	.02	-.12	-.02	-.01	-.04

Note. Each line represents correlations between respective MFQ and MDP scores. Holms-Bonferroni adjustments were again applied by moral domain (Appendix GG).

However, within the positive, fact-checked condition, those with high fairness MFQ scores appeared more likely to use "fairness" related words, $r = .36$, $p = .005$. Furthermore, although high scorers on MFQ ingroup loyalty were less likely to utilise "loyalty" related words in the negative, no-tag condition ($r = -.32$, $p = .01$), correlations between fairness MDP and all MFQ scores (Table 7.9) suggests that they instead utilised "fairness" related words ($r = .34$, $p = .007$). Rather than triggering loyalty-focused decisions (as MFT suggests), negative misinformation appeared to prompt high loyalty valuers to make fairness-based evaluations.

[19] It should be noted, however, that when looking at each condition individually the calculations may be underpowered

Table 7.9

Spearman's Correlations Between Moral Foundation (MFQ) and Fairness (MDP) Scores

		Positive		Negative	
	All Conditions	No Tag	FC Tag	No Tag	FC Tag
n	237	61	58	61	57
Care / Harm	.001	-.11	.23	.08	-.26[a]
Fairness	.01	-.19	.36**	.07	-.08
Loyalty	.08	.03	-.12	.34**	-.02
Authority	.03	.02	-.04	.15	.02
Sanctity	-.04	-.004	-.13	.13	.06

[a] *p*-value less than .05 but not within the threshold set out by holms-bonferroni calculations for within-domain comparisons

7.4. Discussion

The present study sought to demonstrate that the spread and moral judgements of identity-relevant misinformation and disinformation are influenced by valence and the labelling of content as false. The key aims were to better understand the moral processes underlying these judgements and help explain the political asymmetries observed in the previous chapters. The findings indicate that users are more likely to spread inaccurate content that positively frames their ingroup compared to negatively frames (H1). They also judged negatively framed content to be less morally acceptable to spread. Tagging content as false reduced both intentions to amplify spread and the moral acceptability of spreading (H2 & H3). Levels of moral acceptability explained the relationship between valence and spread of misinformation (e.g. no "tag"), where positive misinformation was judged as more acceptable to spread and, in turn, was more likely to be spread. However, when a fact-check was included, the effect of valence on moral judgements was not as strong, with moral acceptability explaining part of the relationship between valence of disinformation and spread (H4). Exploratory analysis identified further differences in engagement with fairness-related values that may help explain these responses.

The present study demonstrates that digital interactions may allow users to acknowledge (e.g. liking) or associate (e.g. sharing) themselves with content containing expressions of ingroup support (including disinformation and misinformation). People

appeared more willing to amplify the spread of positively framed, identity-relevant content than negatively framed, even when the former was labelled as untrue. This expands on previous research demonstrating ingroup biases in the sharing of misinformation (Osmundsen et al., 2021; Pereira et al., 2023) in that the present findings suggest positive misinformation about the ingroup (e.g. without "tag") may not always present the relevant cues needed to make potential identity threats salient. Firstly, participants who viewed positive misinformation were most willing to amplify its spread and gave the most favourable moral judgements. Secondly, those who identified as stronger supporters of their team were most likely to spread positive misinformation. This aligns with the premise of social identity theory, as situations that do not threaten identity may lead high group identifiers to feel safe to engage in identity-expression (Ellemers et al., 2002). These digital interactions may therefore allow people to express the positive distinctiveness of their ingroup when there is no identifiable risk in doing so. Assuming that spreading positive misinformation is not perceived as a threat to the individual, these findings may also indicate that related norms may be focused more on "not spreading inaccurate information" than "not spreading unverified information". As such, spreading positive misinformation may not be perceived to be a norm violation as, in their opinion, the content may be true.

The way participants within the other conditions made their judgements were also in line with social identity theory. Indeed, social identity theory suggests people are motivated to achieve and maintain a positive identity (Tajfel & Turner, 2004) and, as such, they are motivated to defend themselves against identity-threats (Ellemers et al., 2002). Here, participants not only judged content that could undermine the value of their ingroup as less morally acceptable to spread, they also were more likely to make proactive attempts to reduce its spread. This supports previous work finding users are more likely to denounce fake news using comments when the information threatens the ingroup (E. L. Cohen et al., 2020). By not holding identity-benefitting disinformation to the same standards, social

media users may be better able to maintain a positive identity by engaging in moral hypocrisy (see Valdesolo & Desteno, 2007). While it was found here that the relationship between valence and the spread of misinformation (no fact-check) was fully mediated by moral judgements, moral judgements of disinformation only partially explained this relationship. This shift away from reliance on moral judgements to guide behaviour indicates that when learning identity-beneficial content is inaccurate, other (social) factors may create moral conflict with personal moral standards. It may be that users feel that spreading disinformation is wrong generally (as indicated by the effect of "tag" inclusion on moral acceptability scores), but could be able to find ways of morally disengaging when the disinformation is otherwise beneficial, potentially allowing them to contribute to its further spread.

Concerningly, positive disinformation was also viewed as more morally acceptable to spread than either negative misinformation and disinformation. This indicates that content which negatively frames the ingroup may be viewed as a greater problem than false information that may otherwise help the group. This may be because the related identity-concerns (and ultimately their motivated responses), as well as the source of identity threat are likely to differ. Indeed, the identity-threat arising from viewing positive disinformation is ultimately based on potential consequences (e.g. if the user chooses to spread). As previous work has found, people may manage such reputational concerns arising from "fake news" by simply refraining from sharing (Altay et al., 2020). However, any content that negatively frames the ingroup presents a much more immediate threat to the group value, and as such, could be perceived as more serious.

The present study also sought to better understand the underlying basis of these moral judgements, where findings regarding the levels of engagement with the "fairness" domain provided some interesting insight. Notably, the eMFD categorises words such as "fake" and "lie" as fairness related words and as such engagement with this domain may relate to considerations of accuracy. Recent work suggests that people may spread misinformation

because they don't consider accuracy (Pennycook, Epstein, et al., 2021). Here, participants were most likely to spread misinformation that supported their ingroup and attracted lower levels of engagement with "fairness". However, introducing identity-threats appeared to increase engagement and, in turn, influence perceived moral acceptability. This supports Pennycook et al.'s (2021) findings that people do appear to care about accuracy. However, the findings here indicate that accuracy-related considerations may be more likely to occur in response to identity-threats.

Moreover, in line with study two's findings, political asymmetric moral judgements occurred only when identity-beneficial content was known to be inaccurate. Left-leaning participants were more likely to evaluate positive disinformation in the context of "fairness", and this helped explain differences in moral judgements. It could be argued that the inclusion of a label stating that the content was false may have been more likely to present an identity threat to left-leaning participants, but less so for right-leaning participants. However, given political asymmetry did not occur when participants were presented with group-threatening disinformation, right-leaning users do appear to care about disinformation in other contexts. Moreover, as people are motivated to be seen and see themselves as being moral (Jordan et al., 2011), it may also be that right-leaning participants were instead prioritising other values over "fairness" when presented with the positively framed disinformation. Indeed, several participants did highlight the charitable aspect of the core message. Future studies may wish to explore whether ideological differences in value-prioritisation or threat perception are more important in explaining the spread of potentially beneficial disinformation.

Despite previous work suggesting that high scores in binding foundations on the MFQ may be linked to increased susceptibility to misinformation (e.g. Ansani et al., 2021, Lunz Trujillo et al., 2021, Trevors & Duffy, 2020), the present study offers no strong evidence to suggest that is the case. If anything, those who prioritised loyalty (a binding foundation) appeared to be more sceptical of the negatively framed misinformation based

on engagement with fairness words. However, a notable distinction here is that aforementioned studies focused on health or science-based misperceptions whereas the present research focuses on identity. There was also no strong evidence to suggest, in line with MFT, that participants made evaluations in distinct domains at all. Rather, judgements appeared to occur across a spectrum in relation to fairness. However, if value-related processing is indeed interconnected (Turner-Zwinkels et al., 2021), and moral violations which may harm the reputation of the ingroup can trigger loyalty-based evaluations (Leidner & Castano, 2012), then this may explain why those high in loyalty rationalised the lower acceptability of spreading negative misinformation in the context of fairness.

One limitation of this study is that the MFQ scores had generally low reliability, suggesting a lack of consistency in responses across participants. As this may lead to biased estimates, MFQ based findings should be interpreted with caution. This is not, however, uncommon for the MFQ (Graham et al., 2011). The use of the eMFD for short passages of text is also not strongly advised (F. R. Hopp et al., 2021), but this is to increase the chances that moral words are used within said text. As participants were specifically asked about their moral judgements and given the medium strength of effects in relation to "fairness" scores, moral-related terms were likely to be present. Finally, as participants were presented with stimuli across a number of tasks, this may have increased the opportunity to notice any fact-check "tags". Future studies may wish to find ways to time-limit participant exposure to content to increase external validity.

Differences in how social media users evaluate content may help to explain their online behaviour. While interactions with content that benefits the ingroup may be made based on preferences, the present findings suggest that evaluations of potentially problematic content may be better explained in the context of threat saliency. Content that is considered to be beneficial in some way may not itself be considered a threat, but contextual factors (in this instance a fact-check "tag") may produce a perceived-threat, but potentially only for some.

7.4.1. Conclusion

The present study intended to expand on previous findings within this thesis and highlight the importance of social identity in determining moral evaluations and intentions to spread misinformation / disinformation. Fans of five English Premier League teams were assigned to one of four conditions, where an inaccurate post contained a positive or negative story about their own team, and either was or was not accompanied by fact-check information. Both valence and inclusion of a fact-check tag influenced the likelihood that participants would contribute to the onwards spread of disinformation and how morally acceptable they felt it would be to spread. The findings indicate that the contextual cues provided within the content may produce distinct types of identity-directed threats, which in turn may influence considerations of fairness. As shown here, the exact type of threat may be dependent on the viewer's awareness of veracity, as well as being influenced by differences in relation to political ideology and personal values.

Chapter 8. Study Five

8.1. Introduction

This chapter tests the effectiveness of moral reframing and accuracy interventions for reducing intentions to spread identity-beneficial misinformation. In particular, the application of moral reframing interventions relates to the ideological differences found in prior studies in this thesis. The present chapter therefore seeks to understand whether these differences are due to a tendency to associate misinformation with moral values that liberals may more readily prioritise. Specifically, it investigates whether political conservatives might be less accepting of spreading misinformation if it were framed as a violation of other cherished values (e.g. loyalty to ingroup). First, previously researched accuracy and identity-based interventions will be discussed. This is followed by discussion of the concept of "accuracy" itself, specifically how it relates to motivated reasoning and ideological differences in how it may be evaluated. Finally, the use of moral reframing interventions to encourage attitude and behavioural change regarding other politically divisive issues (e.g. climate change) will be explored. The effect of moral appeals (e.g. reframed for binding and individualising values) and accuracy interventions are tested using ANCOVA. The effects for both Democrat and Republican voters are then explored separately. Exploratory analysis looks at the relationships between identity strength, political orientation, and evaluations of spreading misinformation in the context of both interventions.

8.1.1. Current Misinformation Interventions

8.1.1.1 Accuracy Nudges

It has been suggested that the reason people spread misinformation could be due to other factors distracting them from thinking about accuracy. To date, several studies have suggested that drawing attention to the concept of "accuracy" may improve the quality of content shared on social media (Capraro & Celadin, 2022; Epstein et al., 2021; Pennycook,

McPhetres, et al., 2020; Pennycook, Epstein, et al., 2021; Pennycook & Rand, 2022; Roozenbeek, Freeman, et al., 2021). However, the findings present a somewhat mixed picture. For instance, in Pennycook et al. (2020), participants who were asked to rate the perceived accuracy of an unrelated headline prior to being asked to make "sharing" judgements were slightly more likely to "share" true COVID-19 headlines. However, the accuracy nudge did not reduce intentions to share misinformation, yet Pennycook et al. (2020) argued that in increasing the potential availability of accurate information the nudge may help improve the overall quality of content circulating on social media. Even so, the effect sizes observed in the study fall below definitions of "small" (e.g. $d = 0.14$) despite a relatively large sample (over 850 participants) and medium-sized differences between the perceived accuracy of both "true" and "false" headlines (as measured in their initial study). Indeed, there was a large-sized relationship between perceived accuracy and strength of the treatment effect, suggesting participants who viewed the accuracy nudge may have simply been more likely to spread content that appeared more plausible.

This strong relationship between perceived accuracy and treatment effect was found in another set of studies which focused on political headlines (Pennycook et al., 2021). Yet the actual effect of the intervention differed here, as the accuracy nudge appeared to reduce intentions to spread false headlines across three studies. This was also the case in a replication of Pennycook et al.'s (2020) COVID-19 headline study (Roozenbeek, Freeman, et al., 2021). Roozenbeek et al. (2021) suggested this may be because of the time that had passed, and the likelihood that participants were perhaps more aware of COVID-19 misinformation when their data collection occurred. Given the relationship between perceived plausibility and accuracy judgements (e.g. Pennycook & Rand, 2019) it could argued that such nudges promote spread-related decisions based on perceived plausibility rather than accuracy specifically. While users could then still be less likely to spread content they perceive as being inaccurate, that is not necessarily the same as them spreading less "misinformation" generally.

Furthermore, many of the aforementioned accuracy nudge studies focused exclusively on the concept of "sharing" and therefore it is not clear whether the nudge would influence other types of interactions. However, in a similar study, Capraro & Celadin (2022) found accuracy prompts did not influence liking. Given that expression on social media platforms may occur in a variety of ways (e.g. liking, commenting, etc) which may algorithmically contribute to the onward spread of misinformation, there is therefore value in understanding the effects of interventions beyond one single aspect of spread.

Finally, there is evidence to suggest the efficacy of accuracy interventions may be influenced by political orientation. Notably, in a replication (Roozenbeek, Freeman, et al., 2021) and a subsequent meta-analysis of accuracy nudge studies (Rathje et al., 2022) accuracy nudges were found to only be effective for Democrat supporters. However, in their own meta-analysis Pennycook & Rand (2022) suggest that this political asymmetry may only occur in MTurk samples (e.g. non-representative), and that political orientation does not otherwise moderate the treatment effect[20]. It may therefore be the case that accuracy interventions can be effective, but only for some of the population.

8.1.1.2 Identity-Related Interventions

An alternative intervention approach has been to define norms about the spread of misinformation. For instance, some studies have found that specifying desired behaviour may help to reduce misinformation spread (Andı & Akesson, 2020; Gimpel et al., 2021), although others have found conflicting results (Epstein et al., 2021). For instance, Gimpel et al. (2021) found that simply providing information regarding the number of people who had reported a post (e.g. a descriptive norm) had no influence on whether participants would report it themselves. However, when participants were presented with an injunctive norm about reporting (e.g. expressed reporting as a desirable behaviour) there was a small

[20] Notably, Roozenbeek et al.'s (2021) replication study recruited a national quota sample of participants through another platform (e.g. not MTurk) and still found political asymmetry.

increase in intentions to report. Furthermore, when both descriptive and injunctive norms were included this further increased intentions to report misinformation. Outlining behavioural expectations may therefore help reduce the spread of misinformation, including in ways other than reducing intentions to share.

Yet other studies have found descriptive norms to be potentially effective at reducing misinformation spread when framed in relation to the ingroup. For instance, a recent study found that including a "misleading count" (presented alongside other post information such as number of "likes") crowdsourced from fellow ingroup members helped to reduce intentions to spread, and were also more effective than accuracy nudges in doing this (Pretus et al., 2022). Seeing that other users within a personal network chose not to interact with misinformation may also influence sharing behaviour (C. M. Jones et al., 2021). The impact of seeing how others (particularly ingroup members) behave in relation to misinformation may therefore be valuable for influencing spread intentions.

8.1.2. Perceptions of "Accuracy" – Motivations and Political Orientation

Research suggests that people do care about sharing information that is accurate (Pennycook, Epstein, et al., 2021). However, what people believe to be accurate (e.g. "perceived truth") is not necessarily factually true. From a motivated reasoning perspective, such accuracy perceptions may be goal related (Kunda, 1990). When participants are asked to judge the accuracy of information in a study (as in Pennycook et al. (2021)) then, assuming their goal is to perform well in the task, they should be motivated to identify information based on whether it is factually true or not. Arguably, this specific scenario may not reflect users' experiences, goals, and motivations within social media platforms. Instead, as motivated reasoning perspectives would suggest, people may perceive false information as being "accurate" if it helps them achieve, for instance, a social goal. This suggests that incorrect information (from a factual standpoint) could be perceived as accurate in circumstances which may benefit the viewer in other ways.

It is therefore important to consider what motivates people to share accurate information outside of research participation. For instance, accuracy is an important consideration in regards to honesty. Indeed, the intentional sharing of inaccurate information may be considered as "dishonest" (e.g. a violation of honesty). Moreover, people are motivated to act in a way which is seen as "moral" by themselves and others and as such can influence reasoning processes. As such, people tend to assign more value and put more effort into being seen as a moral person than a competent one (Ellemers, 2017). While being perceived as incompetent may lead to negative emotions directed towards others (e.g. anger), being perceived as immoral may produce self-directed negative emotions (e.g. guilt, shame) that may be relatively more challenging to cope with (R. van der Lee et al., 2016). Therefore, the need to act "morally" may at times be more important for motivating users to refrain from interacting with misinformation than any need to be seen as "correct". However, while "honesty" may be a universal moral concept (Mann et al., 2016), leniency towards dishonesty may be influenced by social networks (Mann et al., 2014). Therefore, whether or not a person perceives even the sharing of disinformation to be "dishonest" may be situational.

While in many situations sharing accurate information will be the most morally acceptable option, there will also be times when doing so may conflict with other moral concerns, for instance, when telling the truth could cause harm to a target (a notable motivation for telling "white" or prosocial lies). This can be where individual differences appear. As discussed in the previous chapter, there is some evidence to suggest differences in how political conservatives and liberals prioritise moral values or "foundations" (Graham et al., 2009). For instance, liberals are thought to to prioritise individualising values (e.g. fairness and harm) over binding values (e.g. loyalty to ingroup, sanctity and authority). If that is the case, then, for liberals at least, spreading misinformation may not only block achievement of any fact-based accuracy goals, it may also potentially violate salient values related to upholding "honesty" or "truth". This prioritisation of

individualising values may help explain why accuracy-based interventions appear more effective for liberals generally in contexts that might require the upholding of "fairness" (e.g. politics misinformation) or "harm" (e.g. health misinformation). It may therefore be the case that considering the accuracy of certain content may present a potential violation of a prioritised value.

Conversely, for political conservatives (for whom individualising and binding values are relatively equally valued) accuracy goals may at times be more readily outweighed by other moral concerns. For instance, there may be situations where upholding "loyalty" is prioritised over "fairness" and could mean the sharing of ingroup-affirming disinformation is not always viewed as a moral violation. The effectiveness of a fact-based accuracy intervention may also be diminished. However, rather than simply not caring about fairness-related concepts, research suggests that allowing conservatives to think about the importance of honesty from their own perspectives has the potential to help shift their evaluations about others' dishonesty (Croco et al., 2021). From this perspective, it may be more effective for conservatives to evaluate honesty-related violations in the context of prioritised values such as loyalty, rather than attempting to appeal on the basis of 'fairness' or 'harm'.

8.1.3. Moral Reframing Interventions

The act of spreading content (including misinformation) may not necessarily be viewed as an expression of facts (as in the user may not have considered or indeed intended to claim that the item is objectively true). Indeed, it was observed in study four that people may still be more willing to knowingly spread untrue information (e.g. disinformation) when it is potentially beneficial for an ingroup than unverified information that may be potentially detrimental. Re-framing the "sharing" of content as an endorsement of knowing it is factually true has, however, been shown to be a more effective intervention for reducing the spread of misinformation than accuracy nudges (Capraro &

Celadin, 2022). Their findings indicate that the endorsement intervention worked in a distinctly different manner to an accuracy nudge, rather than simply amplifying the latter's effect. It also demonstrates the potential value in re-framing a user's perception of the meaning of certain actions within social media platforms in reducing misinformation spread. Therefore, it may also be possible to reframe actions such as liking to ensure they are perceived as an action that can boost the visibility of content (including content that is potentially problematic).

Furthermore, information that is relevant to a person is often more persuasive. In light of this, a number of studies tackling often politically-divisive issues have considered whether reframing interventions to be morally relevant can help to reduce political asymmetry. Specifically, "moral reframing" can allow messages about an issue to be tailored in a way which appeals to individuals' moral values (Feinberg & Willer, 2013, 2015; Hurst & Stern, 2020; Voelkel & Feinberg, 2018; Wolsko et al., 2016). For instance, climate change is an issue which is more commonly championed by political liberals. However, reframing the issue in the context of conservative values (e.g. in the context of potential purity violations) has been shown to help encourage political conservatives to engage in pro-environmental behaviour (Hurst & Stern, 2020; Wolsko et al., 2016) and potentially change their climate change beliefs (Feinberg & Willer, 2013). Conversely, arguments which frame military spending in the context of fairness (e.g. providing jobs that help reduce income inequality) may be more appealing to political liberals than traditional arguments focused on authority or loyalty (Feinberg & Willer, 2015). Moral reframing may therefore encourage issues to be evaluated in the context of personally-relevant values, where previous arguments have perhaps failed to engage.

8.1.4. The Present Study

This study will focus exclusively on misinformation which may help users express their social identity. That is, misinformation which favourably positions the ingroup in

comparison to an out-group. As demonstrated in study four, this is the type of content that users may potentially be most likely to spread, even when told it is false. Previous research indicates that interventions which attempt to define new misinformation related norms may be promising for reducing intentions to spread. The present study therefore seeks to test interventions which appeal to participants' moral values (either individualising or binding). Firstly, the individualising appeal frames the spread of unverified content as a potential violation of fairness and harm. Conversely, the binding appeal frames the spread of unverified content in the context of "binding values", for instance a potential loyalty violation. It is predicted that viewing a moral appeal (either binding or individualising) may lead participants to be more careful about spreading potential misinformation than those who do not see either appeal. Therefore, the first pair of hypotheses are:

> **H1a.** Participants exposed to either binding or individualising moral appeals will judge misinformation as less morally acceptable to spread than participants who are not.
>
> **H1b.** Participants exposed to either binding or individualising moral appeal condition will be less likely to contribute to the onward spread of misinformation than participants who are not.

Across the previous chapters there has been evidence of political asymmetry in relation to both intentions to spread and moral judgements of identity-beneficial disinformation. Research also suggests that "moral reframing" may help to close such gaps. Moral appeals may therefore be more effective when they are consistent with the moral values related to a person's ideology. For instance, prior research suggests that re-framing issues in the context of binding values can prove effective for political conservatives, which is of particular interest here given the findings in chapters 5 and 7. It is therefore predicted that people will make more negative evaluations of misinformation

when the appeal is consistent with (rather than opposes) the moral values associated with their political orientation:

H2a. The effect of a moral appeal on moral judgements of misinformation will be stronger when the appeal is consistent with participants' moral values.

H2b. The effect of a moral appeal on intentions to spread misinformation will be stronger when the appeal is consistent with participants' moral values.

H3a. Conservative participants who read a binding moral appeal will judge misinformation as less morally acceptable to spread than other conservatives.

H3b. Conservative participants who read a binding moral appeal will be less likely to spread misinformation than other conservatives.

There is growing evidence that encouraging users to consider accuracy may help reduce the spread of misinformation. Yet, people are also willing to spread disinformation (i.e. knowing that information is misleading or untrue) and therefore the saliency of accuracy may be only part of the picture. Research suggests that associating "accuracy" with a potential norm violation may help improve the efficacy of accuracy nudge. It is therefore predicted that viewing a moral appeal will amplify any negative effects on spread from viewing an accuracy-focused intervention. Therefore, Hypotheses 4 and 5 are:

H4a. The effect of an accuracy intervention on lowering moral judgements of misinformation will be stronger for participants who read a moral appeal.

H4b. The effect of an accuracy intervention on lowering intentions to spread misinformation will be stronger for participants who read a moral appeal.

H5a. The effect of an accuracy intervention on lowering moral judgements of misinformation will be strongest for participants who read a value-consistent moral appeal.

H5b. The effect of an accuracy intervention on lowering intentions to spread misinformation will be strongest for participants who read a value-consistent moral appeal.

Finally, several studies have suggested that prompting social media users to consider accuracy may help reduce intentions to share misinformation. Yet little is known about how it may influence spread in other ways. It is therefore predicted that:

H6a. Participants presented with an accuracy intervention will judge misinformation as less morally acceptable to spread than participants who are not.

H6b. Participants presented with an accuracy intervention will be less likely to spread misinformation than participants who are not.

Additionally, exploratory analysis will look at how strength of identity and political orientation may also influence the effectiveness of interventions and overall intentions to spread misinformation. This notably expands on findings from chapter 7, where strength of identity influenced intentions to spread identity-beneficial misinformation (but not disinformation), and political orientation was related to moral evaluations of identity-beneficial disinformation.

8.2. Method

8.2.1. Development of Stimuli and Pilot Study

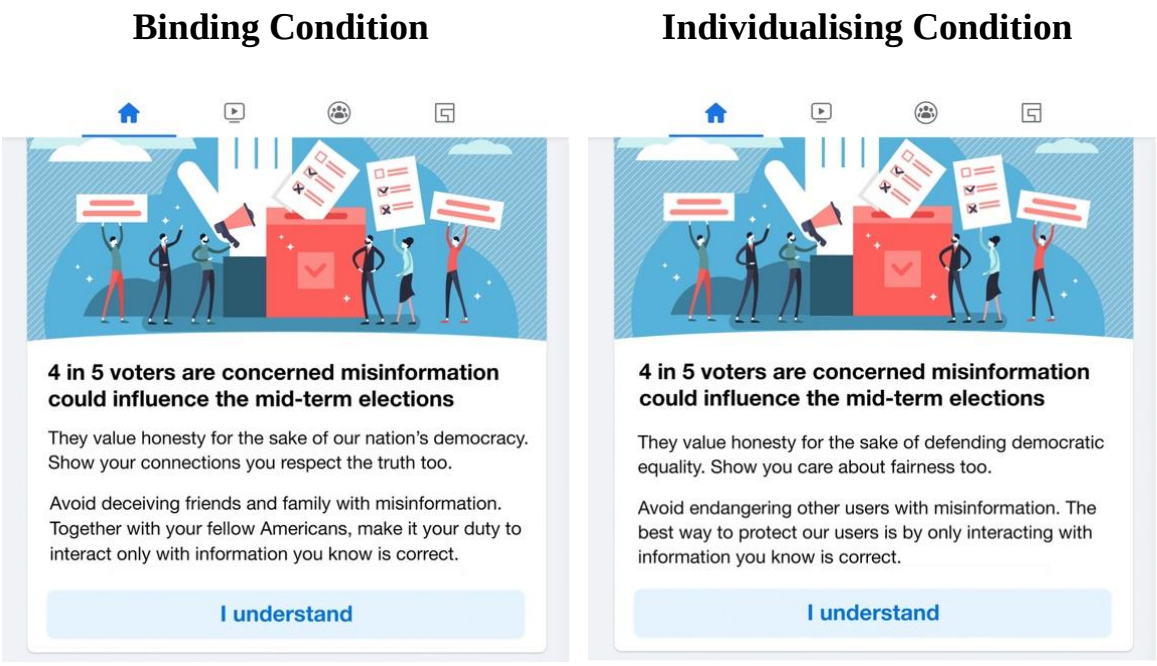
While accuracy interventions have been used in previous misinformation studies, no such moral reframing intervention exists. Therefore, prior to running the main study, two "moral appeals" were developed and tested in a pilot study.

8.2.1.1 Materials

The first moral appeal was intended to appeal to "binding" values, while the other was intended to appeal to "individualising" values. As with previous moral reframing studies, two statements were developed to appeal to the specific values and featured value-relevant words from the Moral Foundations Dictionary (Graham et al., 2009). The materials were designed to mimic an official update by Facebook, accompanied by a generic election graphic and recent statistic regarding misinformation concern. The final moral appeals can be seen in Figure 8.1.

Figure 8.1

Moral Reframing Appeal Stimuli for Study Five



Note. Participants were presented with either the Binding or Individualising appeal or did not see an appeal prior to viewing the misinformation post.

8.2.1.2 Participants

107 participants (48 males) aged 19-77 (M = 37.33, SD = 12.56) were recruited through Prolific to take part in the pilot study. Ethical approval was obtained from the University's Psychology Ethics Committee ETH2223-0568 (Appendix HH). Similar eligibility requirements were used to the main study for consistency. Participants were required to have an active social media account, speak fluent English and currently be residing in the United States. They also had to identify as either a Democrat (N = 54) or Republican (N = 53) voter.

8.2.1.3 Procedure

The study was hosted online using the survey platform Qualtrics. Participants answered basic demographic questions (gender, age and location), as well as confirming their political affiliation (i.e. party). Next, participants were randomly presented with the two moral appeals. First, they were shown the appeal for a minimum of 8 seconds. They were then asked to rate on a 7-point scale how closely the appeals matched different moral concerns (from "Not at all" to "A very large extent"). Ratings were made against each of the moral foundations (e.g. "Caring for / Reducing harm to others", "Respecting authority", "Loyalty to people", "Fairness / equality concerns", "Sanctity / purity concerns"). This wording is similar to previous pilot studies (e.g. Day et al., 2014). Participants were then shown the alternative appeal and asked again to make moral ratings before being thanked and debriefed.

8.2.1.4 Results

Mean binding and individualising ratings for both moral appeals are displayed in Table 8.1.

Table 8.1

Mean Moral Ratings (Binding & Individualising) for Moral Reframing Appeals

		N	Min	Max	Mean	SD
Binding Scores	Individualising Appeal	107	1	7	4.11	1.50
	Binding Appeal	107	1	7	4.37	1.43
Individualising Scores	Individualising Appeal	107	1	7	5.35	1.55
	Binding Appeal	107	1	7	4.91	1.53

There were significant differences between the two moral appeals on both dimensions. The binding appeal received higher ratings for "binding" compared to the individualising appeal, $t(106) = 2.20$, $p < .05$, $d = 0.21$. The individualising appeal received higher ratings for "individualising" compared to the binding appeal, $t(106) = 3.26$, $p < .01$, $d = 0.31$. This suggests that both appeals differ in terms of the moral values expressed.

8.2.1.5 Discussion

Participants judged each moral appeal against a set of binding and individualising moral values. They rated binding appeals to score more highly in binding values compared to the individualising appeal. They also rated the individualising appeal as scoring more highly in individualising values compared to binding values.

8.2.2. Main Study

8.2.2.1 Design

The present study employed a 3x2x2 between-subjects experimental design. The first independent variable (IV) was "Moral Frame", where participants either saw one of the appeals or no appeal prior to viewing the misinformation content (i.e. binding appeal vs individualising appeal vs no appeal). The second IV was "Accuracy Intervention", where participants were or were not presented with an accuracy-related statement alongside the misinformation content. The third IV was political affiliation (e.g. Democrat or Republican supporter). Two Dependent Variables (DV) were used in this study. The first DV was a rating of how morally acceptable participants felt it was to spread the presented

misinformation. The second DV was how likely participants were to contribute to the spread of the presented misinformation. Age, gender, strength of partisan identity, and political orientation were also collected.

8.2.2.2 Materials

8.2.2.2.1 Strength of Identity. As in study four, a single item measure of identity strength was used (Postmes et al., 2013). Participants were asked to state their level of agreement (1 - *strongly disagree* to 7 - *strongly agree*) with the statement "*I identify with being a …. supporter*", updated to reflect their reported political affiliation (i.e. Democrat or Republican).

8.2.2.2.2 Political Orientation. A single item question was used to identify participants political orientation (PO), *"Where would you place yourself on this political spectrum?".* Participants could rate from 1 ("Strongly Liberal") to 7 ("Strongly Conservative").

8.2.2.2.3 Social Media Spread Scale. The version of the scale used in study four was used to measure intentions to contribute to spread. The scale had acceptable reliability across the six conditions, where the lowest alpha was $\alpha = 0.76$.

8.2.2.2.4 Moral Judgements of Spreading. As with previous studies in this thesis, participants were asked to rate how morally acceptable it would be for them to share the content, rated on an 11-point scale *(0 –not at all morally acceptable, 10 - completely morally acceptable).*

8.2.2.3 Procedure

Recruitment took place on Prolific and the study was hosted on Qualtrics. First, participants saw the invitation letter and completed the consent form. Participants confirmed their country of residence before being presented with demographic questions. Next, participants were randomly allocated to one of three conditions (No moral appeal, Binding appeal, Individualising appeal). Participants allocated to a moral appeal condition

were first asked to read the moral appeal carefully and were prevented from proceeding for 8 seconds.

Next, participants were randomly allocated to one of another two conditions (No accuracy intervention, Accuracy intervention). This dictated whether the politically-congruent misinformation (as defined by their registration on Prolific) they were to be presented with would contain an accuracy prompt (Figure 8.2). The narrative within the post was similar to study two's stimuli, with statistics checked against the Major Cities Chiefs Association's (2022) violent crime midyear survey. Participants rated how likely they were to contribute to the online spread of the post using the Social Media Spread Scale and rated how morally acceptable they felt it was to spread. The presentation of the Spread and Moral Judgement blocks were randomised to control for order effects. Finally, participants completed the Political Orientation and Strength of Identity questions before being thanked and debriefed.

Figure 8.2

Misinformation Stimuli for Study Five by Political Affiliation

No Accuracy Nudge

Version for Democrat Voters

Version for Republican Voters

With Accuracy Nudge

Version for Democrat Voters

Version for Republican Voters

8.2.2.4 Participants

524 participants (267 males) aged 19-78 (M = 41.88, SD = 13.67) were recruited through Prolific on 2nd November 2022 (a week before the US mid-term elections) to take part in the study. Ethical approval was obtained from the University's Psychology Ethics Committee ETH2223-0568 (Appendix HH). To ensure enough power for a three-way ANCOVA, a power analysis was conducted using G*Power. Prior studies using accuracy interventions have reported small effect sizes. Therefore, to detect $\eta^2_p = .02$ at 80% power, a minimum of 476 participants was required. This also ensured that enough Republican voters would be recruited to detect $\eta^2_p = .04$ at 80% power in an ANCOVA for H3.

As before, participants were required to have an active social media account (e.g. Facebook, Instagram, etc) and must not have taken part in the pilot. Participants had to identify as either a Democrat or Republican supporter. They also had to be located within the United States and speak fluent English.

Thirteen participants were initially removed based on the pre-registered criteria. Of this, seven participants were removed due to Qualtrics screening tool flagging them as fraudulent and/or bots. Four participants were removed for potentially taking the study twice (Qualtrics duplicate score). Two participants were removed as their political affiliation did not match the recruitment criteria (e.g. Republicans who saw Democrat-congruent stimuli). Furthermore, another three participants were removed due to inauthentic responses on the "spread" responses (e.g. high intentions to amplify the spread and high intentions to prevent spread). However, the presented results continue to apply when these three participants are included in the data set.

Participant demographics for the final sample are shown in Table 8.2.

Table 8.2

Participant Demographics for Study Five by Political Affiliation

	All		Democrat		Republican	
	N	%	N	%	N	%
Total	508	100	256	50.40	252	49.60
Gender						
Female	246	48.40	131	51.20	115	45.60
Male	258	50.80	121	47.30	137	54.40
Non-binary	4	0.80	4	1.60	0	0.00
Education completed						
Some high school or less	2	0.40	1	0.40	1	0.40
High school diploma or GED	67	13.20	31	12.10	36	14.30
Some college, but no degree	84	16.50	40	15.60	44	17.50
Associates or technical degree	58	11.40	22	8.60	36	14.30
Bachelor's degree	207	40.70	103	40.20	104	41.30
Graduate or professional degree	90	17.70	59	23.00	31	12.30

8.2.2.5 Data Analysis

Data analysis for planned tests was pre-registered through AsPredicted.org (#110905, Appendix II). The planned tests in the present study used two 3x2x2 factorial analysis of covariance (ANCOVA) to test H1, H2, H4-H6. “Moral Appeal”, “Political Affiliation” and “Accuracy Intervention” were all between-group factors, with age and gender entered as controls. The first ANCOVA had a DV of Moral Acceptability (a) of spreading misinformation and the second had a DV of Intentions to Spread (b) the misinformation further. H3 was tested using two 2x2 factorial analysis of covariance (ANCOVA) based on Republican voter data only. These tests were followed by exploratory analysis utilising moderated mediation.

8.3. Results

Descriptive statistics are included in Table 8.3. There was some kurtosis and negative skewness in the strength of identity variable, as well as negative kurtosis within the political orientation scores. There was also some skewness and kurtosis in the spread scores within the individual conditions. It is, however, thought that any risks associated with skewness and kurtosis are reduced with a large sample (Tabachnick & Fidell, 2013).

Table 8.3

Summary of Descriptive Statistics

					Range			
	N	*M*	*SD*	*α*	Potential	Actual	Skew	Kurtosis
Age	507	42.12	13.64			19-78	0.58	−0.50
Strength of Identity	508	5.80	1.13		1-7	1-7	−1.31	2.72
Political Orientation	508	3.80	2.12		1-7	1-7	0.08	−1.47
All	507	6.19	1.61		1-11	1-11	0.72	0.90
Spread	507	6.19	1.61	.80	1-11	1-11	0.72	0.90
Moral Judgement	508	6.83	3.32		1-11	1-11	−0.27	−1.10
No Accuracy Label								
No Moral Appeal								
Spread	90	6.66	1.52	.76	1-11	4.33-11	0.94	0.33
Moral Judgement	91	8.35	2.63		1-11	1-11	−0.74	−0.18
Binding Appeal								
Spread	88	6.14	1.67	.81	1-11	1.67-11	0.69	0.69
Moral Judgement	88	6.41	3.71		1-11	1-11	−0.21	−1.40
Individ. Appeal								
Spread	70	6.35	1.56	.78	1-11	3.22-11	1.03	1.03
Moral Judgement	70	6.60	3.42		1-11	1-11	−0.19	−1.13
Accuracy Label								
No Moral Appeal								
Spread	80	6.21	1.58	.78	1-11	2.44-11	0.68	0.76
Moral Judgement	80	7.05	3.51		1-11	1-11	−0.40	−1.15
Binding Appeal								
Spread	83	5.85	1.47	.80	1-11	2.67-10.22	1.19	1.66
Moral Judgement	83	6.11	3.06		1-11	1-11	0.16	−0.97
Individ. Appeal								
Spread	96	5.96	1.72	.82	1-11	1-11	0.42	1.32
Moral Judgement	96	6.39	3.10		1-11	1-11	−0.04	−0.96

8.3.1. Planned Tests

8.3.1.1 Overall Moral Acceptability

To test H1a-H6a, a 3x2x2 between-groups ANCOVA was carried out. This looked at the effects of moral appeal (e.g. "no appeal", "binding appeal", "individualising appeal"), accuracy intervention (e.g. "no intervention", "accuracy intervention"), and political affiliation (e.g. "Democrat" or "Republican") on moral judgements of misinformation. Covariates of age and gender (coded as dummy variable) were included as controls. Levene's test was significant ($p < .05$) suggesting the assumption of homogeneity of variance was violated. A lower significance level (.01) was applied (Tabachnick & Fidell, 2013), with rejections confirmed using 3-way Robust ANOVA. Visual inspections of boxplots showed no outliers (Appendix JJ). Inspections histograms suggest the data are may not be normally distributed, however, ANOVAs are understood to be robust to violations of this assumption (Tabachnick & Fidell, 2013). Neither age nor gender contributed significantly to the model (Table 8.4).

Table 8.4

Three-Way ANCOVA Statistics for Moral Acceptability of Spreading Misinformation

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	12.59	1	12.59	1.26	.003
Gender [a]	29.81	1	29.81	2.99	.01
Moral Appeal	204.33	2	102.16	10.23***	.04
Accuracy	38.64	1	38.64	3.87*	.01
Political Affiliation	235.65	1	235.65	23.61***	.05
MA x AI	41.15	2	20.58	2.06	.01
MA x PA	33.01	2	16.51	1.65	.01
AI x PA	0.18	1	0.18	0.02	.00
MA x AI x PA	17.93	2	8.97	0.90	.004
Residuals	4881.29	489	9.98		

Note. $N = 503$, as four participants identifying as non-binary were excluded from this analysis.

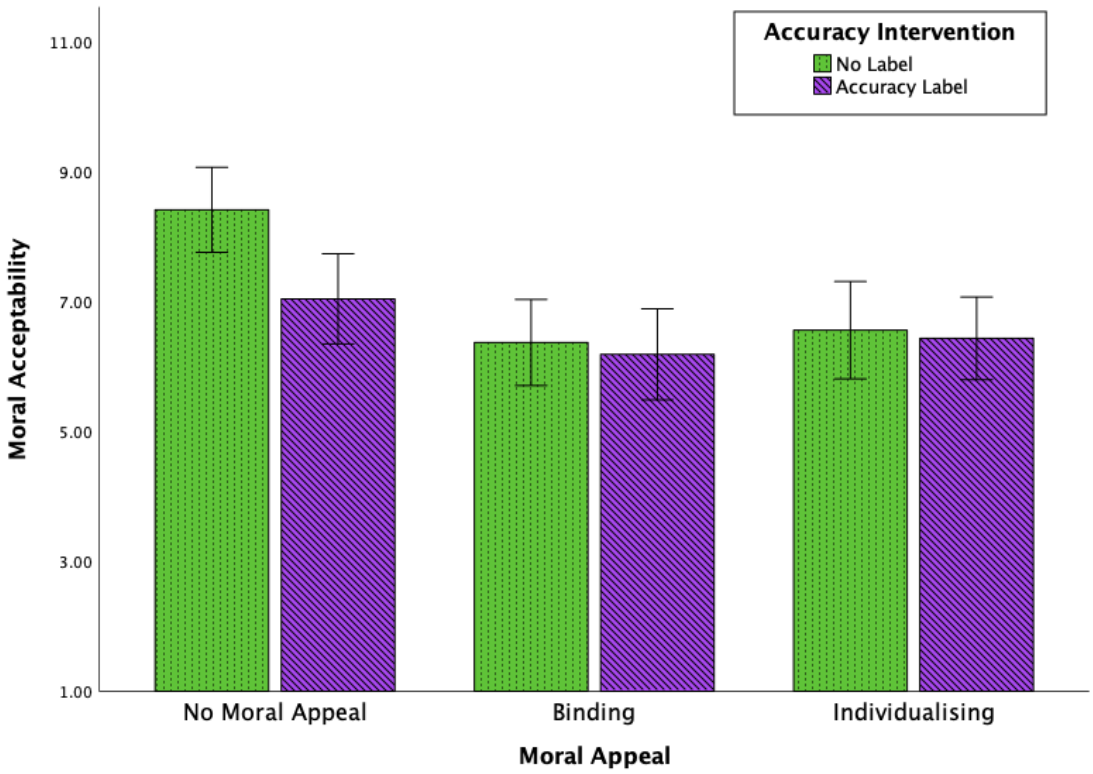
[a] Gender coded as dummy variable, F = 0, M =1.

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

The ANCOVA showed a significant main effect of "Moral Appeal" ($F(2, 489) = 10.23$ $p < .001$, $\eta^2_p = .04$). While this indicates a small effect, it is equal to Ferguson's (2009) minimum recommended effect size in social sciences, suggesting it may be of practical significance. Post-hoc comparisons using the Tukey HSD test (Appendix KK) found that participants who saw either "Binding" ($M = 6.27$, $SE = 0.25$) or "Individualising" ($M = 6.48$, $SE = 0.25$) moral appeals were significantly less likely to feel the misinformation was morally acceptable to spread compared to those who saw no appeal ($M = 7.71$, $SE = 0.24$). There was no significant difference between the two appeals on moral judgements (Figure 8.3). H1a is therefore accepted.

Figure 8.3

Estimated Marginal Means of Moral Judgement by Moral Appeal and Accuracy Nudge



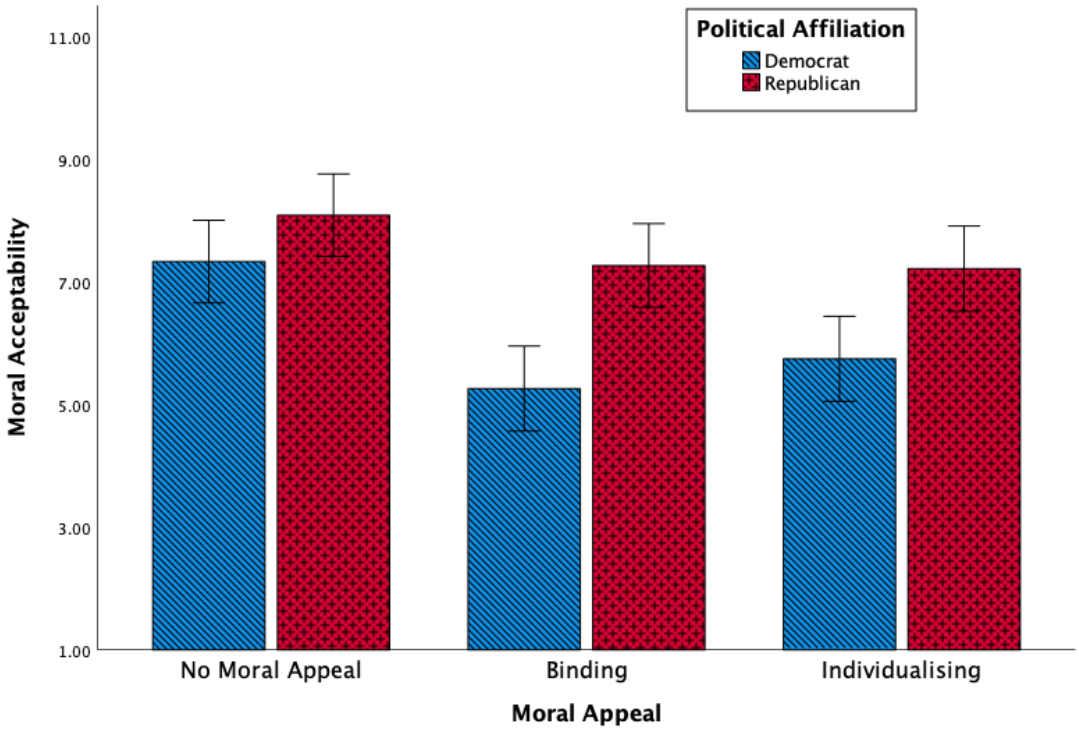
Note. Error bars 95% CI

It had been predicted that moral appeals would be more effective when they were consistent with the moral values associated with participant's political orientation. However, the interaction effect between moral appeal and political affiliation in the

ANCOVA was not significant (F (2, 489) = 1.65, p = .19, η_p^2= .01) and therefore provides no evidence to support H2a. However, there was a small main effect of political affiliation overall (F (1, 489) = 23.61, p < .001, η_p^2= .05) suggesting that the efficacy of the moral appeals were not themselves dependent on PA, but that PA did play a role in moral evaluations of misinformation. As illustrated in Figure 8.4, Republicans appeared to judge the misinformation as more acceptable to spread compared to Democrats.

Figure 8.4

Estimated Marginal Means of Moral Judgement by Moral Appeal and Political Affiliation



Note. Error bars 95% CI

Additionally, moral appeals did not appear to improve the efficacy of the accuracy intervention on moral judgements, as indicated by the non-significant 2-way interaction effect (F (2, 489) = 2.06, p = .13, η_p^2= .01). There was also no evidence to suggest this may differ by political affiliation, as indicated by the non-significant 3-way interaction effect (F(2, 489) = 0.90, p = .41, η_p^2= .004). Therefore there is no evidence to support H4a and H5a.

Finally, there was a significant but small main effect of "Accuracy Intervention" on moral judgements of misinformation (F (2, 489) = 3.87, $p < .05$, $\eta_p^2 = .01$). However, as Levene's test was significant, a 3-way Robust ANOVA was run to confirm the effect. The main effect of AI was found to not be significant (Appendix LL) suggesting that the accuracy intervention may not have influenced moral judgements of spreading misinformation. Therefore there is no evidence to support H6a.

8.3.1.2 Moral Acceptability – Republican Voters

The remaining hypothesis for moral judgements specifically predicted that binding appeals would be more effective for politically-conservative participants (H3a). A 3x2 between-group ANCOVA with factors of "Moral Appeal" and "Accuracy Intervention" was run (Table 8.5). There was no main effect of moral appeal (F (2, 243) = 1.69, $p = .31$, $\eta_p^2 = .01$). Therefore, there is no evidence to support H3a, as viewing either moral appeal had no significant impact on moral judgements for Republican voters.

Table 8.5

Two-Way ANCOVA Statistics for Moral Judgements in Republican Voters

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	10.38	1	10.38	0.917	.004
Gender [a]	2.17	1	2.17	0.191	.001
Moral Appeal	38.34	2	19.17	1.693	.01
Accuracy Intervention	18.36	1	18.36	1.621	.01
MA x AI	26.48	2	13.24	1.169	.01
Residuals	2751.06	243	11.32		

Note. [a] Gender coded as dummy variable, F = 0, M = 1.

8.3.1.3 Overall Intentions to Spread

To test H1b-H6b, a 3x2x2 between-groups ANCOVA was again carried out, but this time measuring intentions to spread misinformation. Covariates of age and gender were included as controls. Levene's test was significant ($p < .05$) suggesting the assumption of homogeneity of variance was violated. A lower significance level (.01) was

applied (Tabachnick & Fidell, 2013), with rejections confirmed using 3-way Robust ANOVA. Visual inspections of histograms suggest the data are somewhat normally distributed (Appendix MM). Neither age nor gender contributed significantly to the model (Table 8.6).

Table 8.6

Three-Way ANCOVA Statistics for Intentions to Spread Misinformation

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	3.45	1	3.45	1.42	.003
Gender [a]	6.77	1	6.77	2.78	.01
Moral Appeal	16.90	2	8.45	3.47*	.01
Accuracy Intervention	15.20	1	15.20	6.24*	.01
Political Affiliation	39.66	1	39.66	16.28***	.03
MA x AI	1.50	2	0.75	0.31	.001
MA x PA	7.72	2	3.86	1.58	.01
AI x PA	0.19	1	0.19	0.08	.000
MA x AI x PA	8.59	2	4.29	1.76	.01
Residuals	1188.81	488	2.44		

Note. N = 502 as four participants identifying as non-binary were excluded from this analysis.

[a] Gender coded as dummy variable, F = 0, M = 1.

*p < .05. ** p < .01. *** p < .001.

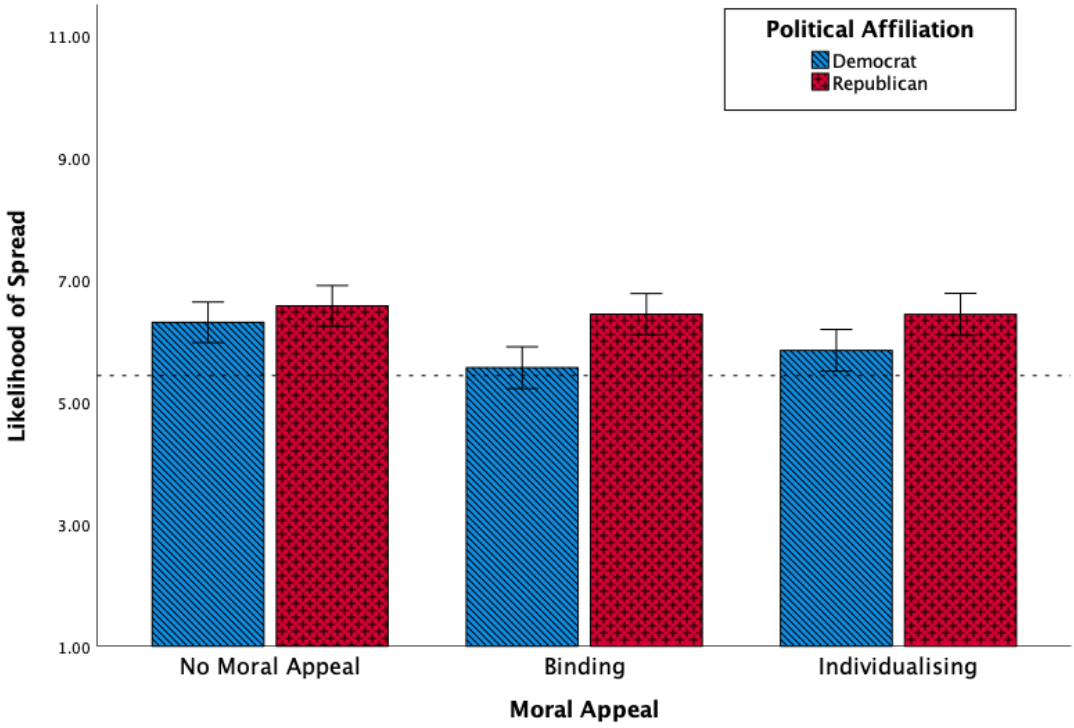
There was a main effect of "Moral Appeal" on intentions to contribute to the spread of misinformation (F (2, 488) = 3.47, p < .05, η_p^2= .01). However, due to the violation of homogeneity, a 3-way Robust ANOVA was run to confirm this decision. The main effect of MA was no longer significant (Appendix NN) suggesting that the moral appeal may not directly influence intentions to spread misinformation. There is therefore no evidence to support H1b.

The 3-way ANCOVA results also indicated there was no significant 2-way interaction effect, suggesting the moral appeals were no more effective when consistent with their political affiliation (F (2, 488) = 1.58, p = .21, η_p^2= .01) and therefore there is no evidence to support H2b (Figure 5). However, again there was a small main effect of

political affiliation overall (F (1, 488) = 16.28, p < .001, η_p^2= .03) suggesting that PA may be important for understanding intentions to spread misinformation.

Figure 8.5

Estimated Marginal Means of Spread by Moral Appeal and Political Affiliation.



Note. Dashed line indicates point at which participants with no intentions to interact in any manner (e.g. amplify or intervene) would fall. Error bars 95% CI.

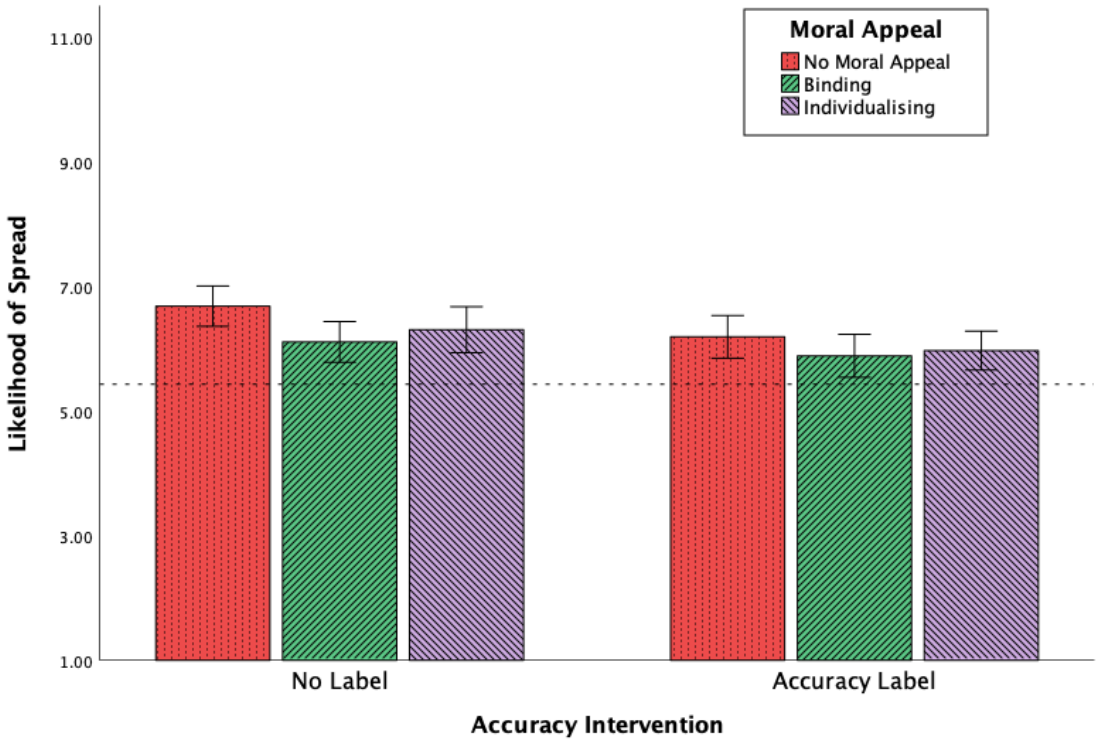
Again, moral appeals did not appear to improve any efficacy of the accuracy intervention on intentions to spread, (F (2, 488) = 0.31, p = .74, η_p^2= .001) and therefore there is no evidence to support H4b. There was also no 3-way interaction effect between political affiliation, moral appeal, and accuracy (F(2, 488) = 1.76, p = .17, η_p^2= .01) and therefore there is no evidence to support H5b.

However, there was a small main effect of "Accuracy Intervention" on intentions to spread misinformation (F (2, 488) = 6.24, p < .05, η_p^2= .03). Again, as Levene's test was significant, the 3-way Robust ANOVA will be referred to (Appendix NN). This confirmed

that the effect of the accuracy intervention was significant (Q = 5.77, p < .05). Reminders to consider accuracy may have a small impact on peoples' intentions to contribute to the spread of misinformation (Figure 8.6).

Figure 8.6

Estimated Marginal Means of Spread by Accuracy Nudge and Moral Appeal.



Note. Dashed line indicates point at which participants with no intentions to interact in any manner (e.g. amplify or intervene) would fall. Error bars 95% CI.

8.3.1.4 Intentions to Spread – Republican Voters

To test whether binding appeals would influence intentions to spread in Republican participants only (H3b), again a 3x2 between-group ANCOVA with factors of "Moral Appeal" and "Accuracy Intervention" was run (Table 8.7). There was no main effect of moral appeal (F (2, 242) = 0.16, p = .85, η_p^2= .001). Therefore, there is no evidence to support H3b, as neither moral appeal had a significant impact on Republican voters' intentions to spread misinformation.

Table 8.7

Two-Way ANCOVA Statistics for Likelihood of Spread in Republican Voters

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	1.14	1	1.14	0.39	.002
Gender [a]	3.13	1	3.13	1.05	.004
Moral Appeal	0.96	2	0.48	0.16	.001
Accuracy Intervention	5.90	1	5.90	1.99	.01
MA x AI	3.06	2	1.53	0.52	.004
Residuals	718.40	242	2.97		

Note. [a] Gender coded as dummy variable, F = 0, M = 1.

8.3.2. Exploratory Analysis

8.3.2.1 Intervention Effects on Democrat Voters

While Republicans neither appeared to be influenced by the moral appeals nor the accuracy intervention, political affiliation was a significant main effect in both ANCOVAs, suggesting that Democrats may have responded differently. Indeed, a 2-way ANCOVA found a medium-sized main effect of moral appeal on Democrat voters' moral evaluations of misinformation, $F(2, 244) = 11.20$, $p < .001$, $\eta^2_p = .08$ (Table 8.8).

Table 8.8

Two-Way ANCOVA Statistics for Moral Judgements in Democrat Voters

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	1.61	1	1.61	0.19	.001
Gender [a]	40.43	1	40.43	4.66*	.02
Moral Appeal	194.30	2	97.15	11.20***	.08
Accuracy Intervention	21.27	1	21.27	2.45	.01
MA x AI	33.64	2	16.82	1.94	.02
Residuals	2115.71	244	8.67		

Note. N = 252 as four participants identifying as non-binary were excluded from this analysis.

[a] Gender coded as dummy variable, F = 0, M = 1.

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

Furthermore, a significant main effect of moral appeal on intentions to spread misinformation was also found for Democrat voters, $F(2, 244) = 6.10$, $p < .01$, $\eta^2_p = .05$ (Table 8.9). Additionally, accuracy interventions appeared to have a small effect on intentions to spread, $F(1, 244) = 4.91$, $p < .05$, $\eta^2_p = .02$). This suggests that both types of intervention may be effective at reducing intentions to spread for some users (e.g. Democrat voters).

Table 8.9

Two-Way ANCOVA Statistics for Likelihood of Spread in Democrat Voters

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	2.47	1	2.47	1.28	.01
Gender	3.44	1	3.44	1.79	.01
Moral Appeal	23.50	2	11.75	6.10**	.05
Accuracy Intervention	9.46	1	9.46	4.91*	.02
MA x AI	6.96	2	3.48	1.81	.02
Residuals	470.23	244	1.93		

Note. N = 252 as four participants identifying as non-binary were excluded from this analysis.

[a] Gender coded as dummy variable, F = 0, M = 1.

8.3.2.2 Relationships with Political Orientation

To further explore the political asymmetry observed in the previous tests, a series of Spearman's correlations were run using the political orientation variable. Overall, these suggested that political orientation had a positive relationship with moral judgements and intentions to spread (Table 8.10), suggesting that the more politically right-leaning an individual is, the more acceptable they may feel it is to spread the misinformation and also more likely to contribute to its spread. Moreover, a number of medium sized effects were observed within certain conditions where a moral appeal had previously been viewed.

Table 8.10

Spearman's Correlations of Moral Judgements & Spread with Political Orientation

		No Accuracy Label			Accuracy Intervention		
	All	No MA	BMA	IMA	No MA	BMA	IMA
Moral Judgement	.27***	.20	.39***	.32**	.11	.27*	.30**
Intentions to Spread	.21***	.09	.34***	.21	.08	.19	.32**

Note. No MA = No Moral Appeal; BMA = Binding Appeal; IMA = Individualising Appeal.

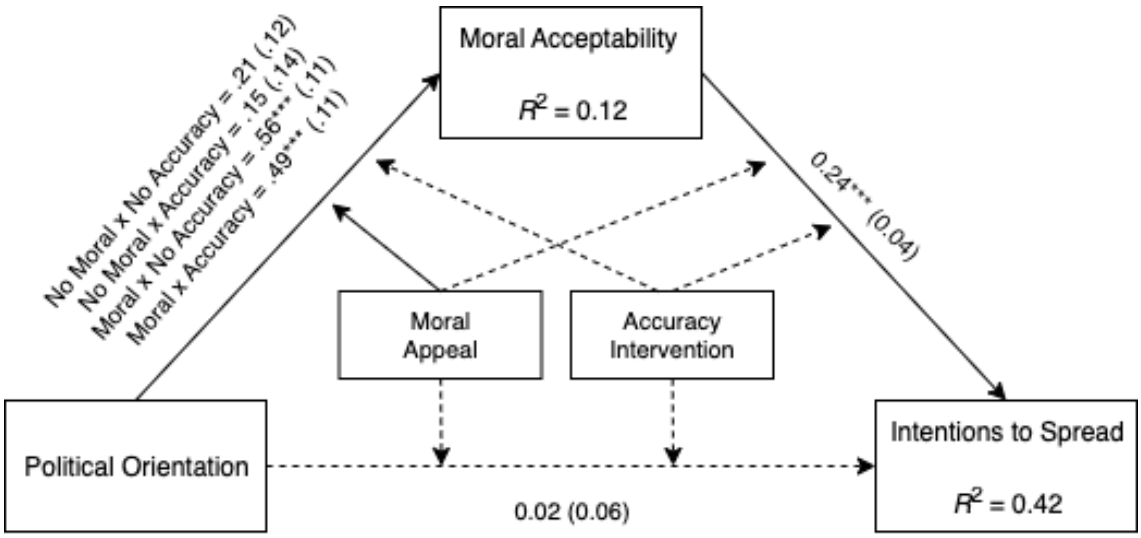
*$p < .05$. ** $p < .01$. *** $p < .001$.

An additive multiple moderation mediation model was then run to better understand the influence of the two interventions on relationships between political orientation *(X)*, intentions to spread *(Y)* and moral judgements (*M)*. Assumptions for regression were again checked, with no violations observed. Age and gender were not significant predictors of either *M* or *Y* and therefore were not included in the model. PROCESS Model 76 allows the conditional effects produced by two separate moderators to be tested on each path (e.g. *a, b,* and *c*). Notably, in this model the impact of one moderator (i.e. *W*) on the strength of one relationship (i.e. *X→M*) is not dependent on the second moderator (i.e. *Z*). As there was no significant difference between the two appeals on moral judgements and intentions to spread, the two appeal conditions were combined into one (e.g. "moral appeal" vs "no moral appeal"). This newly combined "Moral Appeal"

variable (*W)* and "Accuracy Intervention" (*Z)* were entered into the final model (Figure 8.7).

Figure 8.7

Unstandardised Coefficients for the Relationship Between Political Orientation and Likelihood of Spreading Misinformation Mediated by Moral Judgements



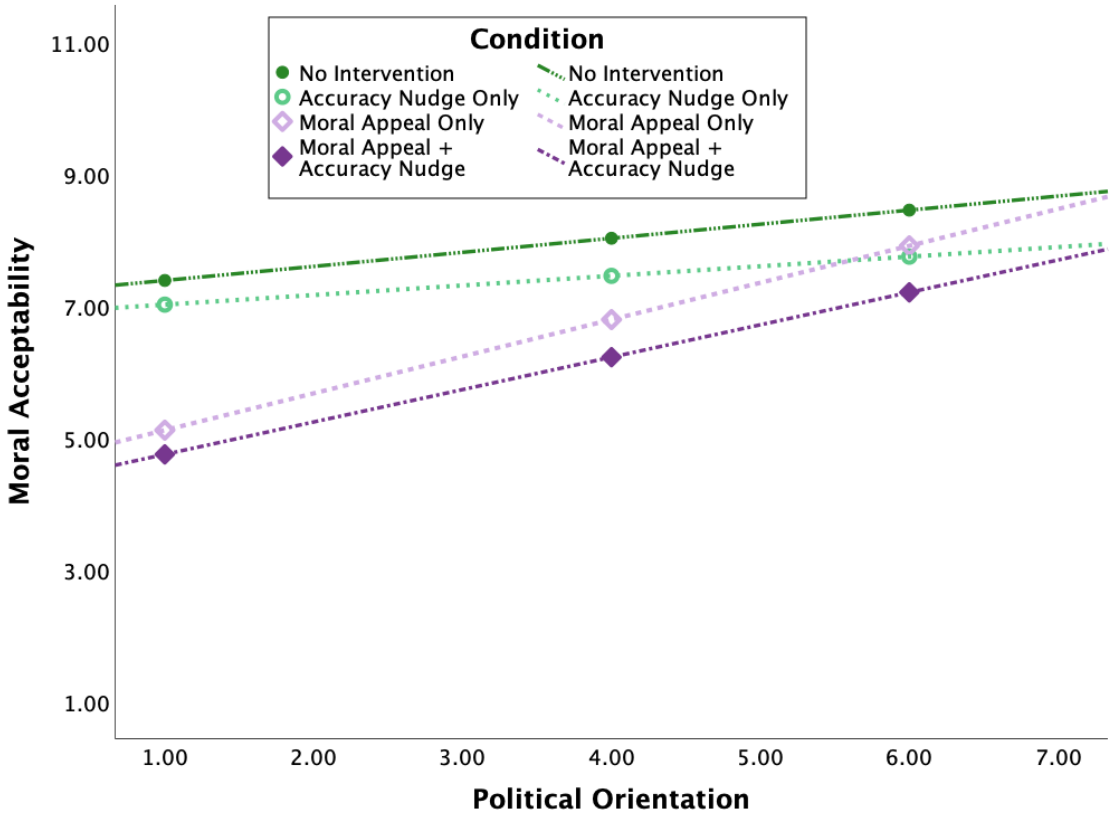
Note. Dashed lines indicate a non-significant path (where $p \geq .05$)

The *a* path saw an interaction effect between political orientation and moral appeal ($B = 0.35$, $B_{SE} = .14$, $t(501) = 2.47$, $p < .05$) indicating that moral appeal moderated the relationship between political orientation and moral judgements. Notably, conditional effects suggest there was no relationship between political orientation and moral judgements when participants did not see a moral appeal (regardless of whether the post was accompanied by an accuracy intervention ($B = 0.15$, $B_{SE} = 0.14$, 95% CI = −0.13, 0.42) or not ($B = 0.21$, $B_{SE} = 0.12$, 95% CI = −0.02, 0.44)). However, when participants saw a moral appeal the effect of political orientation on moral judgements was significant (again, regardless of whether an accuracy intervention was included ($B = 0.49$, $B_{SE} = 0.11$, 95% CI = 0.28, 0.70) or not ($B = 0.56$, $B_{SE} = 0.11$, 95% CI = 0.34, 0.78)). As illustrated in Figure 8.8, the effect of the moral appeal on moral judgements was greatest in those who

identified as “strongly liberal” but appeared to have little to no effect for those who indicated being politically conservative.

Figure 8.8

Conditional Effect of Political Orientation on Moral Judgement



In all, the model for the *a* path accounted for 12% of variance in moral judgements, $F(5, 501) = 15.17$, $p < .001$. Furthermore, the test of highest order unconditional interaction (*X*W) indicated that the moderation effect itself accounted for an R^2 change of 1.12% ($F(1, 501) = 6.08$, $p < .05$) or an $f^2 = 0.01$. While this is of course below Cohen's (1992) definition of a “small effect” size generally, it has been argued this may not apply to moderation effects. Indeed, as the average moderation effect size tends to generally be lower (Aguinis et al., 2005), Kenny (2018) suggests an f^2 of 0.01 may reflect a medium sized moderation effect here.

As in previous studies, moral acceptability was then associated with increased likelihood of spreading disinformation, $B = 0.24$, $B_{SE} = .04$, $t(498) = 5.90$, $p < .001$. Indirect effects between political orientation and intentions to spread via moral judgements only occurred when participants were shown a moral appeal. This was the case whether an accuracy intervention was presented ($B = 0.16$, $B_{SE} = 0.04$, 95% CI = 0.08, 0.24) or was not ($B = 0.16$, $B_{SE} = 0.04$, 95% CI = 0.09, 0.23). A conditional direct effect also occurred when participants saw both the accuracy intervention and moral appeal ($B = 0.09$, $B_{SE} = 0.04$, 95% CI = 0.01, 0.17). Full results can be found in Table 8.11.

Table 8.11

Ordinary Least Squares Regression Coefficients (with Standard Errors) from a First Stage Moderated Mediation Model Predicting Intentions to Spread Misinformation

		Outcome		
		M: Moral Judgements		*Y*: Intentions to Spread
Constant		7.18*** (0.51)		4.64***(0.32)
X: Political Orientation	$a_1 \rightarrow$	0.21 (0.12)	$c'_1 \rightarrow$	0.02* (0.07)
M: Moral Judgements			$b_1 \rightarrow$	0.24*** (0.04)
W: Moral Appeal	$a_2 \rightarrow$	−2.61*** (0.61)		−0.54 (0.34)z
Z: Accuracy Intervention	$a_3 \rightarrow$	−0.29 (0.58)		−0.58* (0.29)
XW: PO x Moral Appeal	$a_4 \rightarrow$	0.35* (0.14)	$c'_2 \rightarrow$	0.05 (0.06)
MW: MJ x Moral Appeal			$b_2 \rightarrow$	0.05 (0.04)
XZ: PO x Accuracy Intervention	$a_5 \rightarrow$	−0.07 (0.14)	$c'_3 \rightarrow$	0.01 (0.06)
XZ: MJ x Accuracy Intervention			$b_3 \rightarrow$	0.05 (0.04)
	R^2	0.12***		0.42***

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

8.3.2.3 Relationships with Strength of Identity

As study four suggested that strong identifiers may be more willing to spread identity-beneficial misinformation (but not disinformation) and therefore understanding whether the two interventions influence this relationship may be beneficial. A series of

correlations suggest that strength of identity (SOI) had a small but consistent relationship with intentions to spread misinformation when no accuracy intervention was shown (Table 8.12).

Table 8.12

Spearmans Correlations of Moral Judgement & Spread with Strength of Identity

		No Accuracy Label			Accuracy Intervention		
	All	No MA	BMA	IMA	No MA	BMA	IMA
Moral Judgement	.17***	.15	.20	.34**	.08	.25*	.04
Intentions to Spread	.16***	.22*	.23*	.31**	.05	.20	−.03

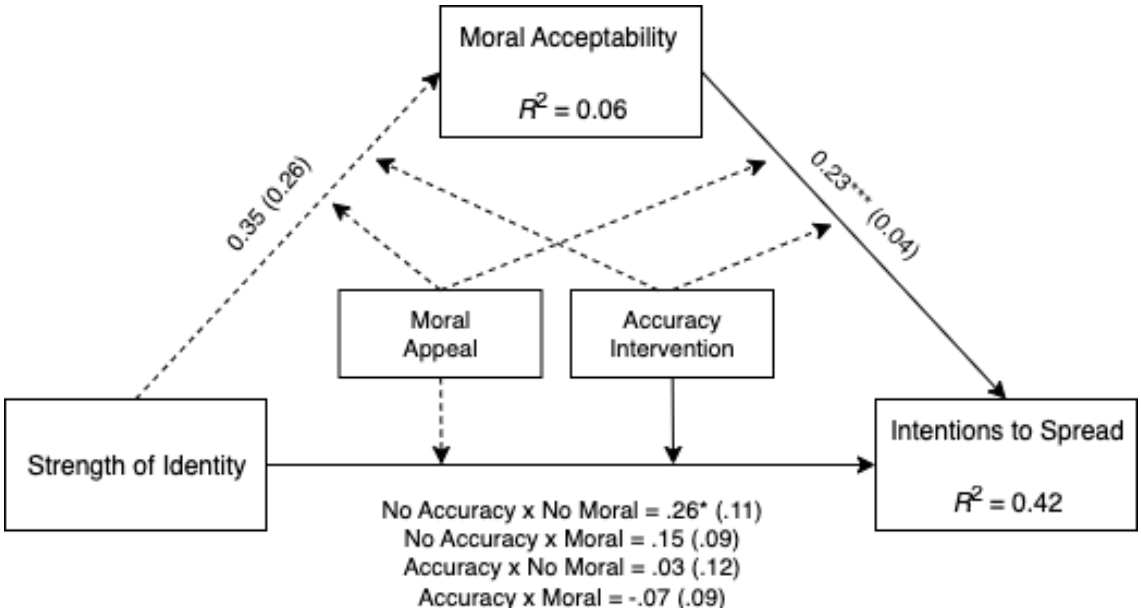
Note. No MA = No Moral Appeal; BMA = Binding Appeal; IMA = Individualising Appeal.

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

To better understand the relationship between strength of identity *(X)*, intentions to spread *(Y)* and moral judgements (*M)*, PROCESS Model 76 was again run. “Moral Appeal” (*W)* and “Accuracy Intervention” (*Z)* were entered into the model as separate moderators (Figure 8.9). Assumptions for regression were again checked, with no violations observed. Gender was a significant predictor of *M* so was included in the model as a control.

Figure 8.9

Unstandardised Coefficients for the Relationship Between Strength of Identity and Likelihood of Spreading Misinformation Mediated by Moral Judgements



Note. Dashed lines indicate a non-significant path (where $p \geq .05$)

As shown in Table 8.13, the effect of SOI on the *a* path was not significant, B = 0.39, B_{SE} = .26, $t(496) = 1.53$, $p = .12$. However, SOI significantly predicted intentions to spread on the *c'* path (B = 0.27, B_{SE} = .11, $t(493) = 2.52$, $p < .05$) suggesting stronger identifiers were more likely to spread identity-beneficial misinformation.

Table 8.13

Ordinary Least Squares Regression Coefficients (With Standard Errors) from a First Stage Moderated Mediation Model Predicting Intentions to Spread Misinformation

		Outcome		
		M: Moral Judgements		*Y*: Intentions to Spread
Constant		5.45 (1.53)		3.21*(0.67)
X: Strength of Identity	$a_1 \rightarrow$	0.39 (0.26)	$c'_1 \rightarrow$	0.27** (0.11)
M: Moral Judgements			$b_1 \rightarrow$	0.23*** (0.04)
W: Moral Appeal	$a_2 \rightarrow$	−2.91 (1.82)		0.25 (0.76)
Z: Accuracy Intervention	$a_3 \rightarrow$	1.74 (1.73)		0.68 (0.66)
XW: SOI x Moral Appeal	$a_4 \rightarrow$	0.28 (0.31)	$c'_2 \rightarrow$	−0.11 (0.12)
MW: MJ x Moral Appeal			$b_2 \rightarrow$	0.07 (0.04)
XZ: SOI x Accuracy Intervention	$a_5 \rightarrow$	−0.40 (0.29)	$c'_3 \rightarrow$	−0.22 (0.12)
XZ: MJ x Accuracy Intervention			$b_3 \rightarrow$	0.06 (0.04)
Gender[a] (Control)		0.55 (0.29)		0.08 (0.11)
	R^2	0.07***		0.42***

Note. N = 503 as four participants identifying as non-binary were excluded from this analysis.

[a] Gender coded as dummy variable, F = 0, M = 1.

*p < .05. ** p < .01. *** p < .001.

There were no significant interaction effects, however the moderation effect of the accuracy intervention on the *c'* path trended towards significance (B = −0.22, B_{SE} = .12, t(493) = −1.93, p = .05), although the confidence intervals crossed zero (95% CI = −0.45, 0.004). This was also the case for the test of highest order unconditional interaction (*X**Z), which indicated that the moderation effect itself (f^2 = 0.006) was on the very edge of significance (F(1, 493) =3.72, p = .05). While any interpretation of this finding should of course be taken with caution, there are known power issues with moderation effects on continuous variables (Kenny, 2018) and so the tests presented here may be

underpowered[21]. Indeed, Kenny (2018) suggests that a moderation effect of $f^2 = 0.006$ would represent a small effect.

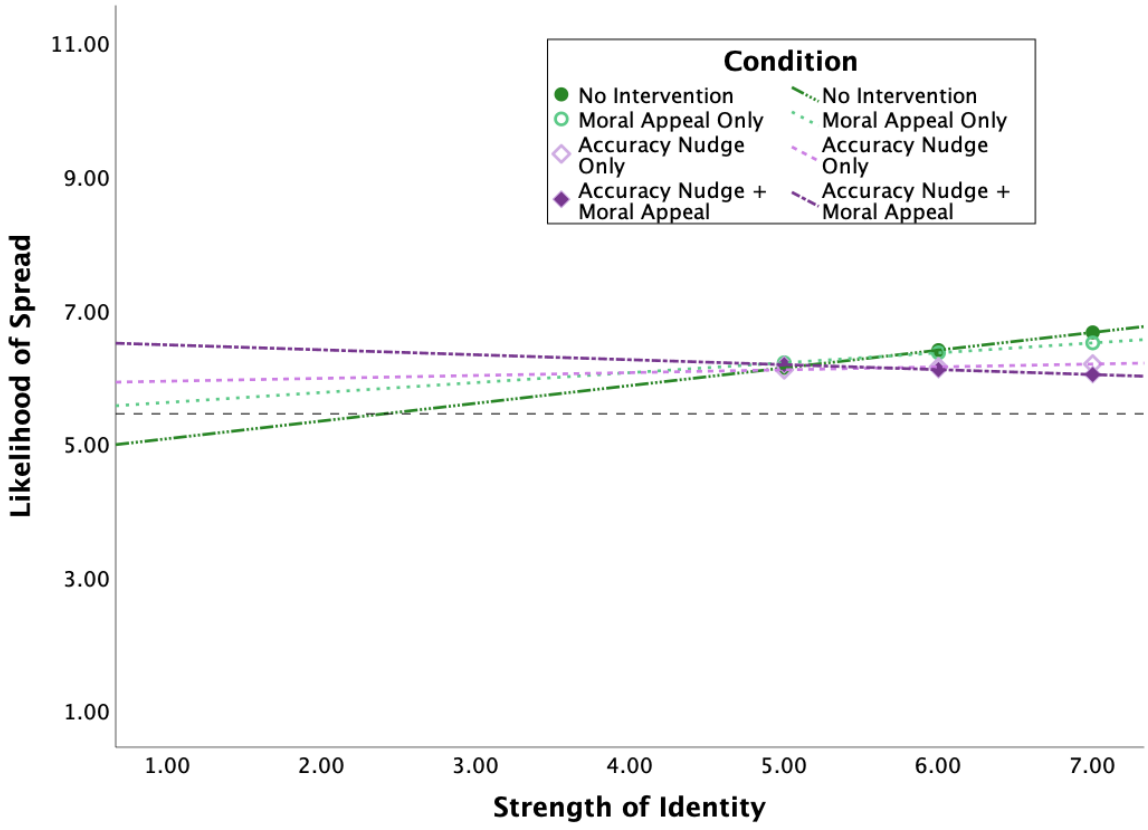
Notably, there was only one instance of a significant conditional direct effect between SOI and intentions to spread: when no intervention was seen ($B = 0.27$, $B_{SE} = 0.11$, 95% CI = 0.06, 0.47). Yet, when an accuracy intervention was presented alongside the content, the effect of SOI on spread was closer to zero (regardless of whether participants previously saw an moral appeal ($B = -0.07$, $B_{SE} = 0.09$, 95% CI = −0.25, 0.11) or not ($B = 0.05$, $B_{SE} = 0.04$, 95% CI = −0.21, 0.29). As illustrated in Figure 8.10, it may be that the accuracy intervention helped to flatten the slope[22].

[21] Indeed, when the same test is run without 'Gender' as a control variable, the coefficient for the interaction effect on the *c'* path is significant, but the coefficient only differs by 0.01 ($B = -0.23$, $B_{SE} = .11$, $t(498) = -2.01$, $p < .05$). Furthermore, without the additional control variable the test of higher order unconditional interaction (X*Z) is significant ($F(1, 501) = 4.04$, $p < .05$), but again with a small change in the effect size ($f^2 = 0.0057$).

[22] Examining the effects of the interventions on the relationship between SOI and spread (e.g. without moral judgements as a mediator – Model 2 in PROCESS) indicates this may be the case. While higher levels of SOI are related to increased intentions to spread ($B = 0.37$, $B_{SE} = .12$, $t(496) = 3.05$, $p < .01$), the moderating effect of seeing an accuracy intervention almost entirely flattens the slope ($B = -0.32$, $B_{SE} = .76$, $t(496) = -2.45$, $p < .05$). This is then confirmed by the test of highest order unconditional interaction (*X*W)* where $f^2 = 0.013$ ($F(1, 493) = 5.99$, $p < .05$), a medium effect as defined by Kenny (2018). It may therefore be plausible to suggest that an accuracy nudge weakens any effect of identity strength on spread, bringing strong identifiers closer to a level of interaction on par with others.

Figure 8.10

Conditional Direct Effect of Strength of Identity on Intentions to Spread Misinformation



Note. Figure illustrates the direct effect between strength of identity and intentions to spread when moral judgement is accounted for as a mediator and gender is included as control. Dashed line indicates point at which participants with no intentions to interact in any manner (e.g. amplify or intervene) would fall.

8.4. Discussion

The present study sought to test whether moral reframing interventions may be an effective method for reducing the spread of misinformation, particularly in the context of political conservatives. Furthermore, the effectiveness of a pre-existing accuracy "nudge" intervention was tested. The findings indicate that moral appeals may have the potential to influence how people make moral evaluations of misinformation (H1a) but may not

directly impact their intentions to spread (H1b). Yet, there was no evidence to suggest that the consistency between moral appeal and participants' moral values influenced evaluations (H2 and H3). Nor was there evidence to suggest the effect of the moral appeal amplified any effect of the accuracy intervention (H4 and H5). However, the accuracy intervention was shown to potentially reduce intentions to spread misinformation (H6b) but not moral evaluations of spreading misinformation (H6a). Exploratory analysis supported the findings that the two interventions played distinct roles. Specifically, the moral appeal moderated the relationship between political orientation and moral judgements, in a way that suggests that they reduced how morally acceptable politically left-leaning participants felt the misinformation was to spread (compared to right-leaning participants). However, the accuracy intervention helped to subdue any effect of strong identification on intentions to spread misinformation.

There were significant differences between supporters of both political parties in relation to moral evaluations and intentions to spread. Analyses at a group level demonstrated that the effectiveness of both interventions related to Democrat voters only. For these participants, the moral appeal had a medium sized effect on lowering perceived moral acceptability (which in turn played an important role in influencing intentions to spread). For Democrats, the moral appeal and accuracy intervention had small effects on lowering intentions to spread. Given the scale of social media and the knowledge that not all users will intend to interact with all content, these findings are potentially promising. Upon viewing an intervention, some participants may have reduced intentions to engage in "amplifying" behaviour (e.g. "likes", sharing, etc), thus clearly lowering its potential spread. Others may have been more likely to actively attempt to intervene in the spread of content (e.g. by "downvoting", etc). However, a sizeable proportion may have simply continued to passively engage with the content (e.g. neither actively amplifying or reducing its spread) regardless of whether they encountered an intervention. While these passive users may not present such a problem in real world contexts (where notably

billions of users are logging into social media platforms each day), their presence is important to consider when interpreting the effect sizes here.

However, while the interventions may have been effective for Democrat voters, for Republican voters neither intervention appeared to have an effect. Exploratory analysis indicated such differences may have occurred at the level of moral reasoning, in that the moral appeals appeared to influence moral judgements (but mostly in left-leaning participants). Given intuitive moral processes are thought to occur pre-cognition (Haidt, 2001) these apparent differences in how information about disinformation is morally evaluated may help provide some explanation for the political asymmetry in previous accuracy nudge studies (e.g. Rathje et al., 2022; Roozenbeek et al., 2021). If an intervention in some way makes moral norm violations more salient for mostly politically left-leaning users, then it could indirectly influence intentions to spread misinformation in a way that potentially presents as political asymmetry.

Furthermore, unlike previous studies (e.g. Feinberg & Willer, 2013; Hurst & Stern, 2020; Wolsko et al., 2016) the use of moral reframing did not help to reduce intentions to spread in political conservatives. This may be due to the source of the message, which here was the social media platform. Indeed, previous work found moral reframing to only be effective for political conservatives when coming from a conservative source, rather than neutral or liberal (Hurst & Stern, 2020). Others have found that when interventions are framed as being from the ingroup they can also become more effective for reducing political conservatives' intentions to share misinformation (Pretus et al., 2020). Future studies may therefore wish to adjust the source of moral appeals to understand whether they have the potential to be an effective intervention across the political spectrum.

The lack of distinctiveness between the two moral appeals did, however, suggest that they may have simply encouraged participants to engage in more deliberative reasoning about the content. While the moral appeals were shown prior to seeing the misinformation content, said appeals had an effect on moral evaluations (regardless of

whether accuracy interventions were produced). Conversely, the accuracy interventions shown alongside the misinformation content asking participants to consider the accuracy of information did not appear to influence moral evaluations to the same extent. Conversely, accuracy interventions appeared to play a small role in reducing intentions to spread. It may be the case that the accuracy interventions influenced decisions, but instead on an intuitive level. Specifically, it appears that the people who were more likely to spread the content (e.g. high identifiers) were only more likely to do so when they did not view any appeal. When strong identifiers viewed an accuracy appeal, their intentions to spread appeared to differ very little from others (and yet their moral evaluations did not appear to be influenced by any of the interventions or by SOI generally). If the two types of intervention do target distinct levels of reasoning, then arguably they may both play a role in reducing misinformation spread. Through behavioural regulation, interventions targeted at a moral level may help promote an intuitive sense of 'wrongness' which may deter users' from spreading unverified content further.

There are of course a number of limitations in this study, mainly that only one piece of misinformation was tested. Additionally, the item was one created specifically for the study rather than a post from social media itself. This was to ensure that the effects of the intervention could be directly comparable between supporters of both parties, however it may not have had the same overall level of "user appeal" that viral misinformation might. Therefore, effects of the intervention may be underestimated. Future studies may therefore wish to test these interventions against "real" misinformation (particularly items which have previously generated higher levels of engagement).

Notably, while moral reframing may have not been effective here, the present study has highlighted the distinct processes underlying users' intentions to spread misinformation on social media. While debate continues in the field regarding whether classical reasoning or identity-based approaches are more important for understanding and tackling the spread of misinformation, multi-faceted approaches may be warranted.

8.4.1. Conclusion

This study tested two types of interventions (moral appeals and accuracy interventions) to understand whether they may help reduce the spread of misinformation on social media. Democrat and Republican voters were assigned to one of six conditions, where they had the potential to be shown a moral appeal (either binding or individualising) and / or an accuracy intervention. Accuracy interventions were found to reduce intentions to spread identity-benefitting misinformation, while moral appeals influenced how morally acceptable participants' felt the misinformation was to spread. The findings suggest that the interventions may influence different types of reasoning (e.g. deliberative vs intuitive). However, the interventions appeared to only be effective for Democrat voters / left-leaning participants, even when moral interventions were tailored to appeal to conservative values.

Chapter 9. General Discussion

9.1. Introduction

The present thesis looked at the influence of social identity on social media users' evaluations and intentions to spread disinformation: firstly, by determining the relationship between identity-related beliefs and users' evaluations of disinformation; next, by investigating whether moral judgements of disinformation are context dependent and influenced by social identity; finally, by understanding whether threats to social identity influence how users interact with and make judgements of disinformation.

Five studies were carried out to address the following research questions (RQ):

1. Are individuals more likely to contribute to the spread of disinformation (e.g. through digital interactions or inaction) when the message appeals to group-related beliefs, attitudes or values? (Studies 1, 3)
2. Do moral judgements of spreading identity-related disinformation differ according to the content's potential impact on achieving positive distinctiveness? (Studies 2, 4, 5)
3. Do perceived identity threats influence the moral judgements of spreading identity-related disinformation? (Studies 4, 5)
4. Do moral processes play a role in user contributions to disinformation spread? (Studies 3, 4, 5)

The present chapter will summarise the findings of the thesis in relation to each of the research questions. The role of belief-consistency on intentions to spread misinformation is discussed, as well as findings relating to group-related beliefs specifically. Ingroup biases in moral judgements are then summarised, before discussing the potential influence of identity threats. Specifically, distinctions between the moral

judgements of misinformation and disinformation are made, before discussing instances where misinformation may have presented a group-directed, self-directed or no threat. In the context of RQ4, the relationships between moral acceptability and intentions to spread are first discussed. This is followed by findings relating to moral intuition and moral reasoning respectively.

Additionally, findings relating to potential ideological differences and moral foundation theory (including the role of "fairness" values) are discussed. Methodological implications are considered as well as implications for current interventions, before presenting limitations and recommendations for future research.

9.2. Research Question 1: Are Individuals More Likely to Contribute to the Spread of Disinformation When the Message Appeals to Group-Related Beliefs, Attitudes, or Values?

9.2.1. Degree of Belief Consistency Influences Intentions to Spread Misinformation

While prior research identified beliefs and believability as predictors of misinformation susceptibility generally, relatively little is known about the use of misinformation to express specific beliefs on social media. However, opinion expression is a key motivation for using social media (Whiting & Williams, 2013), and for sharing misinformation (X. Chen et al., 2015). To understand the relationship between belief-consistent misinformation and spread, studies one and three looked at whether issue-specific beliefs influence the types of misinformation themes and narratives users interact with. Overall, participants reported greater intentions to spread misinformation than other people when the content expressed opinions that were consistent with their own beliefs.

The misinformation presented in study one was distinguished into two overarching themes (e.g. threat of the COVID-19 virus and the UK Government's handling of the pandemic). Within each theme were two sets of misinformation stimuli, each expressing opposing opinions (e.g. stance) on the issue (totalling four sets of misinformation in study

one; two sets in study three). The findings indicated that the relationship between a belief and intentions to spread misinformation were dependent on the exact theme and stance presented. More specifically, participants were more likely to spread misinformation than other people when it was consistent with their beliefs. For instance, higher levels of trust in the UK Government's handling of COVID-19 were related to increased intentions to spread misinformation that was "favourable" towards the UK Government. It also predicted greater moral leniency towards spreading the content even upon learning it was untrue. Yet the opposite was found for unfavourable misinformation, where lower trust in Government was related to greater intentions to spread. Notably, trust levels had no bearing on intended interactions with virus-related misinformation, where instead beliefs about the risk of COVID-19 were more relevant (study one). These findings support previous research demonstrating the role of confirmation bias in interactions with misinformation (e.g. Kim et al., 2019).

One reason people spread misinformation is because they believe it (Buchanan, 2020; Halpern et al., 2019; A. Kim et al., 2019; Pennycook, Bear, et al., 2020). Beliefs represent what a person feels is true (Huber, 2009), which may be why belief-consistent misinformation is sometimes viewed as "more accurate" (Allcott & Gentzkow, 2017; A. Kim et al., 2019), and why people may claim to recall belief-consistent (but false) events (Murphy et al., 2021). Belief-consistent misinformation could be associated with a sense of "truth", that may extend to disinformation (e.g. information disclosed as being false or misleading).

From a motivated reasoning perspective, accuracy-goals can be achieved without factually correct judgements based on objective truth (Leeper & Slothuus, 2014). "Feelings" of accuracy may suffice, even if evidence suggests otherwise. Indeed, where the "general idea" (e.g. "gist") expressed by disinformation is perceived as true, people may be more lenient about spreading it (Effron & Helgason, 2022). Therefore, if the "gist" of belief-consistent disinformation feels true, this may explain participants' leniency towards

spreading it[23] (study one). Indeed, individuals tend to feel they are less vulnerable to misinformation than other people (Corbu et al., 2020; Jang & Kim, 2018; P. L. Liu & Huang, 2020; Ștefăniță et al., 2018). Therefore, compared to misinformation that others share, belief-consistent disinformation may not feel like such a substantial deviation from "the truth".

9.2.2. Spreading Group-Related Beliefs

As would be expected, participants were more willing to spread misinformation that appealed to (rather than conflicted with) group beliefs. In studies one and three, Labour voters reported significantly lower trust in the UK Government's handling of the pandemic than Conservative voters[24]. Labour voters were also more likely to spread misinformation that unfavourably framed the Government. Yet notably, while Conservative voters were more willing to spread "favourable" misinformation in study one, they were not in study three (despite tending to report it was acceptable to do so). Instead, external factors (such as the substantial changes in public opinion at the time (Ipsos MORI, 2021)) may have played a role. While people may feel comfortable expressing their beliefs with likeminded others, social media platforms (SMPs) provide an added layer of visibility compared to the "offline" world that users may be conscious of. Indeed, many behaviours within SMPs are persistent (e.g. recorded for posterity), replicable (e.g. may be copied), and searchable (Boyd, 2008). They may also be observed by a potentially infinite, invisible audience. However, as SMPs also allow individuals to selectively disclose aspects of their identity (which may be what occurred in study three), confirmation bias or "agreement" alone cannot explain why people spread belief-consistent misinformation.

---

[23] While the relationship between moral judgements and intentions to spread disinformation was not established in study one, it was demonstrated in study four where there was a strong relationship between moral judgements and intentions to spread 'positive' disinformation.

[24] The Conservative Party formed the UK Government at the time of data collection for both studies

Indeed, users often need to be considerate of multiple social identities at any one time. While an individual may themselves feel that a certain act (such as spreading belief-consistent misinformation) is morally acceptable, if they perceive that doing so may violate an ingroup norm, they may regulate their behaviour to appease others (Ellemers et al., 2002). Unless there is strong motivation to do otherwise, it may be safer for individuals (from a social identity perspective at least) to simply refrain from amplifying misinformation when they perceive their beliefs to be in the minority. Indeed, this is what previous online research indicates (H. T. Chen, 2018; Fox & Holt, 2018; Woong Yun & Park, 2011; Wu & Atkin, 2018). As groups have a social-regulatory function, the alternative is to risk attracting criticism from an audience (some of whom may be fellow Conservative voters, some of whom may not). Taken together, the research indicates that the beliefs, attitudes, and values of the group (in this instance, a perceived audience) may influence how individuals interact with belief-consistent misinformation in more ways than one.

Finally, how individuals evaluate belief-consistent misinformation that is also connected to their identity may be important. Group beliefs help to define groups and, in some instances, distinguish ingroups from outgroups (Bar-Tal, 1998), and therefore it may be likely that fellow ingroup members spread misinformation that is more consistent with what a user perceives to be "true" than misinformation spread by the outgroup. Such group-related discrepancies in belief-consistency of misinformation may help to explain why people have a tendency to associate disinformation with outgroups (Lyons et al., 2020; Tong et al., 2020; van der Linden et al., 2020). Ingroup biases in evaluations of misinformation will be discussed in more detail over the following sections. However, the findings discussed here may offer useful context for understanding why people feel that they (and their ingroup) are less vulnerable to misinformation than others.

9.3. Research Question 2: Do Moral Judgements of Spreading Identity-Related Disinformation Differ According to the Content's Potential Benefit for the Ingroup?

9.3.1. Ingroup Biases in Moral Judgements of Spread

Social media users are more likely to spread (Osmundsen et al., 2021; Pereira et al., 2023) and believe (Faragó et al., 2020; Neyazi & Muhtadi, 2021; Pereira et al., 2023) misinformation that potentially benefits rather than harms their ingroup. Other research has also observed biases in moral judgements in favour of a political ingroup (Effron, 2018; Helgason & Effron, 2022). The present thesis adds to this body of work by distinguishing between disinformation and misinformation. After controlling for other factors, participants made moral judgements about spreading both misinformation and disinformation in a manner that was preferable to the ingroup (studies two & four). This effect occurred for both political misinformation (study two) and football misinformation (study four).

As expected, participants were more accepting of spreading misinformation that positively framed their ingroup than spreading misinformation that made the ingroup look bad (studies two and four). For misinformation about the outgroup, however, the opposite was true (study two). Indeed, misinformation that made an outgroup look bad was judged to be more acceptable to spread than ingroup-undermining misinformation. This indicates that moral evaluations of spreading misinformation were not simply based on message valence (e.g. positive or negative) or how the ingroup was framed specifically, but in relation to the viewers' social identity.

The uniformity of participant judgements in response to relatively small adjustments within the stimuli are consistent with Social Identity Theory (SIT) (Tajfel & Turner, 2004). Indeed, moral leniency was not exclusively based on the potential usefulness of content for expressing ingroup membership (e.g. social creativity). Participants were also more lenient towards spreading misinformation that would facilitate

social comparison strategies (e.g. positive differentiation). SIT suggests such acts of positive differentiation are competitive (Tajfel & Turner, 2004), and therefore the spread of identity-related misinformation may not be simply motivated by a need to express identity or heightened belief-consistency. Leniency towards spreading misinformation that undermines an outgroup may therefore be a competitive strategy for group members to achieve or maintain relative superiority.

Notably, awareness of disinformation "status" does not necessarily alleviate ingroup bias. Indeed, Pereira et al. (2023) suggest people are more likely to believe and share political misinformation that positively positions the ingroup (vs. outgroup) in such a manner that identity-concerns appeared to be prioritised above accuracy. The present work supports this finding by demonstrating ingroup bias in the context of disinformation (studies two & four). It did not matter whether participants were explicitly told previously viewed information was false or if the post itself was labelled with a fact-check message; they were still more morally lenient about spreading content that benefitted the ingroup. Given that individuals are motivated to achieve or maintain positive self-esteem (Tajfel & Turner, 2004), false information that facilitates this aim may be perceived favourably compared to false information that could be detrimental for the ingroup (and in turn, the self-concept). Indeed, the present findings indicate that people may be more tolerant towards spreading false information that supports the ingroup than potentially true information that criticises the ingroup (study four). Moral evaluations of misinformation and disinformation may therefore prioritise identity over accuracy.

9.4. Research Question 3: Do Identity Threats Influence the Moral Judgments of Spreading Identity-Related Disinformation?

The findings discussed so far indicate that moral judgements of misinformation may be flexible in relation to a user's identity. Rather than people simply being more or less lenient towards misinformation, it may be that the contextual basis against which these

evaluations are made changes in relation to the impact content has on identity. As in, whether misinformation and disinformation encountered within an SMP (a social environment) presents an identity-threat. Indeed, information (and therefore arguably misinformation and disinformation) can be "both a source of threat and source of potential resources to deal with threats" (Ellemers et al., 2002, p. 166). For instance, while some information may present no perceivable threat, other information may appear to threaten the value of the group (e.g. group-directed threat). Furthermore, as groups define and regulate moral behaviour (see Ellemers, 2017), amplifying the spread of certain information may lead to social repercussions, and therefore threaten an individual's position within their group (e.g. self-directed threat). The general differences between moral evaluations of "misinformation" and "disinformation" will be discussed, before focusing on how moral evaluations of misinformation appeared to be made in the context of group-directed threats, the absence of identity-threats, and self-directed threats.

9.4.1. Distinctions Between Evaluations of Disinformation and Misinformation

As people perceive disinformation as being harmful (J. W. Cheng et al., 2021), it was anticipated that participants would judge disinformation as significantly less morally acceptable to spread than comparable misinformation (studies two & four). Moreover, when people spread disinformation, their perceived intentions influence other's evaluations of them (R. Young et al., 2023). For instance, spreading false information to intentionally deceive others is likely to be viewed less favourably than accidentally spreading false information to help others. This means that knowledge of disinformation "status" is likely to shift the context against which any judgement is made.

Yet, the "misinformation" conditions presented here provided no indications to suggest that the information presented was factually accurate. The large effect sizes that occurred in studies two & four are therefore notable. While it is not possible to ascertain from the present work whether such judgements were comparable to evaluations of "true"

information, Pennycook et al. (2021) have previously argued that the reason that people spread misinformation is because they don't consider accuracy. The findings here may support this. For instance, study five suggests a moral appeal encouraging the viewer to only interact with factually accurate information may shift moral evaluations and help to reduce intentions to spread misinformation (at least in political liberals). Therefore, thinking about the possibility that information could be inaccurate may help to reduce the spread of misinformation.

Yet, as previously discussed, the present work also indicates that people may be more morally lenient towards spreading disinformation that is consistent with their beliefs (study one), supports their ingroup (studies two & four), or undermines an outgroup (study two). Therefore, learning that a social media post contains "disinformation" does not appear to produce a "deontological judgement"[25], where the sharing of any disinformation would be judged equally regardless of any potential impact. Instead, the findings here indicate that individuals may make moral evaluations of disinformation on a case-by-case basis. This may mean that it is possible to view accuracy as important and think disinformation is "wrong" to spread; however, the extent to which these standards are upheld may not always be consistent.

9.4.2. Group-Directed Threats: Motivated to Question Content Accuracy

As people are motivated to "downplay the credibility" of group-threatening information to minimise any negative impact to the self (Ellemers et al., 2002, p. 177), users may automatically consider accuracy when they encounter group-threatening misinformation. As study four indicates, this may include misinformation that undermines an ingroup. By analysing participants' free-text responses using the Extended Moral Foundations Dictionary (F. R. Hopp et al., 2021) it was shown that participants in

[25] Deontological judgements focus on the morality of the action itself, rather than the potential consequences. In this instance, the act of spreading disinformation would need to be seen as "wrong" to do, regardless of whether it may lead to a positive outcome for the individual.

conditions presenting ingroup undermining misinformation (and disinformation) were more likely to consider "fairness" related words (such as "fake" and "lie") than participants who viewed ingroup supporting misinformation. In turn, increased considerations of "fairness" (e.g. higher fairness domain score) were related to harsher moral judgements. Moreover, according to the principles of Moral Foundations Theory (Graham et al., 2011), a greater tendency to value ingroup loyalty should predict increased sensitivity to ingroup-directed threats. Here it was found that valuing ingroup loyalty was related to increased engagement with "fairness" considerations when presented with ingroup undermining misinformation. These findings indicate that ingroup undermining misinformation (and disinformation) may be perceived as group-directed threats and, as such, may automatically attract considerations of accuracy.

Notably, previous research suggests people may be more likely to judge politically incongruent headlines as "false" (regardless of veracity), an effect that appears to be unaffected by tendency to engage with deliberative reasoning (Batailler et al., 2021). Within SMPs, users may also be more likely to scrutinise content that undermines rather than supports an ingroup (Huntington, 2020). They may also be more likely to associate the term "fake news" with outgroup media sources (Axt et al., 2020; Tong et al., 2020; van der Linden et al., 2020), who may at times publish information that is seen to threaten the ingroup. As study four in the present thesis demonstrates, awareness of information being false may not be required to make these kinds of associations. Together, these findings indicate that people are likely to consider accuracy-related concepts on SMPs (and potentially even deliberate over the accuracy of content), but not necessarily in a way that is useful for reducing disinformation spread.

9.4.3. No Threat: Content May Provide Users with Opportunities for Self-Expression

While people may readily consider accuracy when presented with group-threatening information, findings within this thesis indicate that ingroup supporting

misinformation may not present such a clear identity-threat. The first is that participants who viewed misinformation that positively framed their ingroup were least likely to consider "fairness" when making their judgements (study four). Rather than using words such as "evidence" or "facts", they often tended to use words such as "good" or "positive". The difference between fairness domain scores when viewing ingroup supporting misinformation (compared to in other conditions) might indicate that evaluations were contextually different from those made in other conditions. As moral acceptability scores for ingroup supporting misinformation were also notably higher than for the other conditions, it may be that participants did not perceive potential issues (e.g. a "fairness" violation) with this content.

Another indicator was that strong identifiers were more likely to spread identity-beneficial misinformation. While motivations for expressing identity can change in response to identity-threat (Ellemers et al., 2002), when there is no perceived identity threat strong identifiers may be motivated to express their identity to achieve or maintain a positive self-concept (Tajfel & Turner, 2004). Indeed, both studies four and five show evidence that strong identifiers were more likely to spread ingroup supporting misinformation than other people, but only within this context. This adds to previous work suggesting that strong identifiers may be more likely to believe (Anthony & Moulding, 2019; Sanchez & Dunning, 2021) and spread (Osmundsen et al., 2021) identity-affirming misinformation.

Given that strong identifiers may be more motivated to spread identity-beneficial content generally (e.g. not necessarily misinformation), it may be that misinformation that does not present a clear identity-threat is treated much like any other information on social media. For instance, people are more likely to spread misinformation shared by people they trust (Bringula et al., 2022; Buchanan & Benson, 2019; Sterrett et al., 2019), that is about topics they care about (Sterrett et al., 2019) and as a means of expressing an opinion (X. Chen et al., 2015; Schaewitz et al., 2020). SMPs are, after all, environments designed

with personal expression, and the development and maintenance of interpersonal relationships in mind. One reason that people amplify misinformation may therefore be because it blends in amongst a user's feed in a way that is difficult to detect.

9.4.4. Self-Directed Threat: Potentially "Useful" Content May Present a Dilemma

While users may have prosocial motivations for sharing identity-benefitting misinformation, the present findings indicate that learning the content is inaccurate (e.g. disinformation) can influence evaluations and behaviour. Indeed, identity-beneficial disinformation (e.g. supported an ingroup and / or undermined an outgroup) was judged as less acceptable to spread than beneficial misinformation (e.g. studies two, four & five). Knowing information is or may be false can therefore have an impact. Yet, as seen in both studies two and four, the act of spreading identity-beneficial disinformation was judged as more acceptable than spreading disinformation (and misinformation) that undermined the ingroup.

There may therefore be moral ambiguity surrounding spreading disinformation that could otherwise be useful achieving certain goals (e.g. framing the ingroup in a positive light). Unlike spreading group-threatening information (which studies two and four indicate may be relatively unacceptable), identity-beneficial disinformation may not directly target the value of the group and therefore may not be judged as harshly. Such leniency may also produce a moral dilemma (where an individual's desires or needs conflict with the desires or needs of others), leaving individuals to weigh up perceived outcomes. Such "consequential judgements" may explain why identity-beneficial disinformation was judged as more acceptable to spread than disinformation that undermined the ingroup.

However, a person's ability to spread identity-beneficial disinformation does still require a level of willingness to potentially deceive others (e.g. a key source of the self-directed threat) to achieve an otherwise beneficial outcome for the self or ingroup. As

such, when disinformation does not directly threaten the value of an ingroup, any potential benefits may be weighed up against the risk of negative consequences for the self. This may include exclusion from an ingroup if fellow group members judge the act to be antisocial. As such, in the face of self-directed threats, highly committed individuals may be motivated to behave in a prototypical manner to avoid rejection (Ellemers et al., 2002).

The findings in studies four and five do suggest that identity-beneficial disinformation has the potential to present such a self-directed threat (even when expressing ingroup "love"). As previously noted, strong identifiers appeared to be more likely to spread identity-beneficial misinformation (e.g. no identity threat) than other people. Yet, when the same information featured a fact-check or accuracy prompt tag there was no significant relationship between strength of identity (SOI) and spread. Again, SOI had no significant relationship with moral evaluations. If users think that such warning "tags" on identity-beneficial disinformation content may also be seen by others, this could threaten the self to the extent that it outweighs any perceived benefits gained from spreading it further. Indeed, previous work indicates people avoid spreading disinformation to protect their reputation (Altay et al., 2020). Consequently, refraining from interacting may likely be the safest option to maintain a positive self-concept.

Similarly, certain narratives may produce a self-directed threat (without the need for warning-tags). In the present thesis, perceived risk of COVID-19 did not predict intentions to spread misinformation that sought to minimise the severity of the virus (study one). However, lower perceived risk did predict more lenient moral judgements of spreading the content after learning it was untrue. The overall lower interaction and moral acceptability rates indicate that participants may have intended to refrain from spreading "minimising" misinformation, even when it was consistent with their beliefs. While it is not possible to ascertain here whether participants felt others may be critical of this content, previous work suggests users self-censor online when they perceive their views to be in the minority (H. T. Chen, 2018; Fox & Holt, 2018; Woong Yun & Park, 2011; Wu &

Atkin, 2018). Research also suggests social media users may adapt their behaviour to conform to perceived norms (e.g. Bradshaw et al., 2021; Colliander, 2019; Jahng et al., 2021; Woong Yun & Park, 2011). Indeed, groups play an important role in regulating social behaviour and so, even when a person feels that an act is acceptable, what the group thinks matters (Ellemers, 2017). It is also suggested people are more lenient towards group-benefitting deceptions, but more critical towards self-benefitting lies (Fu et al., 2008). Certain narratives in belief-consistent misinformation may therefore also present a self-directed threat: for instance, where the association with disinformation is well-publicised.

9.5. Research Question 4: Do Moral Processes Play a Role in the Contribution to Disinformation Spread?

Much research in this area has focused on understanding how cognitive processes influence susceptibility to and intentions to spread misinformation. Yet, such work has primarily focused on the identification of disinformation. However, as the findings in the present thesis demonstrate, people may be more willing to spread information they know to be inaccurate than information they feel is morally "wrong" to share. Given people often care more about being perceived as "moral" than "correct" (Ellemers et al., 2008) then moral cognition may be important for guiding their behaviour within digital social environments. The relationship between moral evaluations of misinformation and intentions to spread the content are therefore discussed, before looking at how moral cognition may help explain the evaluations in the first instance.

9.5.1. Levels of Acceptability Guide Users' Intentions to Spread Misinformation

Moral evaluations help guide behaviour (Bandura, 1991b), which may explain why people are more likely to spread misinformation when they feel it is morally acceptable to do so (Effron & Raj, 2020; Helgason & Effron, 2022). In the present thesis, this relationship was confirmed across three studies (studies three to five), using different target

populations (e.g. social media users in England, English premier league fans, and US-based Democrat and Republican voters) and themes of misinformation (e.g. political and football). As previously discussed, changes in context (e.g. misinformation becoming disinformation) can impact levels of perceived moral acceptability. Such changes were also related to differences in intentions to spread (studies four & five). Additionally, moral acceptability was found to mediate any relationship with spread (sometimes entirely). Individuals may therefore be guided by their moral evaluations when choosing to interact with misinformation, and their decisions to spread are also sensitive to situation-based adjustments in moral evaluations.

9.5.2. Possible Reliance on Moral Intuition to Guide Evaluations of Content

The present thesis provides several indications that people may rely on moral intuitions (e.g. feeling a sense of "right" or "wrong") to help determine the appropriateness of spreading misinformation further. Furthermore, that these moral intuitions may be guided in part by the previously discussed identity threats. Moral intuitions are rapid, affective processes that can quickly alert individuals to potential moral violations and encourage them to respond appropriately (Haidt, 2001). They may not always be consciously experienced, but at their strongest, moral intuitions may produce moral emotions (Haidt & Kesebir, 2010). Notably, moral intuitions may also be prompted without any active reasoning (Haidt, 2001). Therefore, people may be unaware of the automatic, moral evaluations of social media content that they are constantly making, until a feature prompts a consciously experienced response.

The first indication that this may be the case is that "non-threatening" misinformation was generally judged as acceptable to spread, however, "group-threatening" misinformation was much less so (studies 2 and 4). As previously discussed, encountering identity-threats may produce negative affect (Ellemers et al., 2002). Furthermore, seeing others (including the ingroup) being treated unfairly can encourage

moral emotions such as anger, disgust, and contempt, all believed to motivate punishment[26] (see FeldmanHall et al., 2018). Viewing group-threatening information may therefore prompt negative affect, that, if strong, could be interpreted as a sense of "wrongness". Indeed, even disinformation that supported the ingroup was judged as more acceptable to spread than misinformation that undermined the ingroup (study four). Simply put, intuitive processes may lead ingroup-undermining information to be automatically evaluated as being less moral, regardless of whether it is inaccurate or not.

Another indicator of reliance on moral intuition to guide evaluations of misinformation was how participants made their moral judgements when there were no perceived "threats". It has been argued that people make constant evaluations of people, situations and objects, and therefore moral intuitions may allow individuals to make rapid evaluations (see Haidt & Kesebir, 2010) that arguably may keep a person safe. However, moral intuitions do not always manifest in a person's consciousness (Haidt, 2001). It is also thought that people make rapid assessments about the morality of intended behaviour, both against personal moral standards and in the context of the group (Bandura, 1991b). Therefore, unless they sense that behaviour may be considered "wrong", people may feel relatively free to engage in an act. For instance, as previously discussed, strong identifiers (who are generally more motivated to express their identity (Tajfel & Turner, 2004)) were only more likely to spread identity-beneficial misinformation (studies 4 and 5). This was also the condition with the highest levels of moral acceptability overall, where responses generally sat towards the very top end of the scale.

This might suggest that, on the whole, participants did not sense there was anything "wrong" with spreading certain types of misinformation. Without affective cues indicating

[26] Notably, these are emotions often associated with responses to misinformation (Barfar, 2019; Pulido et al., 2020; Vosoughi et al., 2018). Whether the content depicts a group-directed threat or contains information that directly threatens the value of the ingroup may be important for distinguishing response. Tackling the former may encourage amplification, whereas the opposite has been shown here for the latter.

otherwise, identity-beneficial misinformation may feel "safe" to spread. In contrast, identity-beneficial disinformation was judged as less morally acceptable to spread. The inclusion of fact-checks or accuracy nudges also appeared to remove the significant influence of SOI on spread intentions. It may therefore be that affective responses after viewing such interventions disrupt the "safe" feeling, not necessarily through promoting deliberation, but perhaps due to prompting an intuitive sense of "wrongness" that may guide user's spread-related behaviour[27]. Indeed, previous work has suggested that one reason that individuals give for reporting misinformation is a "funny feeling" (Gimpel et al., 2021), indicating that, at the very least, decisions to intervene in the spread of content may be reliant on having experienced some kind of conscious cue.

Yet it is unlikely that identity-beneficial disinformation elicits the same intuitive response as group-threatening content. Indeed, self-directed threats are thought to be more closely related to anticipatory moral emotions such as guilt and shame, which may encourage norm-compliance (see FeldmanHall et al., 2018). Previous work has also shown anticipated guilt to be important for predicting unethical behaviour online (T. Kim et al., 2022; X. Wang & McClung, 2012). Therefore, if people experience a sense of "wrongness" when faced with disinformation, there may be important psychological distinctions in how such intuitions manifest that may help to further explain the findings from this thesis.

9.5.3. Moral Reasoning May Help or Hinder Disinformation Spread

When identity-beneficial disinformation produces a moral dilemma, individuals are arguably faced with several choices. They may feel it is not worthwhile to engage in more effortful reasoning processes, and simply scroll on. Alternatively, they may weigh up the moral arguments on both sides using reasoning processes. This could be perceived as a

---

[27] This is somewhat consistent with Pennycook's "inattention account", which suggests users may be more likely to spread misinformation when inattentive to accuracy and that such interventions draw focus back to "accuracy" (Pennycook, McPhetres, et al., 2020; Pennycook, Epstein, et al., 2021).

positive, given that previous work suggests that susceptibility to misinformation is due to a lack of reasoning (e.g. Pennycook & Rand, 2019). However, in the context of moral thinking, reasoning processes have been linked to an increased likelihood of consequential (e.g. "greater good") judgements (J. D. Greene et al., 2008; Paxton et al., 2012) and moral hypocrisy (Valdesolo & DeSteno, 2008). Therefore, engagement with moral reasoning has the potential to lead individuals to make decisions that favour spreading disinformation above competing outcomes.

For instance, when individuals are presented with belief-consistent disinformation, they may be able to utilise the sense that it feels "generally true" to rationalise an argument for it being acceptable to spread further (study one). In doing this they may also be able to protect the self-concept (e.g. their "moral self"). Indeed, previous EEG research found people allocate greater cognitive resources when viewing belief-consistent disinformation, and still continue to uphold said beliefs when asked to make accuracy judgements (Moravec et al., 2019). Arguably, even if people experience discomfort at the thought of spreading disinformation, moral reasoning may allow them to rationalise its spread in a way that helps to maintain a positive self-concept. While moral reasoning may of course help reduce the likelihood of engaging in antisocial behaviour online in some contexts (e.g. Wang et al., 2016), further work is required to better understand its potential impact on disinformation spread.

9.6. Other Findings

9.6.1. Self-Directed Threats Could Help Explain Political Asymmetry

To date, several studies have observed a potential relationship between political ideology and susceptibility to misinformation (Baptista et al., 2021; De Keersmaecker & Roets, 2019; Ecker & Ang, 2019; Garrett & Bond, 2021; A. Guess et al., 2019). However, others have argued that alternative explanations, such as a greater availability of identified conservative-leaning misinformation may explain previous findings (Garrett & Bond,

2021; A. Guess et al., 2019). Across the present thesis, asymmetries based on political orientation were observed for moral judgements of disinformation but not for misinformation.

Similarly to studies two, four & five, previous experimental research showed participants misinformation that was customised to their identity (Ryan & Aziz, 2021). As in, other than ingroup name, participants saw the same post regardless of their group affiliation. They found no notable difference between Republicans and Democrats in their levels of belief in misinformation statements. This corresponds with findings in the present thesis, where participants appeared to make biased evaluations of identity-relevant misinformation, regardless of their political orientation.

Where differences did occur, however, was in the context of disinformation. For instance, in study two, after learning that a previously viewed post was false, Conservative voters continued to make moral judgements that reflected ingroup bias, whereas Labour voters did not. While these findings somewhat conflict with previous work where Democrat voters made biased moral judgements in favour of ingroup politicians who lie (De Keersmaecker & Roets, 2019), prior work indicates this may be explained by cultural differences (as well as who is thought to be engaging in the deception). Indeed, research suggests that US voters may be more lenient towards politicians lying (Swire et al., 2017; Swire-Thompson et al., 2020) in a way that voters in other countries (including the UK) may not (Aird et al., 2018; Prike et al., 2023). Aird et al. (2018) suggests this may be because politics in the US are more polarised than elsewhere. Therefore, the current political climate may also influence how people make judgements of disinformation.

However, study five of the present thesis was carried out in the US. Here, showing Democrat voters an accuracy nudge had no influence on their moral judgements of ingroup-beneficial misinformation (only spread). In the context of De Keersmaecker & Roets' (2019) findings, it may be that Democrat voters are able to dismiss potential inaccuracies when making their moral evaluations of ingroup-beneficial misinformation,

but factor in any potential self-directed threats (e.g. other people's evaluations) when deciding on whether to spread[28]. In contrast, both moral appeals appeared to help reduce Democrats' moral judgements of spreading disinformation (and in turn, influenced intentions to spread). Such appeals may potentially work by reminding political liberals within politically polarised climates of their existing "fairness" values in a way that accuracy nudges cannot.

However, despite tailoring an appeal for Republican voters, neither moral appeal influenced their judgements. In support of previous findings (e.g. Rathje et al., 2022; Roozenbeek et al., 2021), the use of an accuracy nudge was not effective for reducing intentions to spread in Republican voters. As previous research found that political conservatives were more willing to evaluate photographic evidence in a way that prioritised the ingroup above presented reality (Schaffner & Luks, 2018), there may be important differences related to political orientation that influence evaluations of false information that could otherwise benefit the ingroup.

Furthermore, previous research has suggested political conservatives may generally be more morally lenient towards politicians' lying (De Keersmaecker & Roets, 2019), and more likely to spread "fake news" (Baptista et al., 2021). Yet, there was no indication to suggest that was the case here. Indeed, the findings from study one indicates that Conservative voters were less likely to interact with misinformation that negatively framed the UK Government (e.g. their ingroup) than Labour voters. They also felt it was less acceptable to spread the "unfavourable" disinformation compared to Labour voters. Moreover, unlike De Keersmaecker & Roets (2019), Conservative voters were also less accepting of spreading disinformation framed to support an outgroup party compared to

[28] Given individuals wish to be seen as "moral" by other people (Ellemers et al., 2008; Pagliaro et al., 2011), they may hold differing standards for their own behaviour than for others (e.g. politicians). Whereas, research suggests US voters may expect politicians to lie (Swire-Thompson et al., 2020).

disinformation supporting the ingroup or undermining said outgroup (study two). Given that data collection for that study was carried out the day before an election, any leniency political conservatives may have towards spreading disinformation may not apply when doing so is at the ingroup's expense.

Finally, to rule out the possibility of partisanship-related norms, study three presented fans of five football teams with football-related disinformation. Again, a relationship between political orientation and moral judgements was found. Specifically, right-leaning participants were more accepting of sharing identity-beneficial disinformation, and this was partially explained by lower tendency to engage with the "fairness" domain. Overall, the findings suggest that there may be underlying differences in how political conservatives and liberals evaluate identity-beneficial disinformation that extend beyond partisanship.

9.6.2. Moral Foundations Theory and Disinformation

Previous work has suggested that political ideology is related to meaningful differences in moral values (Graham et al., 2009). Specifically, Moral Foundations Theory (MFT) suggests that moral reasoning occurs in relation to various, discrete moral "foundations" which are supposedly universal (Graham et al., 2011). Research suggests the tendency to engage with each foundation may differ across the ideological spectrum (Graham et al., 2009). Moreover, recent studies suggest an increased tendency to engage with "binding" foundations[29] ("Loyalty", "Authority" and "Purity") may predict increased susceptibility to misinformation (Ansani et al., 2021; Piejka & Okruszek, 2020; Trevors & Duffy, 2020). However, exploratory analysis in study four does not support these findings, as no significant relationships between scores on the Moral Foundations Questionnaire (MFQ), and overall intentions to spread or moral judgments were found.

[29] This collection of "binding foundations" also tend to be more associated with political conservatives (Graham et al., 2009).

However, the findings across the present thesis illustrate how moral evaluations of misinformation can be situational. MFT initially proposed a modular approach to moral thinking (Haidt & Joseph, 2008). As in, higher MFQ scores on a particular foundation would indicate an increased tendency to engage in foundation-related cognition when said foundation was triggered (e.g. loyalty judgements would occur in a discrete "loyalty" location when prompted by a loyalty cue). However, again, the findings do not support this. In fact, when viewing ingroup-undermining misinformation, tendency to engage with loyalty foundations had a negative relationship with engagement with the "loyalty" domain (study four). In other words, they were less likely to use "loyalty" related words. This suggests that judgements for those who more readily prioritise ingroup loyalty were unlikely to occur exclusively within the proposed "loyalty" domain.

Yet, there was evidence suggesting certain "foundations" may be of interest within specific contexts. In study four, higher fairness MFQ scores were related to increased use of "fairness" related words, but only when presented with ingroup supporting disinformation. As in, those who prioritise fairness may be more likely than others to consider fairness when presented with potentially useful disinformation. Moreover, higher MFQ loyalty scores were also related to increased engagement with the "fairness" domain, but when presented with information that made their ingroup look bad (e.g. ingroup undermining misinformation). This is notable, given individuals who value loyalty should respond more readily to a group-directed threats (Graham et al., 2011; Haidt & Joseph, 2008).

However, framing the spread of disinformation as a potential binding violation was not enough to encourage Republican voters (who tend to more readily prioritise binding foundations such as loyalty than liberals (Graham et al., 2009)) to adjust their moral judgements. Yet, previous moral reframing interventions have effectively influenced beliefs and behaviour of political conservatives in relation to climate change (Feinberg & Willer, 2013; Hurst & Stern, 2020; Wolsko et al., 2016). As the present findings indicate

political orientation may influence moral evaluations of ingroup beneficial disinformation (study four) rather than disinformation generally (study one), appeals to binding values (including ingroup loyalty) may be ineffective at competing with disinformation that may assist in fulfilling a similar goal.

9.6.2.1 The Importance of "Fairness"

While it could of course be that the moral appeals in study five were not persuasive enough to help convince Republican voters of any moral violations, the issue itself (e.g. "disinformation) and its association with "fairness" values could be important to understanding why users spread disinformation. Notably, "fairness" values can be variably applied (Haidt & Graham, 2007) and may be overridden by competing concerns (e.g. Shaw et al., 2012; Waytz et al., 2013). This is likely to present problems in the context of disinformation. For instance, people may more easily dismiss fairness concerns when viewing belief-consistent disinformation, as well as other disinformation that potentially "feels" true (e.g. Effron & Raj, 2020). As political conservatives also tend to value all five moral foundations to a similar degree (Graham et al., 2009), the greater ease to which fairness may be overridden could also explain why the moral appeals were not effective for Republican voters in study five. Arguably, identity-beneficial misinformation may itself present more appealing loyalty cues that simply outweigh "fairness" concerns. However, in other contexts (for instance, when misinformation threatens an ingroup) the expectation may be for others (e.g. those who may wish to spread the content) to act "fairly" (as in by not spreading it further). This may explain why participants who prioritised loyalty were more likely to make justifications based on "fairness" when they saw ingroup-undermining misinformation (study four). Therefore "fairness" may be a key reason for not spreading disinformation, but a willingness to tolerate unfair behaviour may also explain why users are willing to spread it.

Furthermore, it has been previously argued that liberals and conservatives have distinct moral belief systems. Unlike conservatives, liberals appear to have a greater tendency to engage with individualising foundations (e.g. fairness and harm) compared to binding foundations (e.g. loyalty, sanctity, purity) (Graham et al., 2009). If liberals can prioritise "fairness" concerns when presented with identity-beneficial disinformation, then this may help explain the political asymmetries seen here. Indeed, previous work has shown a relationship between fairness concerns and whistleblowing in the face of competing loyalty values (Waytz et al., 2013). Notably, as fairness concerns encourage people to engage in reciprocal altruism (Haidt & Graham, 2007), people chose to forgo any potential benefits of spreading identity-beneficial disinformation to act in a "fair" way.

A need to act "fairly" may also influence how people spread belief-consistent disinformation. In the present thesis, for instance, it was found that belief consistency did not predict moral judgements of disinformation attempting to maximise the threat of COVID-19 (study one). However, research also suggests tendency to prioritise fairness has also been associated with likelihood of conforming to COVID-19 restrictions (Chan, 2021). If, as the findings here suggest, spreading disinformation may be associated with violating "fairness" values then it might help explain why belief-consistency was not a significant predictor in this context. For some participants, the decision may not have been straightforward and therefore may have produced a moral dilemma. If so, this may mean that while people who prioritise fairness may tend to abstain from spreading identity-relevant disinformation in some contexts, they may be themselves vulnerable to fairness-related disinformation narratives. Given that disseminators of disinformation have previously been observed presenting themselves as activists seeking to achieve equality (François et al., 2019), understanding how users resolve moral dilemmas presented by fairness-related disinformation may be an important avenue for future research.

Finally, the present thesis found that liberals were as likely as conservatives to make biased moral judgements of misinformation (studies two and four). Previous work

has suggested that moral foundations may act as cognitive schemas, drawing attention to relevant cues, encoding (e.g. adjusting moral evaluations), and triggering relevant behaviour (Süssenbach et al., 2019). They also demonstrated using eye-tracking that tendency to engage with "harm" foundations was related to greater attention to harm violations. However, there was no relationship between harm foundation engagement and attention when harm cues were not present, suggesting the relevant foundation must be made salient. Therefore, unless misinformation presents a fairness-related "cue", liberals would be no more likely than others to consider fairness. Indeed, this is what was found in study four when participants viewed ingroup-supporting misinformation.

By making "fairness" salient, the moral appeals in study five may have helped draw the attention of participants who more readily prioritise fairness (in this instance, Democrat voters) to potential "fairness" violations in identity-beneficial misinformation. In turn, this may have helped to reduce their moral judgements, and subsequently, intentions to spread the misinformation further. Concerningly, however, Süssenbach et al. (2019) also found that people who prioritise binding foundations may avoid giving attention to individualising-related cues. Therefore, if disinformation is associated with "fairness", then high binders may also avoid certain references to disinformation. This may also explain why the moral appeals and accuracy nudges in study five had no influence on Republican voters (who tend to prioritise binding foundations more readily) for instance. The potential implications of this on current interventions will be discussed shortly.

9.6.3. Methodological Implications

9.6.3.1 Belief Consistency or Specific Beliefs

The present thesis indicates potential methodological issues arising from looking at beliefs (or belief-related categories such as group membership, ideological stance) in relation to misinformation susceptibility. Adding to the existing research in this area (e.g. A. Kim et al., 2019; A. Kim & Dennis, 2019; Schaewitz et al., 2020; Tsang, 2020; Vegetti

& Mancosu, 2020), studies one and three demonstrated that the level of consistency between a person's beliefs and what is expressed by misinformation can influence users' evaluations and intentions to spread the content further.

Specifically, studies one and three add to the existing literature in a number of ways. Firstly, a number of papers have previously used political orientation scales (Buchanan, 2020; Faragó et al., 2020; Helmus et al., 2020) and political party affiliation (Helgason & Effron, 2022) to represent political "beliefs". However, the specific beliefs held by individuals at each point across the political spectrum (and indeed within a party) will differ. Such heterogeneity is important to acknowledge in the context of understanding user interactions with disinformation, as personal "relevance" will likely be important. Additionally, some disinformation campaigns have attempted to create division within groups by targeting differences in beliefs (Barry, 2022) and therefore broad political views may not best account for interactions in this context.

Indeed, previous research looking at real-life Twitter interactions with disinformation has demonstrated the benefits of focusing on specific groups situated within broad-ideological categories for understanding user-interactions (Freelon et al., 2022). Others have also shown that certain political-attitudes (e.g. towards abortion) may influence intentions to interaction with relevant misinformation (A. Kim et al., 2019; A. Kim & Dennis, 2019). However, such attitudes tend to be established and therefore may be relatively stable. As disseminators of disinformation may swiftly take advantage of crisis situations, there is arguably also a need for research on more recently established and less stable beliefs. The COVID-19 pandemic is an example of such a crisis, and therefore arguably the beliefs measured in studies one and three were relatively new compared to beliefs formed over decades or even a lifetime.

Furthermore, a number of studies have previously employed beliefs as predictors of misinformation susceptibility generally, i.e. where beliefs have been presented as one-way predictors of misinformation susceptibility (Roozenbeek et al., 2020; Saling et al., 2021;

Scherer et al., 2021). As the present findings illustrate, there may be a chance that stimuli presented to participants may simply have been more consistent with any significant beliefs. Therefore, unless belief-consistency is controlled for in some way, it may be difficult to distinguish between true indicators of potential vulnerability and the effects of belief-consistency.

The findings also add to recent work demonstrating the importance of moving beyond broad categories of "misinformation" by drawing focus onto specific narratives and topics (e.g. Freelon et al., 2022; Hameleers et al., 2021; Morosoli et al., 2022). Arguably, as the present findings suggest, expressing that specific beliefs increase the likelihood of spreading "misinformation" (as is the norm) has the potential to cause real issues not least in terms of replicability. Had studies one and three only presented participants with misinformation that undermined the UK Government, it would not be reasonable to claim that low trust increases the likelihood of spreading "misinformation" because, as illustrated here, that was not the case for three other categories of misinformation. Not only is this a logical step from a methodological viewpoint, but it may also help to emphasise the need to expand the focus of misinformation research beyond current dominant narratives (e.g. US partisanship, COVID-19, etc).

9.6.3.2 Benefits of Developing Stimuli to Test Group-Differences

While stimuli in two studies in the present thesis consisted of content (containing false or misleading information) taken from social media, for three studies the materials were developed and created to facilitate the use of experimental designs. These were based on real-life disinformation narratives and revealed useful insights into how situations presented by misinformation (e.g. identity-threats) may influence moral evaluations, as well as ideological differences in judgements of disinformation. While the use of real-world disinformation as stimuli is important for generalisability (Pennycook, Binnendyk, et al., 2021), it would not be possible to test for such differences through their use alone.

9.6.3.3 Social Media Spread Scale

A scale was developed during this thesis which aimed to measure people's intentions to contribute to the spread of a piece of social media content. Given the variety of ways that users can help to amplify content on social media, focusing on only single actions may arguably produce challenges in regard to power and generalisability. As seen in study one, previous Likert-based measures that only consider responses that amplify content can also present challenges such as faux floor effects and make it difficult to distinguish attempts to intervene from passive social media use. These challenges were overcome by developing a novel scale that also incorporates potential steps to minimise the spread of social media content, which was used across studies three to five and showed good reliability.

9.6.4. Implications for Current Interventions

The findings in the present thesis indicate that people may not make evaluations of social media content in a consistent way, which may have implications for current interventions. Firstly, while the present findings provide some support for the argument that individuals spread misinformation because they do not consider accuracy, they also indicate that there may be instances where individuals may not be motivated to. Arguably, questioning the accuracy of certain types of information (e.g. ingroup supporting) could potentially threaten identity if it does prove to be untrue. As people are motivated to avoid cognitive discomfort (see McGrath, 2017 for a review) certain situations may require incentives to successfully encourage objective accuracy-based objectives.

Another consideration is that much of the relevant psychological research around accuracy assessments of misinformation (that have informed the development of interventions) is based on assessments of "plausibility" (e.g. Pennycook & Rand, 2019). For instance, it is thought that individuals who have a greater tendency to engage with deliberation may use plausibility cues to distinguish between "real" and "fake" headlines.

Such studies potentially present stimuli where "real" information is deemed more plausible than the presented "fake" headlines. Rather than improving individuals' ability to identify misinformation, deliberation may lead individuals to rely on a "sense" of plausibility (suggesting even high deliberators may rely on some kind of "feeling"). There is even evidence to suggest that such approaches may only help to improve the identification of "true" headlines, rather than necessarily aiding individuals with identifying misinformation (Batailler et al., 2021; Martel et al., 2020). In other words, deliberation about the accuracy of social media content may help individuals overturn false alarms (Batailler et al., 2021), however, it may be ineffective for aiding in evaluations of misinformation with no identifiable cues (e.g. plausibility, identity-threat, moral violation, etc). Indeed, the present thesis adds to previous work suggesting that perceiving the "gist" of disinformation as true may increase how morally acceptable people think it is to share (Effron & Helgason, 2022). As much of the content individuals view on social media will be personally relevant, exclusive reliance on such cues to identify misinformation may have serious implications.

While previous work suggests that people do care about sharing accurate information online (Pennycook, Epstein, et al., 2021), considering why this is the case may be helpful for understanding the usefulness of interventions such as "accuracy nudges" and fact-check tags. Given that people are motivated to be seen, and see themselves, as "moral" (Ellemers et al., 2008; Pagliaro et al., 2011), individuals may be averse to sharing inaccurate information which has the potential to negatively impact an individual's moral identity (and potentially their social identities). As reputational concerns are one reason people avoid spreading misinformation (Altay et al., 2020), users may be hesitant to spread information they identify as "implausible" in case their audience also notices the "implausibility". Indeed, reputational concerns may explain why strong identifiers in the present work were only willing to spread ingroup-supporting misinformation that did not feature a fact-check or an accuracy nudge. While the fact-check was found to adjust moral

evaluations, the accuracy-nudge did not, suggesting that in some instances individuals may care more about protecting the self than whether information is inaccurate or not. Similarly, a recent study suggests that including a "misleading count" alongside other post metrics (e.g. like counts) may be effective at reducing intentions to spread (Pretus et al., 2022). This suggests there may be potential for harnessing "tags" and visible metrics in a way that potentially minimises identity-related motivations for spreading misinformation. As people tend to refrain from spreading their beliefs when they are perceived to be in the minority (H. T. Chen, 2018; Fox & Holt, 2018; Woong Yun & Park, 2011; Wu & Atkin, 2018), public awareness of disinformation narratives may also have a similar effect.

Finally, it was found that the efficacy of certain interventions such as "accuracy-nudges" and moral appeals may be influenced by a person's tendency to engage with certain moral foundations. Given the relationship between moral foundations and political ideology (Graham et al., 2013), this may help explain instances where misinformation interventions were found to be less effective for political conservatives (Rathje et al., 2022; Roozenbeek, Freeman, et al., 2021). People who more readily prioritise "fairness" may be more attentive to misinformation interventions to potentially avoid violating values they deem important, leading them to potentially update their evaluations of identity-beneficial misinformation in a way that helps to reduce spread intentions. However, while moral reframing did not appear to be effective for Republican voters, political conservatives did appear to care about spreading disinformation in certain circumstances (when it clearly undermined the value of the ingroup). Rather than attempting to appeal to pre-existing values and standards, there may be ways to help individuals (and groups) develop clearer moral standards and values around disinformation that could be particularly beneficial for those who do not readily prioritise "fairness" and during moral development.

9.6.5. Limitations

There are of course a number of potential limitations in the present thesis. One key limitation is that only "false or misleading" information was used within the studies, meaning that participants were not asked to make judgements of any "factual" information. As a result, it is not possible to ascertain the extent to which judgements made in relation to the misinformation and disinformation presented here may have differed from those of accurate information. However, arguments for including "factual" information in studies tend to focus on discernment (e.g. ability to distinguish between true and false information), and may not be necessary when focusing on misinformation exclusively (Pennycook, Binnendyk, et al., 2021). Moreover, in reality, disinformation can be difficult to detect and often includes information that is difficult to verify (Colley et al., 2020). Indeed, disinformation based on crime rates (the focus of stimuli in two studies here) is a key example of this. However, it should still be noted that claims and comparisons made within the present work only apply in relation to disinformation and misinformation specifically.

Another limitation is that the present studies utilised self-report data in the context of hypothetical behaviour. While this may not be a perfect approach, it does provide a greater level of insight and sensitivity that more "realistic" approaches currently cannot. For instance, it may have been possible to run the studies within a simulated social media feed, or potentially within a social media platform itself. However, firstly, only certain actions would be measurable within these contexts. Indeed previous work (e.g. Gimpel et al., 2021; Jones et al., 2021; Pennycook, Epstein, et al., 2021) using such approaches have focused primarily on one or two digital actions (e.g. reposting, reporting, etc). These are only some of the ways that users can contribute to the spread of disinformation. They would not necessarily be any better at answering the present research questions, which relate to why users "spread" disinformation. For instance, it would not be possible to capture whether a user has shared the post with users through private message or cross-

platform posting. Within a ‘real’ platform it would also not be possible to ascertain whether users made private attempts to have the post taken down. Indeed, for many actions only the user and the platform are likely to have visibility. While simulated social media feeds may be updated in the future to reflect the breadth of options, studies run within real social media platforms are unlikely to ever observe the full range of user-interactions.

Moreover, such alternative approaches often measure behaviour using dichotomous outcomes (e.g. “shared” vs “did not share”). While capturing “behaviour” may be preferable to a self-report scale measuring “likelihood” of engaging in such behaviour, it is also important to consider the context presented here. Interactions with content on social media are relatively uncommon and highly influenced by other factors (for instance, relationship with the original poster). Furthermore, there is a high likelihood that participants would be presented with stimuli that they would not usually encounter within their own algorithmically determined social media feeds. It is therefore difficult to know whether these “realistic” approaches would be a substantial improvement from a validity perspective.

A final limitation was that participants only rated one misinformation item in studies 2, 4 and 5. However, unlike in studies 1 and 3 (where participants rated 12 items) these were designed for the specific experimental manipulation. Furthermore, whether a digital environment contains heterogenous or homogenous information can influence evaluations of misinformation (Gill & Rojas, 2020) and may produce ideological asymmetry in susceptibility (Rhodes, 2022). Therefore, the potential impact of introducing additional items within this kind of experimental design needs careful consideration. Notably, the primary effects of interest were replicated here more than once using different stimuli, and within well-powered studies. The only exceptions are of course the work relating to the Moral Foundations Questionnaire, linguistic content analysis, and moral appeals. Care should therefore be taken when interpreting these findings.

9.6.6. Recommendations for Future Research

While on the whole it appears that the belief consistency of misinformation influences whether people intend to spread it further, this was not always the case. It was proposed here that people whose beliefs were consistent with misinformation that minimised the seriousness of COVID-19 (who also remained more likely to judge the content as morally acceptable to spread after learning it was untrue) may have intended to refrain from spreading the content due to social concerns. Indeed, there is a substantial body of working indicating that people refrain from sharing their beliefs online when perceiving that the majority do not share said beliefs (e.g. H. T. Chen, 2018; Fox & Holt, 2018; Woong Yun & Park, 2011; Wu & Atkin, 2018). However, the current work cannot determine whether that was indeed the case. Given the public are becoming increasingly aware of "misinformation", understanding how levels of perceived awareness of specific "misinformation narratives" impact users intention to spread belief-consistent misinformation may be useful for informing future awareness campaigns and training.

Furthermore, here people made harsher moral judgements and were more likely to make fairness-considerations when misinformation threatened the ingroup compared to when it did not. In line with previous work, it was therefore proposed that people could potentially make accuracy considerations when viewing group-threatening information (misinformation or otherwise). However, what is not known is whether identity-based considerations of "accuracy" may allow people to feel they are careful about the information they spread online. As in, if people automatically question the accuracy of information that threatens an ingroup, is this enough to make them feel that they objectively question the accuracy of information online. Future research may therefore wish to better understand how such judgements and decisions are made over the course of a user's time within a platform, and the impact this has on any self-perceptions regarding "careful" use in the context of misinformation.

Additionally, exploring how people resolve moral dilemmas presented by disinformation is likely to be a useful avenue for future enquiry for a number of reasons. Firstly, previous work has suggested that users may make themselves more vulnerable to spreading misinformation by not engaging with reasoning processes. Yet, moral reasoning processes may also help people rationalise engaging in behaviour they otherwise feel is immoral. As the present findings indicate that people can be morally lenient towards spreading certain types of identity-relevant disinformation (e.g. belief-consistent, identity-beneficial, etc), looking at the impact on deliberation within such scenarios could prove valuable. Moreover, moral dilemma-based approaches may prove useful for better understanding the circumstances under which people are more willing to spread disinformation, but particularly if doing so may threaten their social identity. This may provide important information about whether people can perceive spreading disinformation to help achieve a greater good, or if they simply do not care about (or potentially even experience) any resulting moral violations.

Finally, it may also be helpful to test whether real-life disinformation does produce moral dilemmas in their "intended audiences". While the present findings indicated that people who more strongly value "fairness" may be less likely to spread the identity-beneficial disinformation presented here, they may be more vulnerable to disinformation which expresses competing fairness values (such as condemning police brutality, etc). As such tactics are being employed by those who disseminate disinformation, it is important to understand whether even users who appear less willing to spread disinformation can be persuaded by bad actors to do so when framed as supporting "a greater good".

9.7. Conclusion

The present thesis sought to explore the influence of social identity on social media users' evaluations and intentions to spread disinformation. It makes several unique contributions to the literature. Firstly, studies one and three demonstrated that the greater

the consistency between an individual's specific beliefs about a current issue and opinions expressed within misinformation content, the more likely they are to spread it further on social media. This was demonstrated using a novel scale developed as part of this thesis. It was also shown that people may be more morally lenient about spreading belief-consistent disinformation; even when they know the information is untrue. This work adds to the literature by making important connections between previous work on intentions to spread misinformation that supports political attitudes and misinformation that is perceived as "more accurate". By focusing on less stable, issue-specific beliefs, the present work also significantly contributes to our understanding of why people may amplify disinformation disseminated during crisis and emergency situations, as well as rapidly changing social environments.

This thesis also supports recent work demonstrating that biased evaluations of disinformation may be important for understanding why users go on to spread it. However, it contributes further by demonstrating that moral evaluations of misinformation can be situational and may change in relation to the viewer's social identity. People may be less accepting of spreading false content (misinformation and disinformation) that undermines rather than benefits their ingroup (studies two and four). However, the present findings indicate this could be because such information threatens the value of the ingroup and, as such, may be more readily associated as a potential "fairness" violation (study four). Conversely, moral evaluations of ingroup benefitting misinformation and disinformation may instead be determined by whether spreading it is perceived as a threat to a user's position in the ingroup (e.g. a self-directed threat). It may therefore be that social media users may feel spreading false information is "wrong", but that the basis against which any judgements of false information are made may be context specific. As in, the reasons a user may refrain from spreading ingroup-undermining misinformation are not necessarily the same as for disinformation that benefits the ingroup.

Furthermore, the present thesis also contributes to our understanding of apparent political asymmetries in both disinformation spread and efficacy rates of prior misinformation interventions. Indeed, as demonstrated in studies four and five, there may be differences in how individuals across the political spectrum perceive otherwise-useful disinformation to be a self-directed threat (which in turn, may influence how morally acceptable they judge it to spread). This may help explain previous work that found differences in how members of different political groups evaluate disinformation (study two). The present thesis makes a significant contribution, however, by demonstrating that the person-level effect of political orientation is sustained in relation to non-political groups (study four), i.e. underlying moral processes related to political orientation may influence how people evaluate certain types of known disinformation. Finally, study five adds to previous research that found misinformation interventions to be less effective for right-leaning participants. The present work also makes another unique contribution by showing that tailoring misinformation interventions to focus on moral values more frequently associated with political conservatism may also be ineffective. Together, these findings may help explain why previous research has indicated a greater availability of politically conservative disinformation on social media and could also provide useful insight into what makes certain individuals more susceptible to spreading disinformation generally.

References

AAP Factcheck. (2020, April 1). Coronavirus: Italy's COVID-19 situation is dire but a photo of coffins is not from the coronavirus pandemic. *News.com.au.* https://www.news.com.au/lifestyle/health/health-problems/coronavirus-italys-covid19-situation-is-dire-but-a-photo-of-coffins-is-not-from-the-coronavirus-pandemic/news-story/46281d651b1cad273e034cdef75d443d

Abendroth, J., & Richter, T. (2020). Mere plausibility enhances comprehension: The role of plausibility in comprehending an unfamiliar scientific debate. Journal of *Educational Psychology, 113*(7), 1304–1322. https://doi.org/10.1037/edu0000651

Abrams, D., Marques, J., Bown, N., & Dougill, M. (2002). Anti-norm and pro-norm deviance in the bank and on the campus: Two experiments on subjective group dynamics. *Group Processes & Intergroup Relations, 5*(2), 163–182. https://doi.org/10.1177/1368430202005002922

Abrams, D., Marques, J. M., Bown, N., & Henson, M. (2000). Pro-norm and anti-norm deviance within and between groups. *Journal of Personality and Social Psychology, 78*(5), 906–912. https://doi.org/10.1037/0022-3514.78.5.906

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications, 5*(1), 15. https://doi.org/10.1057/s41599-019-0224-y

Adelman, L., & Dasgupta, N. (2019). Effect of threat and social identity on reactions to ingroup criticism: defensiveness, openness, and a remedy. *Personality and Social Psychology Bulletin, 45*(5), 740–753. https://doi.org/10.1177/0146167218796785

Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: Evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health, 21*(1). https://doi.org/10.1186/s12889-020-10103-x

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *The Journal of Applied Psychology, 90*(1), 94–107. https://doi.org/10.1037/0021-9010.90.1.94

Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences, 182*, Article 111074. https://doi.org/10.1016/j.paid.2021.111074

Ahmed, S., Chen, V. H. H., & Chib, A. I. (2021). Xenophobia in the time of a pandemic: social media use, stereotypes, and prejudice against immigrants during the COVID-19 crisis. *International Journal of Public Opinion Research, 33*(3), 637–653. https://doi.org/10.1093/ijpor/edab014

Aird, M. J., Ecker, U. K. H., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society Open Science, 5*(12), 180593. https://doi.org/10.1098/rsos.180593

Alderman, L. (2021, May 26). *Influencers Say They Were Urged to Criticize Pfizer Vaccine.* The New York Times. https://www.nytimes.com/2021/05/26/business/pfizer-vaccine-disinformation-influeners.html

Allen-Kinross, P. (2020a, September 28). Covid-19 has killed more people than obesity in the UK this year. *Full Fact.* https://fullfact.org/online/coronavirus-obesity-mortality/

Allen-Kinross, P. (2020b, December 4). Vaccine approval isn't quicker because of Brexit. *Full Fact.* https://fullfact.org/health/coronavirus-vaccine-brexit/

Ali, K., Li, C., Zain-ul-abdin, K., & Zaffar, M. A. (2022). Fake news on Facebook: Examining the impact of heuristic cues on perceived credibility and sharing intention. *Internet Research, 32*(1), 379–397. https://doi.org/10.1108/INTR-10-2019-0442

Ali, K., Zain-ul-abdin, K., Li, C., Johns, L., Ali, A. A., & Carcioppolo, N. (2019). Viruses going viral: impact of fear-arousing sensationalist social media messages on user engagement. *Science Communication, 41*(3), 314–338. https://doi.org/10.1177/1075547019846124

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

Allen, V. L., & Wilder, D. A. (1975). Categorization, belief similarity, and intergroup discrimination. *Journal of Personality and Social Psychology, 32*(6), 971–977. https://doi.org/10.1037/0022-3514.32.6.971

Allen-Kinross, P. (2020, May 20). *This viral UK Facebook post claiming 31 million job losses probably refers to the USA.* Full Fact. https://fullfact.org/online/31m-lose-jobs/

Al-Rawi, A. (2021). Political memes and fake news discourses on Instagram. *Media and Communication, 9*(1), 276–290. https://doi.org/10.17645/MAC.V9I1.3533

Altay, S., Hacquin, A.-S. S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society, 24*(6), 1303–1324. https://doi.org/10.1177/1461444820969893

Amira, K., Wright, J. C., & Goya-Tocchetto, D. (2021). In-group love versus out-group hate: which is more important to partisans and when? *Political Behavior, 43*(2), 473–494. https://doi.org/10.1007/s11109-019-09557-6

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: visual imagery and moral judgment. *Psychological Science, 23*(8), 861–868. https://doi.org/10.1177/0956797611434965

Amnesty International. (2022). *Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya* (ASA 16/5933/2022). https://www.amnesty.org/en/documents/asa16/5933/2022/en/

Andı, S., & Akesson, J. (2020). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism, 9*(1), 106–125. https://doi.org/10.1080/21670811.2020.1847674

Andreasen, M. B. (2021). 'Rapeable' and 'unrapeable' women: The portrayal of sexual violence in Internet memes about #MeToo. *Journal of Gender Studies, 30*(1), 102–113. https://doi.org/10.1080/09589236.2020.1833185

Ansani, A., Marini, M., Cecconi, C., Dragoni, D., Rinallo, E., Poggi, I., & Mallia, L. (2021). Analyzing the perceived utility of covid-19 countermeasures: The role of pronominalization, moral foundations, moral disengagement, fake news embracing, and health anxiety. *Psychological Reports, 125*(5), 2591–2622. https://doi.org/10.1177/00332941211027829

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

APA Dictionary of Psychology. (n.d.). Objective Reality. Retrieved 6 February 2023, from https://dictionary.apa.org/objective-reality

Apuke, O. D., & Omar, B. (2021). Social media affordances and information abundance: Enabling fake news sharing during the COVID-19 health crisis. *Health Informatics Journal, 27*(3). https://doi.org/10.1177/14604582211021470

Aronson, P., & Jaffal, I. (2022). Zoom memes for self-quaranteens: generational humor, identity, and conflict during the pandemic. *Emerging Adulthood, 10*(2), 519–533. https://doi.org/10.1177/21676968211058513

Ask, K., & Abidin, C. (2018). My life is a mess: Self-deprecating relatability and collective identities in the memification of student issues. *Information Communication and Society, 21*(6), 834–850. https://doi.org/10.1080/1369118X.2018.1437204

Axt, J. R., Landau, M. J., & Kay, A. C. (2020). The psychological appeal of fake-news attributions. *Psychological Science, 31*(7), 848–857. https://doi.org/10.1177/0956797620922785

Baek, Y. M., Kang, H., & Kim, S. (2019). Fake news should be regulated because it influences both "others" and "me": How and why the influence of presumed influence model should be extended. *Mass Communication and Society, 22*(3), 301–323. https://doi.org/10.1080/15205436.2018.1562076

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

Bandura, A. (1991a). Social cognitive theory of moral thought & action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (Vol. 1, pp. 69–128). Lawrence Erlbaum Associates, Inc.

Bandura, A. (1991b). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Baptista, J. P., Correia, E., Gradim, A., & Piñeiro-Naval, V. (2021). The influence of political ideology on fake news belief: The Portuguese case. *Publications, 9*(2), 23. https://doi.org/10.3390/publications9020023

Barber, A. (2020). Lying, Misleading, and Dishonesty. *The Journal of Ethics, 24*(2), 141–164. https://doi.org/10.1007/s10892-019-09314-1

Barfar, A. (2019). Cognitive and affective responses to political disinformation in Facebook. *Computers in Human Behavior, 101*, 173–179. https://doi.org/10.1016/j.chb.2019.07.026

Barry, E. (2022, September 18). *How Russian trolls helped keep the women's march out of lock step.* The New York Times. https://www.nytimes.com/2022/09/18/us/womens-march-russia-trump.html

Bar-Tal, D. (1998). Group beliefs as an expression of social identity. In S. Worchel, J. F. Morales, D. Páez & J. C. Deschamps (Eds.), *Social identity: International perspectives* (pp. 93–113). Sage Publications, Inc. https://doi.org/10.4135/9781446279205.n7

Basol, M., Roozenbeek, J., & Van Der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition, 3*(1), 2. https://doi.org/10.5334/joc.91

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2021). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Bauer, P. C., & Clemm von Hohenberg, B. (2021). Believing and sharing information by fake sources: An experiment. *Political Communication, 38*(6), 647–671. https://doi.org/10.1080/10584609.2020.1840462

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review, 11*(2), 167–203. https://doi.org/10.1177/1088868307301033

Bliuc, A.-M., Bouguettaya, A., & Felise, K. D. (2021). Online intergroup polarization across political fault lines: An integrative review. *Frontiers in Psychology, 12.* Article 641215. https://doi.org/10.3389/fpsyg.2021.641215

Bode, L., & Vraga, E. K. (2018). See something, say something: correction of global health misinformation on social media. *Health Communication, 33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bode, L., Vraga, E. K., & Tully, M. (2020). Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School (HKS) Misinformation Review, 1*(4). https://doi.org/10.37016/mr-2020-026

Bonetto, E., Varet, F., & Troïan, J. (2019). To resist or not to resist? Investigating the normative features of resistance to persuasion. *Journal of Theoretical Social Psychology, 3*(3), 167–175. https://doi.org/10.1002/jts5.44

Boot, A. B., Dijkstra, K., & Zwaan, R. A. (2021). The processing and evaluation of news content on social media is influenced by peer-user commentary. *Humanities and Social Sciences Communications, 8*(1), Article 209. https://doi.org/10.1057/s41599-021-00889-5

Boyd, D. M. (2008). Why youth (heart) Social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), *Youth, Identity, and Digital Media* (pp. 119–142). The MIT Press. www.doi.org/10.1162/dmal.9780262524834.119

Brader, T., & Marcus, G. E. (2013). Emotion and Political Psychology. In L. Huddy, D. O. Sears & J. S. Levy (Eds.), *The Oxford Handbook of Political Psychology* (pp. 165–204). https://doi.org/10.1093/OXFORDHB/9780199760107.013.0006

Bradley, M. M., & Lang, P. J. (2002). Measuring Emotion: Behavior, Feeling, and Physiology. In R. D. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion (pp. 242–276). Oxford University Press.*

Bradshaw, A. S., Shelton, S. S., Wollney, E., Treise, D., & Auguste, K. (2021). Pro-vaxxers get out: Anti-vaccination advocates influence undecided first-time, pregnant, and new mothers on Facebook. *Health Communication, 36*(6), 693–702. https://doi.org/10.1080/10410236.2020.1712037

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General, 149*(4), 746–756. https://doi.org/10.1037/xge0000673

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., Van Bavel, J. J., & Fiske, S. T. (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

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology, 49*(5), 811–821. https://doi.org/10.1016/j.jesp.2013.04.005

Brennan, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020, April 7). Types, sources, and claims of COVID-19 misinformation. *Reuters Institute for the Study of Journalism*. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Brey, P. (2014). The physical and social reality of virtual worlds. In M. Grimshaw (Ed.), *The Oxford Handbook of Virtuality* (pp. 42–54). Oxford University Press.

Bringula, R. P., Catacutan-Bangit, A. E., Garcia, M. B., Gonzales, J. P. S., & Valderama, A. M. C. (2022). "Who is gullible to political disinformation?": Predicting susceptibility of university students to fake news. *Journal of Information Technology and Politics, 19*(2), 165–179. https://doi.org/10.1080/19331681.2021.1945988

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

Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *PLoS ONE, 15*(10), Article e0239666. https://doi.org/10.1371/journal.pone.0239666

Buchanan, T., & Benson, V. (2019). Spreading disinformation on Facebook: do trust in message source, risk propensity, or personality affect the organic reach of "fake news"? *Social Media + Society, 5*(4), 1–9. https://doi.org/10.1177/2056305119888654

Bucknell Bossen, C., & Kottasz, R. (2020). Uses and gratifications sought by pre-adolescent and adolescent TikTok consumers. *Young Consumers, 21*(4), 463–478. https://doi.org/10.1108/YC-07-2020-1186

Burkhardt, J. M. (2017). History of Fake News. *Library Technology Reports, 53*(8), Article 8. https://journals.ala.org/index.php/ltr/article/view/6497

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F., Wellings, K., Johnson, A. M., & Erens, B. (2016). Using the web to collect data on sensitive behaviours: a study looking at mode effects on the British national survey of sexual attitudes and lifestyles. *PLOS ONE, 11*(2). https://doi.org/10.1371/journal.pone.0147983

Bussey, K. (1999). Children's categorization and evaluation of different types of lies and truths. *Child Development, 70*(6), 1338–1347. https://doi.org/10.1111/1467-8624.00098

Cantarero, K., & Szarota, P. (2017). When is a lie more of a lie? Moral judgment mediates the relationship between perceived benefits of others and lie-labelling. *Polish Psychological Bulletin, 48*(2), 315–325. https://doi.org/10.1515/ppb-2017-0036

Cao, F., Zhang, J., Song, L., Wang, S., Miao, D., & Peng, J. (2017). Framing effect in the trolley problem and footbridge dilemma: Number of saved lives matters. *Psychological Reports, 120*(1), 88–101. https://doi.org/10.1177/0033294116685866

Capraro, V., & Celadin, T. (2022). "I think this news is accurate": Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin*, Article 01461672221117691. https://doi.org/10.1177/01461672221117691

Carvalho, S. W., & Luna, D. (2014). Effects of national identity salience on responses to ads. *Journal of Business Research, 67*(5), 1026–1034. https://doi.org/10.1016/j.jbusres.2013.08.009

Cascio, J., & Plant, E. A. (2015). Prospective moral licensing: Does anticipating doing good later allow you to be bad now? *Journal of Experimental Social Psychology, 56*, 110–116. https://doi.org/10.1016/j.jesp.2014.09.009

Chan, E. Y. (2021). Moral foundations underlying behavioral compliance during the COVID-19 pandemic. *Personality and Individual Differences, 171*, Article 110463. https://doi.org/10.1016/j.paid.2020.110463

Chang, C. (2021). Fake news: audience perceptions and concerted coping strategies. *Digital Journalism, 9*(5), 636–659. https://doi.org/10.1080/21670811.2021.1923403

Chen, H. T. (2018). Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media and Society, 20*(10), 3917–3936. https://doi.org/10.1177/1461444818763384

Chen, X., Sin, S. C. J., Theng, Y. L., & Lee, C. S. (2015). Why students share misinformation on social media: Motivation, gender, and study-level differences. *Journal of Academic Librarianship, 41*(5), 583–592. https://doi.org/10.1016/j.acalib.2015.07.003

Cheng, J. W., Mitomo, H., Kamplean, A., & Seo, Y. (2021). Lesser evil? Public opinion on regulating fake news in Japan, South Korea, and Thailand – A three-country comparison. *Telecommunications Policy, 45*(9), 102185. https://doi.org/10.1016/j.telpol.2021.102185

Cheng, Y., & Chen, Z. F. (2020). The influence of presumed fake news influence: Examining public support for corporate corrective response, media literacy interventions, and governmental regulation. *Mass Communication and Society, 23*(5), 705–729. https://doi.org/10.1080/15205436.2020.1750656

Cheung-Blunden, V., Sonar, K. U., Zhou, E. A., & Tan, C. (2021). Foreign disinformation operation's affective engagement: Valence versus discrete emotions as drivers of tweet popularity. *Analyses of Social Issues and Public Policy, 21*(1), 980–997. https://doi.org/10.1111/asap.12262

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Chua, T. H. H., & Chang, L. (2016). Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior, 55*, 190–197. https://doi.org/10.1016/j.chb.2015.09.011

Chung, M., & Kim, N. (2021). When i learn the news is false: How fact-checking information stems the spread of fake news via third-person perception. *Human Communication Research, 47*(1), 1–24. https://doi.org/10.1093/hcr/hqaa010

Ciaramelli, E., Muccioli, M., Làdavas, E., & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2*(2), 84–92. https://doi.org/10.1093/scan/nsm001

Cislak, A., Cichocka, A., Wojcik, A. D., & Milfont, T. L. (2021). Words not deeds: National narcissism, national identification, and support for greenwashing versus genuine proenvironmental campaigns. *Journal of Environmental Psychology, 74*, Article 101576. https://doi.org/10.1016/j.jenvp.2021.101576

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

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

Cohen, E. L., Atwell Seate, A., Kromka, S. M., Sutherland, A., Thomas, M., Skerda, K., & Nicholson, A. (2020). To correct or not to correct? Social identity threats increase willingness to denounce fake news through presumed media influence and hostile media perceptions. *Communication Research Reports, 37*(5), 263–275. https://doi.org/10.1080/08824096.2020.1841622

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155.

Colley, T., Granelli, F., & Althuis, J. (2020). Disinformation's societal impact: Britain, Covid, and beyond. *Defence Strategic Communications, 8,* 89-140. https://doi.org/10.30966/2018.riga.8.3.

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior, 97*, 202–215. https://doi.org/10.1016/j.chb.2019.03.032

Committee On Intelligence United States Senate. (2019). *Russian active measures campaigns and interference in the 2016 U.S. election: Vol. 2. Russia's use of social media with additional views.* https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

Coppolino Perfumi, S., Bagnoli, F., Caudek, C., & Guazzini, A. (2019). Deindividuation effects on normative and informational social influence within computer-mediated-communication. *Computers in Human Behavior, 92*, 230–237. https://doi.org/10.1016/j.chb.2018.11.017

Corbu, N., Oprea, D. A., Negrea-Busuioc, E., & Radu, L. (2020). 'They can't fool me, but they can fool the others!' Third person effect and fake news detection. European *Journal of Communication, 35*(2), 165–180. https://doi.org/10.1177/0267323120903686

Creyer, E. H., Bettman, J. R., & Payne, J. W. (1990). The Impact of accuracy and effort feedback and goals on adaptive decision behavior. *Journal of Behavioral Decision Making, 3*(1), 1–16. https://doi.org/10.1002/bdm.3960030102

Dabbous, A., Aoun Barakat, K., & de Quero Navarro, B. (2022). Fake news detection and social media trust: A cross-cultural perspective. *Behaviour and Information Technology, 41*(14), 2953–2972. https://doi.org/10.1080/0144929X.2021.1963475

DataReportal. (2023). *Digital 2023: Global Overview Report.* https://datareportal.com/reports/digital-2023-global-overview-report

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly, 47*(1), 1–15. https://doi.org/10.1086/268763

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality & Social Psychology Bulletin, 40*(12), 1559–1573. https://doi.org/10.1177/0146167214551152

De Keersmaecker, J., & Roets, A. (2019). Is there an ideological asymmetry in the moral approval of spreading misinformation by politicians? *Personality and Individual Differences, 143*, 165–169. https://doi.org/10.1016/j.paid.2019.02.003

De Kimpe, L., Walrave, M., Verdegem, P., & Ponnet, K. (2022). What we think we know about cybersecurity: An investigation of the relationship between perceived knowledge, internet trust, and protection motivation in a cybercrime context. *Behaviour and Information Technology, 41*(8), 1796–1808. https://doi.org/10.1080/0144929X.2021.1905066

DeCook, J. R. (2018). Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology, 43*(4), 485–504. https://doi.org/10.1080/17439884.2018.1544149

Derbaix, C., & Decrop, A. (2011). Colours and scarves: An ethnographic account of football fans and their paraphernalia. *Leisure Studies, 30*(3), 271–291. https://doi.org/10.1080/02614367.2010.527356

Di Domenico, G., Nunan, D., Sit, J., & Pitardi, V. (2021). Free but fake speech: When giving primacy to the source decreases misinformation sharing on social media. *Psychology and Marketing, 38*(10), 1700–1711. https://doi.org/10.1002/mar.21479

Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review, 1*(1). https://doi.org/10.37016/mr-2020-001

Digital Culture Media and Sport Committee. (2019). *Disinformation and 'fake news': Final report (HC1791*). The House of Commons. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

DiResta, R., Shaffer, D. K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2019). *The tactics & tropes of the Internet Research Agency*. New Knowledge. https://digitalcommons.unl.edu/senatedocs/2/

Ditrich, L., & Sassenberg, K. (2017). Kicking out the trolls – Antecedents of social exclusion intentions in Facebook groups. *Computers in Human Behavior, 75*, 32–41. https://doi.org/10.1016/j.chb.2017.04.049

Dong, M., van Prooijen, J. W., & van Lange, P. A. M. (2019). Self-enhancement in moral hypocrisy: Moral superiority and moral identity are about better appearances. *PLoS ONE, 14*(7), Article e0219382. https://doi.org/10.1371/journal.pone.0219382

Douglas, K. M., & McGarty, C. (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology, 40*(3), 399–416. https://doi.org/10.1348/014466601164894

Douglas, K. M., & McGarty, C. (2002). Internet identifiability and beyond: A model of the effects of identifiability on communicative behavior. *Group Dynamics: Theory, Research, and Practice, 6*(1), 17–26. https://doi.org/10.1037/1089-2699.6.1.17

Duffy, A., Tandoc, E., & Ling, R. (2020). Too good to be true, too good not to share: The social utility of fake news. *Information Communication and Society, 23*(13), 1965–1979. https://doi.org/10.1080/1369118X.2019.1623904

Duffy, B., & Allington, D. (2020). *The Trusting, the Dissenting and the Frustrated: How the UK is dividing as lockdown is eased*. The Policy Institute.

https://www.kcl.ac.uk/policy-institute/assets/how-the-uk-is-dividing-as-the-lockdown-is-eased.pdf

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015). Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study. *Journal of Medical Internet Research, 17*(6), Article e144. https://doi.org/10.2196/jmir.4343

Dupuv, B. (2019, May 24). Cristiano Ronaldo did not donate money to Palestinians during Ramadan. *AP NEWS*. https://apnews.com/article/fact-checking-5223040944

Duran, G., Dochez, S., Tapiero, I., & Michael, G. A. (2020). Opinions, actions and emotions: Does the content of lies affect their detectability? *Psychology, Crime and Law, 26*(10), 927–949. https://doi.org/10.1080/1068316X.2020.1742341

Ecker, U. K. H., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology, 40*(2), 241–260. https://doi.org/10.1111/pops.12494

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

Effron, D. A. (2018). It Could Have Been True: How Counterfactual Thoughts Reduce Condemnation of Falsehoods and Increase Political Polarization. *Personality and Social Psychology Bulletin, 44*(5), 729–745. https://doi.org/10.1177/0146167217746152

Effron, D. A., & Conway, P. (2015). When virtue leads to villainy: Advances in research on moral self-licensing. *Current Opinion in Psychology, 6*, 32–35. https://doi.org/10.1016/j.copsyc.2015.03.017

Effron, D. A., & Helgason, B. A. (2022). The moral psychology of misinformation: Why we excuse dishonesty in a post-truth world. *Current Opinion in Psychology, 47*, Article 101375. https://doi.org/10.1016/j.copsyc.2022.101375

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science, 31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Einav, S., Levey, A., Patel, P., & Westwood, A. (2020). Epistemic vigilance online: Textual inaccuracy and children's selective trust in webpages. *British Journal of Developmental Psychology, 38*(4), 566–579. https://doi.org/10.1111/bjdp.12335

Elder, T. J., Sutton, R. M., & Douglas, K. M. (2005). Keeping it to ourselves: Effects of audience size and composition on reactions to criticisms of the ingroup. *Group Processes & Intergroup Relations, 8*(3), 231–244. https://doi.org/10.1177/1368430205053940

Ellemers, N. (2017). *Morality and the Regulation of Social Behavior.* Routledge. https://doi.org/10.4324/9781315661322

Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology, 24*(1), 160–193. https://doi.org/10.1080/10463283.2013.841490

Ellemers, N., Pagliaro, S., Barreto, M., & Leach, C. W. (2008). Is it better to be moral than smart? The effects of morality and competence norms on the decision to work at group status improvement. *Journal of Personality and Social Psychology, 95*, 1397–1410. https://doi.org/10.1037/a0012628

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and Social Identity. *Annual Review of Psychology, 53,* 161–186. https://doi.org/10.1146/annurev.psych.53.100901.135228

Ellemers, N., & van den Bos, K. (2012). Morality in groups: On the social-regulatory functions of right and wrong. *Social and Personality Psychology Compass, 6*(12), 878–889. https://doi.org/10.1111/spc3.12001

Endresen, A., Campbell, A., Torresson, B., & Terry, C. (2020). Sorting fact from fiction without source evaluation is a 50-50 guess in the disinformation age. *Psi Chi Journal of Psychological Research, 25*(3), 213–223. https://doi.org/10.24839/2325-7342.jn25.3.213

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School (HKS) Misinformation Review, 2*(3). https://doi.org/10.37016/mr-2020-71

Evanega, S., Lynas, M., Adams, J. & Smolenyak, K. (2020) *Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic'.* Cornell Alliance for Science. https://allianceforscience.org/wp-content/uploads/2020/09/Evanega-et-al-Coronavirus-misinformationFINAL.pdf

Facebook. (n.d.). *What influences the order of posts in my Facebook News Feed?* Retrieved September 12, 2021. https://www.facebook.com/help/520348825116417/?helpref=uf_share

Facebook. (2021, May 26). *Taking action against people who repeatedly share misinformation.* https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation/

Faragó, L., Kende, A., & Krekó, P. (2020). We only believe in news that we doctored ourselves: the connection between partisanship and political fake news. *Social Psychology, 51*(2), 77–90. https://doi.org/10.1027/1864-9335/a000391

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin and Review, 26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Feinberg, M., & Willer, R. (2013). The Moral Roots of Environmental Attitudes. *Psychological Science, 24*(1), 56–62. https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence? *Personality and Social Psychology Bulletin, 41*(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Feldman, S. (2013). Political ideology. In L. Huddy, D. O. Sears, & J. S. Levy (Eds.), *The Oxford handbook of political psychology* (2nd ed., pp. 1–40). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199760107.013.0019

FeldmanHall, O., Son, J.-Y., & Heffner, J. (2018). Norms and the flexibility of moral action. *Personality Neuroscience, 1*, Article e15. https://doi.org/10.1017/pen.2018.13

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538. https://doi.org/10.1037/a0015808

Fisher, L. (2020, April 6). No 10 attacks Russian claims of Boris Johnson ventilator. *The Times.* https://www.thetimes.co.uk/article/coronavirus-no-10-attacks-russian-fake-news-about-boris-johnson-ventilator-q73qm7f6s

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fissel, E. R., Fisher, B. S., & Nedelec, J. L. (2021). Cyberstalking perpetration among young adults: An assessment of the effects of low self-control and moral disengagement. *Crime and Delinquency, 67*(12), 1935–1961. https://doi.org/10.1177/0011128721989079

Foster, M. D. (2015). Tweeting about sexism: The well-being benefits of a social media collective action. *British Journal of Social Psychology, 54*(4), 629–647. https://doi.org/10.1111/bjso.12101

Fox, J., & Holt, L. F. (2018). Fear of isolation and perceived affordances: The spiral of silence on social networking sites regarding police discrimination. *Mass Communication and Society, 21*(5), 533–554. https://doi.org/10.1080/15205436.2018.1442480

François, C., Nimmo, B., & Eib, C. S. (2019). *The IRA CopyPasta Campaign: Russian accounts posing as Americans on Instagram targeted both sides of polarizing issues*. Graphika. https://www.graphika.com/reports/copypasta

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2022). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review, 40*(3), 560–578. https://doi.org/10.1177/0894439320914853

Frimer, J. A., Gaucher, D., & Schaefer, N. K. (2014). Political conservatives' affinity for obedience to authority is loyal, not blind. *Personality and Social Psychology Bulletin, 40*(9), 1205–1214. https://doi.org/10.1177/0146167214538672

Frimer, J. A., Haidt, J., Dehghani, M., & Boghrati, R. (2017). *Moral foundations dictionaries for linguistic analyses, 2.0.* [Unpublished manuscript]. https://osf.io/xakyw

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology, 72,* 1–12. https://doi.org/10.1016/j.jesp.2017.04.003

Fu, G., Evans, A. D., Wang, L., & Lee, K. (2008). Lying in the name of the collective good: A developmental study. *Developmental Science, 11*(4), 495–503. https://doi.org/10.1111/j.1467-7687.2008.00695.x

Full Fact. (2020). *Report on the Facebook Third Party Fact Checking programme.* https://fullfact.org/media/uploads/tpfc-2020.pdf

Galli, M., & Gorn, G. (2011). Unconscious transfer of meaning to brands. *Journal of Consumer Psychology, 21*(3), 215–225. https://doi.org/10.1016/j.jcps.2010.12.004

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances, 7*(23), 1–10. https://doi.org/10.1126/sciadv.abf1234

Gawronski, B., & Beer, J. S. (2017). What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience, 12*(6), 626–632. https://doi.org/10.1080/17470919.2016.1248787

Gill, H., & Rojas, H. (2020). Chatting in a mobile chamber: Effects of instant messenger use on tolerance toward political misinformation among South Koreans. *Asian Journal of Communication, 30*(6), 470–493. https://doi.org/10.1080/01292986.2020.1825757

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems, 38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the Moral Domain. *Journal of Personality and Social Psychology, 101*(2), 366–385. https://doi.org/10.1037/a0021847

Greene, C. M., Nash, R. A., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory, 29*(5), 587–604. https://doi.org/10.1080/09658211.2021.1923754

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Grimmelikhuijsen, S., & Knies, E. (2015). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences, 83*(3), 583–601. https://doi.org/10.1177/0020852315585950

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the*

National Academy of Sciences of the United States of America, 117(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Gunther, A. C., & Mundy, P. (1993). Biased optimism and the third-person effect. *Journalism & Mass Communication Quarterly, 70*(1), 58–67. https://doi.org/10.1177/107769909307000107

Gupta, N., Rigotti, L., & Wilson, A. (2021). *The experimenters' dilemma: Inferential preferences over populations*. arXiv. https://doi.org/10.48550/arXiv.2107.05064

Hagey, K., & Horwitz, J. (2021, September 15). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal.* https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J., & Graham

, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In S. Stich, P. Carruthers, & S. Laurence (Eds.), *The innate mind: Vol. 3. Foundations and the future* (pp. 367–392). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195332834.003.0019

Haidt, J., & Kesebir, S. (2010). Morality. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 797–832). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy002022

Halpern, D., Valenzuela, S., Katz, J., & Miranda, J. P. (2019). From belief in conspiracy theories to trust in others: Which factors influence exposure, believing and sharing fake news. In G. Meiselwitz (Ed.), *HCII 2019: Social computing and social media. design, human behavior and analytics: Vol. 11578. Lecture notes in computer science* (pp. 217–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-21902-4_16

Hameleers, M. (2019). Susceptibility to mis- and disinformation and the effectiveness of fact-checkers: Can misinformation be effectively combated? *Studies in Communication and Media, 8*(4), 523–546. https://doi.org/10.5771/2192-4007-2019-4-523

Hameleers, M. (2020). My reality is more truthful than yours: radical right-wing politicians' and citizens' construction of "fake" and "truthfulness" on social media—Evidence from the United States and the Netherlands. *International Journal of Communication, 14*, 1135–1152. https://ijoc.org/index.php/ijoc/article/view/12463

Hameleers, M. (2020). Populist disinformation: Exploring intersections between online populism and disinformation in the US and the Netherlands. *Politics and Governance, 8*(1), 146–157. https://doi.org/10.17645/pag.v8i1.2478

Hameleers, M., Humprecht, E., Möller, J., & Lühring, J. (2021). Degrees of deception: The effects of different types of COVID-19 misinformation and the effectiveness of corrective information in crisis times. *Information Communication and Society.* https://doi.org/10.1080/1369118X.2021.2021270

Harmon-Jones, E. (2000). Cognitive dissonance and experienced negative affect: Evidence that dissonance increases experienced negative affect even in the absence of aversive consequences. *Personality and Social Psychology Bulletin, 26*(12), 1490–1501. https://doi.org/10.1177/01461672002612004

Harris, E. A., & Van Bavel, J. J. (2021). Preregistered replication of "feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority". *Psychological Science, 32(*3), 451-458. https://doi.org/10.1177/0956797620968792

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*(4), 643–659. https://doi.org/10.1037/a0023589

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016). One click, many meanings: interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media, 60*(1), 171–187. https://doi.org/10.1080/08838151.2015.1127248

Heering, M. S., Travaglino, G. A., Abrams, D., & Goldsack, E. (2020). "If they don't listen to us, they deserve it": The effect of external efficacy and anger on the perceived legitimacy of hacking. *Group Processes and Intergroup Relations, 23*(6), 863–881. https://doi.org/10.1177/1368430220937777

Helgason, B. A., & Effron, D. A. (2022). It might become true: How prefactual thinking licenses dishonesty. *Journal of Personality and Social Psychology*, *123*(5), 909–940. https://doi.org/10.1037/pspa0000308

Helmus, T. C., Marrone, J. V., Posard, M. N., & Schlang, D. (2020). Russian propaganda hits its mark: Experimentally testing the impact of Russian propaganda and counter-interventions. *RAND Corporation*. https://doi.org/10.7249/rra704-3

Hichy, Z., Mari, S., & Capozza, D. (2008). Pronorm and antinorm deviants: A test of the subjective group dynamics model. *The Journal of Social Psychology*, *148*(5), 641–644. https://doi.org/10.3200/SOCP.148.5.641-644

Hinton, P. R., Perry R., McMurray, I., & Brownlow, C. (2014). *SPSS explained* (2nd ed.). Routledge.

Ho, S. S., Goh, T. J., & Leung, Y. W. (2020). Let's nab fake science news: Predicting scientists' support for interventions using the influence of presumed media influence model. *Journalism*, *23*(4), 910–928. https://doi.org/10.1177/1464884920937488

Hogg, M. A., & Abrams, D. (1998). *Social identifications: A social psychology of intergroup relations and group processes.* Routledge.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, *53*(1), 232–246. https://doi.org/10.3758/s13428-020-01433-0

Hopp, T. (2022). Fake news self-efficacy, fake news identification, and content sharing on Facebook. *Journal of Information Technology and Politics*, *19*(2), 229–252. https://doi.org/10.1080/19331681.2021.1962778

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). *The IRA, social media and political polarization in the United States, 2012-2018.* Computational propaganda research project. https://digitalcommons.unl.edu/senatedocs/1/

Huang, C.-C., Lin, T.-C., & Lin, K.-J. (2009). Factors affecting pass-along email intentions (PAEIs): Integrating the social capital and social cognition theories. *Electronic Commerce Research and Applications*, *8*(3), 160–169. https://doi.org/10.1016/j.elerap.2008.11.001

Huang, J., Su, S., Zhou, L., & Liu, X. (2013). Attitude toward the viral ad: Expanding traditional advertising models to interactive advertising. *Journal of Interactive Marketing*, *27*(1), 36–46. https://doi.org/10.1016/j.intmar.2012.06.001

Huber, F. (2009). Belief and degrees of belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 1–33). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9198-8_1

Human Rights Watch. (2020). *Mexico: Mexicans need accurate COVID-19 information.* https://www.hrw.org/news/2020/03/26/mexico-mexicans-need-accurate-covid-19-information

Huntington, H. E. (2020). Partisan cues and internet memes: Early evidence for motivated skepticism in audience message processing of spreadable political media. *Atlantic Journal of Communication, 28*(3), 194–208. https://doi.org/10.1080/15456870.2019.1614589

Hurst, K., & Stern, M. J. (2020). Messaging for environmental action: The role of moral framing and message source. *Journal of Environmental Psychology, 68,* Article 101394. https://doi.org/10.1016/j.jenvp.2020.101394

Innes, M. (2020). Soft Facts and Digital Behavioural Influencing After the 2017 Terror Attacks Full Report. *Centre for Research and Evidence on Security Threats.* https://crestresearch.ac.uk/resources/soft-facts-full-report/

Ipsos MORI. (2021). *Political Monitor—December 2021*. https://www.ipsos.com/en-uk/least-2-3-britons-think-government-doing-bad-job-managing-immigration-nhs-and-levelling

Iyer, A., Jetten, J., & Haslam, S. A. (2012). Sugaring o'er the devil: Moral superiority and group identification help individuals downplay the implications of ingroup rule-breaking. *European Journal of Social Psychology, 42*(2), 141–149. https://doi.org/10.1002/ejsp.864

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE, 7*(8), Article e42366. https://doi.org/10.1371/journal.pone.0042366

Jahng, M. R., Stoycheff, E., & Rochadiat, A. (2021). They said it's "fake": Effects of discounting cues in online comments on information quality judgments and information authentication. *Mass Communication and Society, 24*(4), 527–552. https://doi.org/10.1080/15205436.2020.1870143

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior, 80,* 295–302. https://doi.org/10.1016/j.chb.2017.11.034

Jennings, F. J., Bramlett, J. C., McKinney, M. S., & Hardy, M. M. (2020). Tweeting along partisan lines: Identity-motivated elaboration and presidential debates. *Social Media + Society*, *6*(4). https://doi.org/10.1177/2056305120965518

Jennings, W., Valgarðsson, V. O., Stoker, G., Devine, D., Gaskell, J., & Evans, M. (2020). *Political trust and the COVID-19 crisis: Pushing populism to the backburner?* TrustGov Project. https://trustgov.net/s/Published-report-covid_and_trust.pdf

John, B. (2020). *The first six months of the pandemic, as told by the fact checks*. First Draft. https://firstdraftnews.org/latest/the-first-six-months-of-the-pandemic-as-told-by-the-fact-checks/

Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, *45*(1), 20–31. https://doi.org/10.1037/0022-3514.45.1.20

Johnson, S. C. (2013, April 23). Analysis: False White House tweet exposes instant trading dangers. *Reuters*. https://www.reuters.com/article/us-usa-markets-tweet-idUSBRE93M1FD20130423

Jones, C. M., Diethei, D., Schöning, J., Shrestha, R., Jahnel, T., & Schüz, B. (2021). *Social reference cues can reduce misinformation sharing behaviour on social media.* PsyArXiv. https://doi.org/10.31234/osf.io/v6fc9

Jones, M. K., Calzavara, L., Allman, D., Worthington, C. A., Tyndall, M., & Iveniuk, J. (2016). A comparison of web and telephone responses from a national HIV and AIDS survey. *JMIR Public Health and Surveillance, 2*(2), Article e37. https://doi.org/10.2196/publichealth.5184

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist, 65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin, 37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making, 8*(4), 407–424. https://doi.org/10.2139/ssrn.2182588

Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology, 36*(S1), 1–43. https://doi.org/10.1111/pops.12244

Kahan, D. M., Jamieson, K. H., Landrum, A., & Winneg, K. (2017). Culturally antagonistic memes and the Zika virus: An experimental test. *Journal of Risk Research, 20*(1), 1–40. https://doi.org/10.1080/13669877.2016.1260631

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy, 1*(1), 54–86. https://doi.org/10.1017/bpp.2016.2

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732–735. https://doi.org/10.1038/nclimate1547

Kahneman, D., & Frederick, S. (2012). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/cbo9780511808098.004

Kahneman, D., Schkade, D., & Sunstein, C. R. (1998). Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty, 16*(1), 49–86. https://doi.org/10.1023/A:1007710408413

Kalimeris, D., Bhagat, S., Kalyanaraman, S., & Weinsberg, U. (2021). Preference amplification in recommender systems. *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 805–815. https://doi.org/10.1145/3447548.3467298

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kaye, L. K., Rousaki, A., Joyner, L. C., Barrett, L. A. F., & Orchard, L. J. (2022). The Online Behaviour Taxonomy: A conceptual framework to understand behaviour in computer-mediated communication. *Computers in Human Behavior, 137*, Article 107443. https://doi.org/10.1016/j.chb.2022.107443

Kelly, D. (2019). Evaluating the news: (Mis)perceptions of objectivity and credibility. *Political Behavior, 41*(2), 445–471. https://doi.org/10.1007/s11109-018-9458-4

Kemp, S. (2020). *Digital 2020: The United Kingdom*. DataReportal. https://datareportal.com/reports/digital-2020-united-kingdom

Kenny, D. A. (2017). *MedPower: An interactive tool for the estimation of power in tests of mediation* [Computer software]. https://davidakenny.shinyapps.io/MedPower/

Kenny, D. A. (2018, September 15). Moderator variables: An introduction. *David A. Kenny.* https://davidakenny.net/cm/moderation.htm

Khan, M. L., & Idris, I. K. (2019). Recognise misinformation and verify before sharing: A reasoned action and information literacy perspective. *Behaviour and Information Technology, 38*(12), 1194–1212. https://doi.org/10.1080/0144929X.2019.1578828

Kilpatrick, C. (2014, September 11). Fraud accused moved £38,000 of Linfield supporters' club money into his own bank account, court is told. *Belfast Telegraph.* https://www.belfasttelegraph.co.uk/news/northern-ireland/fraud-accused-moved-38000-of-linfield-supporters-club-money-into-his-own-bank-account-court-is-told/30578021.html

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

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems, 36*(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

Kim, C., & Yang, S.-U. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review, 43*(2), 441–449. https://doi.org/10.1016/j.pubrev.2017.02.006

Kim, T., Lee, H., Kim, M. Y., Kim, S., & Duhachek, A. (2022). AI increases unethical consumer behavior due to reduced anticipatory guilt. Journal of the Academy of Marketing Science, 51(4), 785-801. https://doi.org/10.1007/s11747-021-00832-9

Kivikangas, J. M., Fernández-Castilla, B., Järvelä, S., Ravaja, N., & Lönnqvist, J. E. (2021). Moral foundations and political orientation: Systematic review and meta-analysis. *Psychological Bulletin, 147*(1), 55–94. https://doi.org/10.1037/bul0000308

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911. https://doi.org/10.1038/nature05631

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality, 46*(2), 184–194. https://doi.org/10.1016/j.jrp.2012.01.006

Krämer, B., & Conrad, J. (2017). Social ontologies online: The representation of social structures on the internet. *Social Media + Society, 3*(1). https://doi.org/10.1177/2056305117693648

Krishna, R. (2020, November 14). The infection fatality rate for Covid-19 is higher than 0.1%. *Full Fact.* https://fullfact.org/health/toby-young-ifr-tweet/

Kross, E., Verduyn, P., Sheppes, G., Costello, C. K., Jonides, J., & Ybarra, O. (2021). Social media and well-being: pitfalls, progress, and next steps. *Trends in Cognitive Sciences*, *25*(1), 55–66. https://doi.org/10.1016/j.tics.2020.10.005

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Laato, S., Islam, A. K. M. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, *29*(3), 288–305. https://doi.org/10.1080/0960085X.2020.1770632

Lada, A., Wang, M., & Yan, T. (2021). H*ow does news feed predict what you want to see?* Facebook. https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

Lanius, C., Weber, R., & MacKenzie, W. I. (2021). Use of bot and content flags to limit the spread of misinformation among social networks: A behavior and attitude survey. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00739-x

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. Competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Lee, E.-J. (2007). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, *57*(2), 385–403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

Lee, S., Forrest, J. P., Strait, J., Seo, H., Lee, D., & Xiong, A. (2020). Beyond cognitive ability: Susceptibility to fake news is also explained by associative inference. *Extended abstracts of the conference on human factors in computing systems*, 1–8. https://doi.org/10.1145/3334480.3383077

Lee, S.-Y., Hansen, S. S., & Lee, J. K. (2016). What makes us click "like" on Facebook? Examining psychological, technological, and motivational factors on virtual endorsement. *Computer Communications*, *73*(B), 332–341. https://doi.org/10.1016/j.comcom.2015.08.002

Leeper, T. J., & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Political Psychology, 35*(S1), 129–156. https://doi.org/10.1111/pops.12164

Leidner, B., & Castano, E. (2012). Morality shifting in the context of intergroup violence. *European Journal of Social Psychology, 42*(1), 82–91. https://doi.org/10.1002/ejsp.846

Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin, 36*(8), 1115–1129. https://doi.org/10.1177/0146167210376391

Leiser, A. (2022). Psychological perspectives on participatory culture: Core motives for the use of political internet memes. *Journal of Social and Political Psychology, 10*(1), 236-252. https://doi.org/10.5964/jspp.6377

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology, 53*, 107–117. https://doi.org/10.1016/j.jesp.2014.03.005

Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science, 16*(3), 190–195. https://doi.org/10.1111/j.0956-7976.2005.00802.x

LinkedIn. (n.d.). *Visibility and Impact of Your Social Activity on the LinkedIn Feed*. Retrieved September 12, 2021. https://www.linkedin.com/help/linkedin/answer/a523397

Liu, P. L., & Huang, L. V. (2020). Digital disinformation about covid-19 and the third-person effect: Examining the channel differences and negative emotional outcomes. *Cyberpsychology, Behavior, and Social Networking, 23*(11), 789–793. https://doi.org/10.1089/cyber.2020.0363

Liu, Y., Rui, J. R., & Cui, X. (2017). Are people willing to share their political opinions on Facebook? Exploring roles of self-presentational concern in spiral of silence. *Computers in Human Behavior, 76*, 294–302. https://doi.org/10.1016/j.chb.2017.07.029

London Elects. (2021). *Results 2021. London elects - Mayor of London & London Assembly elections*. https://www.londonelects.org.uk/im-voter/election-results/results-2021

Lowe-Calverley, E., & Grieve, R. (2018). Thumbs up: A thematic analysis of image-based posting and liking behaviour on social media. *Telematics and Informatics, 35*(7), 1900–1913. https://doi.org/10.1016/j.tele.2018.06.003

Lunz Trujillo, K., Motta, M., Callaghan, T., & Sylvester, S. (2021). Correcting misperceptions about the MMR vaccine: Using psychological risk factors to inform targeted communication strategies. *Political Research Quarterly, 74*(2), 464–478. https://doi.org/10.1177/1065912920907695

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research, 49*(2), 171–195. https://doi.org/10.1177/0093650220921321

Lyons, B. A., Merola, V., & Reifler, J. (2020). How Bad is the Fake News Problem? In B. A. Lyons, V. Merola, & J. Reifler (Eds.), *The Psychology of Fake News* (pp. 11–26). Routledge. https://doi.org/10.4324/9780429295379-3

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences of the United States of America, 118*(23), Article e2019527118. https://doi.org/10.1073/pnas.2019527118

Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanku, E., Tully, M., Umejei, E., & Uzuegbunam, C. (2021). Motivations for sharing misinformation: A comparative study in six sub-Saharan African countries. *International Journal of Communication, 15,* 1200–1219. https://ijoc.org/index.php/ijoc/article/view/14801

Mahoney, C. (2020). Is this what a feminist looks like? Curating the feminist self in the neoliberal visual economy of Instagram. *Feminist Media Studies*, *22*(3), 1–17. https://doi.org/10.1080/14680777.2020.1810732

Major Cities Chiefs Association. (2022). *Violent crime survey—National totals. Midyear comparison: January 1 to June 30, 2022, and 2021*. https://majorcitieschiefs.com/wp-content/uploads/2022/08/MCCA-Violent-Crime-Report-2022-and-2021-Midyear.pdf

Mann, H., Garcia-Rada, X., Hornuf, L., Tafurt, J., & Ariely, D. (2016). Cut from the same cloth: Similarly dishonest individuals across countries. *Journal of Cross-Cultural Psychology, 47*(6), 858–874. https://doi.org/10.1177/0022022116648211

Mann, H., Garcia-Rada, X., Houser, D., & Ariely, D. (2014). Everybody else is doing it: Exploring social transmission of lying behavior. *PLOS ONE, 9*(10), Article e109591. https://doi.org/10.1371/journal.pone.0109591

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

Mashuri, A., Putra, I. E., Kavanagh, C., Zaduqisti, E., Sukmawati, F., Sakdiah, H., & Selviana, S. (2022). The socio-psychological predictors of support for post-truth collective action. *Journal of Social Psychology, 162*(4), 504–522. https://doi.org/10.1080/00224545.2021.1935678

Masullo, G. M., Lu, S., & Fadnis, D. (2021). Does online incivility cancel out the spiral of silence? A moderated mediation model of willingness to speak out. *New Media and Society, 23*(11), 3391–3414. https://doi.org/10.1177/1461444820954194

Maxey, S. (2021). Limited spin: When the public punishes leaders who lie about military action. *Journal of Conflict Resolution, 65*(2–3), 283–312. https://doi.org/10.1177/0022002720961517

McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass, 11*(12), Article e12362. https://doi.org/10.1111/spc3.12362

Melro, A., & Pereira, S. (2019). Fake or not fake? Perceptions of undergraduates on (DIS)Information and critical thinking. *Medijske Studije, 10*(19), 46–67. https://doi.org/10.20901/ms.10.19.3

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy and Internet, 12*(2), 165–183. https://doi.org/10.1002/poi3.214

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on Instagram: The impact of trusted endorsements on message credibility. *Social Media + Society, 6*(2). https://doi.org/10.1177/2056305120935102

Merrill, J. B., & Oremus, W. (2021, October 26). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *The Washington Post*. https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Meter, D. J., & Bauman, S. (2018). Moral disengagement about cyberbullying and parental monitoring: Effects on traditional bullying and victimization via cyberbullying involvement. *Journal of Early Adolescence, 38*(3), 303–326. https://doi.org/10.1177/0272431616670752

Metropolitan Police. (2020). *MPS FY 2019/20 crime statistics*. https://www.met.police.uk/sd/stats-and-data/met/year-end-crime-statistics-19-20/

Michael, R. B., & Breaux, B. O. (2021). The relationship between political affiliation and beliefs about sources of "fake news". *Cognitive Research: Principles and Implications, 6*(1), 6. https://doi.org/10.1186/s41235-021-00278-1

Molden, D. C., Lee, A. Y., & Higgins, E. T. (2008). Motivations for promotion and prevention. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (1st ed., pp. 169–187). The Guilford Press.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33–43. https://doi.org/10.1037/0022-3514.81.1.33

Moore, A., Hong, S., & Cram, L. (2021). Trust in information, political identity and the brain: An interdisciplinary fMRI study. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*(1822), Article 20200140. https://doi.org/10.1098/rstb.2020.0140

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

Morosoli, S., Van Aelst, P., Humprecht, E., Staender, A., & Esser, F. (2022). Identifying the drivers behind the dissemination of online misinformation: A study on political attitudes and individual characteristics in the context of engaging with misinformation on social media. *American Behavioral Scientist.* Advanced online publication. https://doi.org/10.1177/00027642221118300

Moskalenko, S., McCauley, C., & Rozin, P. (2006). Group identification under conditions of threat: College students' attachment to country, family, ethnicity, religion, and university before and after September 11, 2001. *Political Psychology, 27*(1), 77–97. https://doi.org/10.1111/j.1467-9221.2006.00450.x

Mosseri, A. (2021, June 8). *Shedding more light on how Instagram works*. Instagram. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works

Motta, M., Callaghan, T., Sylvester, S., & Lunz-Trujillo, K. (2023). Identifying the prevalence, correlates, and policy consequences of anti-vaccine social identity. *Politics, Groups, and Identities, 11*(1), 108–122. https://doi.org/10.1080/21565503.2021.1932528

Mun, I. B., & Kim, H. (2021). Influence of false self-presentation on mental health and deleting behavior on Instagram: The mediating role of perceived popularity.

Frontiers in Psychology, 12, Article 660484. https://doi.org/10.3389/fpsyg.2021.660484

Murphy, G., Murray, E., & Gough, D. (2021). Attitudes towards feminism predict susceptibility to feminism-related fake news. *Applied Cognitive Psychology, 35*(5), 1182–1192. https://doi.org/10.1002/acp.3851

Myrick, J. G., & Erlichman, S. (2020). How audience involvement and social norms foster vulnerability to celebrity-based dietary misinformation. *Psychology of Popular Media*, *9*(3), 367–379. https://doi.org/10.1037/ppm0000229

Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications, 5*(1), 56. https://doi.org/10.1186/s41235-020-00251-4

Nauroth, P., Gollwitzer, M., Bender, J., & Rothmund, T. (2015). Social identity threat motivates science- discrediting online comments. *PLoS ONE, 10*(2), Article e0117476. https://doi.org/10.1371/journal.pone.0117476

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

Nee, R. C., & De Maio, M. (2019). A 'presidential look'? An analysis of gender framing in 2016 persuasive memes of Hillary Clinton. *Journal of Broadcasting & Electronic Media*, *63*(2), 304–321. https://doi.org/10.1080/08838151.2019.1620561

Neill Hoch, I. (2020). Russian Internet Research Agency disinformation activities on Tumblr: Identity, privacy, and ambivalence. *Social Media + Society, 6*(4), Article 2056305120961783. https://doi.org/10.1177/2056305120961783

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media + Society, 6*(1), Article 2056305119897322. https://doi.org/10.1177/2056305119897322

Neyazi, T. A., & Muhtadi, B. (2021). Selective belief: How partisanship drives belief in misinformation. *International Journal of Communication, 15*, 1286–1308. https://ijoc.org/index.php/ijoc/article/view/15477

Nilsson, A., Erlandsson, A., & Västfjäll, D. (2019). The Complex Relation Between Receptivity to Pseudo-Profound Bullshit and Political Ideology. *Personality and*

Social Psychology Bulletin, 45(10), 1440–1454. https://doi.org/10.1177/0146167219830415

Nimmo, B., François, C., Eib, C. S., & Ronzaud, L. (2020). *The Case of the Inauthentic Reposting Activists.* Graphika. https://graphika.com/reports/the-case-of-the-inauthentic-reposting-activists

Nimmo, B., François, C., Eib, C. S., Ronzaud, L., Smith, M., Lederer, T., Carter, A., & Mcaweeney, E. (2020). *IRA in Ghana: Double Deceit*. Graphika. https://graphika.com/reports/ira-in-ghana-double-deceit

Nimmo, B., Francois, C., Eib, C. S., & Tamora, L. (2020). *Operation Red Card*. Graphika. https://www.graphika.com/reports/operation-red-card

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

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media and Society, 18*(8), 1491–1507. https://doi.org/10.1177/1461444814563519

O'Brien, T. C., Palmer, R., & Albarracin, D. (2021). Misplaced trust: When trust in science fosters belief in pseudoscience and the benefits of critical evaluation. *Journal of Experimental Social Psychology, 96,* Article 104184. https://doi.org/10.1016/j.jesp.2021.104184

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The ineffectiveness of fact-checking labels on news memes and articles. *Mass Communication and Society, 23*(5), 682–704. https://doi.org/10.1080/15205436.2020.1733613

Ofcom. (2022). *News Consumption in the UK: 2022*. Ofcom. https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption

Office for National Statistics. (2021). *The nature of violent crime in England and Wales: Year ending March 2020.* https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/thenatureofviolentcrimeinenglandandwales/yearendingmarch2020

Ohme, J., Hameleers, M., Brosius, A., & Van der Meer, T. (2021). Attenuating the crisis: The relationship between media use, prosocial political participation, and holding misinformation beliefs during the COVID-19 pandemic. *Journal of Elections, Public*

Opinion and Parties, 31(S1), 285–298. https://doi.org/10.1080/17457289.2021.1924735

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review, 115*(3), 999–1015. https://doi.org/10.1017/S0003055421000290

Packer, D. J. (2014). On not airing our dirty laundry: Intergroup contexts suppress ingroup criticism among strongly identified group members. *British Journal of Social Psychology, 53*(1), 93–111. https://doi.org/10.1111/bjso.12017

Pagliaro, S., Ellemers, N., & Barreto, M. (2011). Sharing moral values: Anticipated ingroup respect as a determinant of adherence to morality-based (but not competence-based) group norms. *Personality and Social Psychology Bulletin, 37*(8), 1117–1129. https://doi.org/10.1177/0146167211406906

Pagliaro, S., Ellemers, N., Barreto, M., & Di Cesare, C. (2016). Once dishonest, always dishonest? The impact of perceived pervasiveness of moral evaluations of the self on motivation to restore a moral reputation. *Frontiers in Psychology, 7*, Article 586. https://doi.org/10.3389/fpsyg.2016.00586

Paisana, M., Pinto-Martinho, A., & Cardoso, G. (2020). Trust and fake news: Exploratory analysis of the impact of news literacy on the relationship with news content in Portugal. *Communication and Society, 33*(2), 105–117. https://doi.org/10.15581/003.33.2.105-117

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Panjwani, A. (2020a, September 9). Matt Hancock gets test and trace figures wrong again. *Full Fact.* https://fullfact.org/health/september-2020-test-trace/

Panjwani, A. (2020b, November 24). Research suggesting increased Covid-19 risk for dog-owners may be barking up the wrong tree. *Full Fact.* https://fullfact.org/health/dog-risk-coronavirus/

Parkinson, M., & Byrne, R. M. J. (2018). Judgments of moral responsibility and wrongness for intentional and accidental harm and purity violations. *Quarterly Journal of Experimental Psychology, 71*(3), 779–789. https://doi.org/10.1080/17470218.2016.1276942

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and Reasoning in Moral Judgment. *Cognitive Science, 36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications, 6* (1), Article 24. https://doi.org/10.1186/s41235-021-00292-3

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*, Article 2104. https://doi.org/10.3389/fpsyg.2018.02104

Penney, J. (2020). Its my duty to be like 'This is Wrong'": Youth political social media practices in the Trump Era. *Journal of Computer-Mediated Communication, 24*(6), 319–334. https://doi.org/10.1093/jcmc/zmz017

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science, 66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology, 7*(1). https://doi.org/10.1525/collabra.25293

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865–1880. https://doi.org/10.1037/xge0000465.supp

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*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*(1), Article 1. https://doi.org/10.1038/s41467-022-30073-5

Pereira, A., Harris, E., & Van Bavel, J. J. (2023). Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations, 26*(1), 24–47. https://doi.org/10.1177/13684302211030004

Peterson, C. C., Peterson, J. L., & Seeto, D. (1983). Developmental changes in ideas about lying. *Child Development, 54*(6), 1529–1535. https://doi.org/10.1111/j.1467-8624.1983.tb00069.x

Piejka, A., & Okruszek, Ł. (2020). Do you believe what you have been told? Morality and scientific literacy as predictors of pseudoscience susceptibility. *Applied Cognitive Psychology, 34*(5), 1072–1082. https://doi.org/10.1002/acp.3687

Pierce, L., Rogers, T., & Snyder, J. A. (2016). Losing hurts: The happiness impact of partisan electoral loss. *Journal of Experimental Political Science, 3*(1), 44–59. https://doi.org/10.1017/XPS.2015.8

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology, 52*(4), 597–617. https://doi.org/10.1111/bjso.12006

Poytner. (2021). *Fighting the Infodemic: The #CoronaVirusFacts Alliance.* https://www.poynter.org/coronavirusfactsalliance/

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Bavel, J. J. V. (2022). *The misleading count: An identity-based intervention to mitigate the spread of partisan misinformation*. PsyArXiv. https://doi.org/10.31234/osf.io/7j26y

Prike, T., Reason, R., Ecker, U. K. H., Swire-Thompson, B., & Lewandowsky, S. (2023). Would I lie to you? Party affiliation is more important than Brexit in processing political misinformation. *Royal Society Open Science, 10*(2), 220508. https://doi.org/10.1098/rsos.220508

Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health, 17*(7), 2430. https://doi.org/10.3390/ijerph17072430

Rabinowitz, M., Latella, L., Stern, C., & Jost, J. T. (2016). Beliefs about childhood vaccination in the United States: Political ideology, false consensus, and the illusion of uniqueness. *PLOS ONE, 11*(7), Article e0158382. https://doi.org/10.1371/journal.pone.0158382

Rana, M., & O'Neill, S. (2020, October 16). Russians spread fake news over Oxford coronavirus vaccine. *The Times*. https://www.thetimes.co.uk/article/russians-spread-fake-news-over-oxford-coronavirus-vaccine-2nzpk8vrq

Rathje, S., Roozenbeek, J., Traberg, C. S., Bavel, J. J. V., & Linden, D. S. van der. (2022). Meta-analysis reveals that accuracy nudges have little to no effect for U.S. conservatives: Regarding Pennycook et al. (2020) [Letter to the editor]. *Psychological Science.* https://doi.org/10.25384/SAGE.12594110.v2

Rathje, S., van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences of the United States of America, 118*(26), Article e2024292118. https://doi.org/10.1073/pnas.2024292118

Rhodes, S. C. (2022). Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication, 39*(1), 1–22. https://doi.org/10.1080/10584609.2021.1910887

Rojas Torrijos, J. L., & Mello, M. S. (2021). Football misinformation matrix: a comparative study of 2020 winter transfer news in four European sports media outlets. *Journalism and Media*, *2*(4). https://doi.org/10.3390/journalmedia2040037

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science, 32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online

misinformation: Solomon revisited. *Educational and Psychological Measurement, 81*(2), 340–362. https://doi.org/10.1177/0013164420940378

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world: Susceptibility to COVID misinformation. *Royal Society Open Science, 7*(10). https://doi.org/10.1098/rsos.201199

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Rosca, V. (2011). Corporate social responsibility in English football: History and present. *Management & Marketing, 6*(2), 327–346. http://www.managementmarketing.ro/pdf/articole/229.pdf

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision Making, 16*(2), 484–504. https://doi.org/10.31234/osf.io/cgsx6

Rossini, P., Stromer-Galley, J., Baptista, E. A., & Veiga de Oliveira, V. (2021). Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media and Society, 23*(8), 2430–2451. https://doi.org/10.1177/1461444820928059

Ryan, T. J., & Aziz, A. R. (2021). Is the political right more credulous? Experimental evidence against asymmetric motivations to believe false political information. *Journal of Politics, 83*(3), 1168–1172. https://doi.org/10.1086/711133

Rykov, Y. G., Meylakhs, P. A., & Sinyavskaya, Y. E. (2017). Network structure of an AIDS-denialist online community: Identifying core members and the risk group. *American Behavioral Scientist, 61*(7), 688–706. https://doi.org/10.1177/0002764217717565

Saling, L. L., Mallal, D., Scholer, F., Skelton, R., & Spina, D. (2021). No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. *PLOS ONE, 16*(8), Article e0255702. https://doi.org/10.1371/journal.pone.0255702

Salvatore, J., & Morton, T. A. (2021). Evaluations of science are robustly biased by identity concerns. *Group Processes and Intergroup Relations, 24*(4), 568–582. https://doi.org/10.1177/1368430221996818

Salvi, C., Iannello, P., Cancer, A., McClay, M., Rago, S., Dunsmoor, J. E., & Antonietti, A. (2021). Going viral: how fear, socio-cognitive polarization and problem-solving

influence fake news detection and proliferation during COVID-19 pandemic. *Frontiers in Communication, 5*, Article 562588. https://doi.org/10.3389/fcomm.2020.562588

Sanchez, C., & Dunning, D. (2021). Cognitive and emotional correlates of belief in political misinformation: Who endorses partisan misbeliefs? *Emotion, 21*(5), 1091–1102. https://doi.org/10.1037/emo0000948

Schaeffer, K. (2019). *Share of Americans who favor stricter gun laws has increased since 2017*. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/10/16/share-of-americans-who-favor-stricter-gun-laws-has-increased-since-2017/

Schaewitz, L., Kluck, J. P., Klösters, L., & Krämer, N. C. (2020). When is disinformation (in)credible? Experimental findings on message characteristics and individual differences. *Mass Communication and Society, 23*(4), 484–509. https://doi.org/10.1080/15205436.2020.1716983

Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? *Public Opinion Quarterly, 82*(1), 135–147. https://doi.org/10.1093/poq/nfx042

Schaffner, B. F., & Roche, C. (2017). Misinformation and motivated reasoning: Responses to economic news in a politicized environment. *Public Opinion Quarterly, 81*(1), 86–110. https://doi.org/10.1093/poq/nfw043

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin, 41*(8), 1147–1163. https://doi.org/10.1177/0146167215591501

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Scherer, L. D., McPhetres, J., Pennycook, G., Kempe, A., Allen, L. A., Knoepke, C. E., Tate, C. E., & Matlock, D. D. (2021). Who is susceptible to online health misinformation? A test of four psychosocial hypotheses. *Health Psychology, 40*(4), 274–284. https://doi.org/10.1037/hea0000978

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition, 124*(3), 325–333. https://doi.org/10.1016/j.cognition.2012.06.004

Shahab, L., McGowan, J. A., Waller, J., & Smith, S. G. (2018). Prevalence of beliefs about actual and mythical causes of cancer and their association with socio-demographic and health-related characteristics: Findings from a cross-sectional survey in England.

European Journal of Cancer, 103, 308–316. https://doi.org/10.1016/j.ejca.2018.03.029

Shaw, A., DeScioli, P., & Olson, K. R. (2012). Fairness versus favoritism in children. *Evolution and Human Behavior, 33*(6), 736–745. https://doi.org/10.1016/j.evolhumbehav.2012.06.001

Shi, J., Ghasiya, P., & Sasahara, K. (2021). Psycho-linguistic differences among competing vaccination communities on social media. *APSIPA Transactions on Signal and Information Processing, 11*(2), Article e15. https://doi.org/10.1561/116.00000056

Silverman, C. (2017, October 11). Facebook says its fact checking program helps reduce the spread of a fake story by 80%. *Buzzfeed*. https://www.buzzfeednews.com/article/craigsilverman/facebook-just-shared-the-first-data-about-how-effective-its

Smith, C. A. (2019). Weaponized iconoclasm in Internet memes featuring the expression 'Fake News'. *Discourse and Communication, 13*(3), 303–319. https://doi.org/10.1177/1750481319835639

Smith, C. N., & Seitz, H. H. (2019). Correcting misinformation about neuroscience via social media. *Science Communication, 41*(6), 790–819. https://doi.org/10.1177/1075547019890073

Smith, K. B., Alford, J. R., Hibbing, J. R., Martin, N. G., & Hatemi, P. K. (2017). Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science, 61*(2), 424–437. https://doi.org/10.1111/ajps.12255

Smith, L. E., Amlôt, R., Lambert, H., Oliver, I., Robin, C., Yardley, L., & Rubin, G. J. (2020). Factors associated with adherence to self-isolation and lockdown measures in the UK: a cross-sectional survey. *Public Health, 187*, 41–52. https://doi.org/10.1016/j.puhe.2020.07.024

Smith, N. (2020). *Securing a Brighter Future: The role of youth services in tackling knife crime.* All-Party Parliamentary Group for Knife Crime & Violence Reduction. http://www.preventknifecrime.co.uk/wp-content/uploads/2020/03/Securing-a-brighter-future-the-role-of-youth-services-in-tackling-knife-crime-v.2.pdf

Smith, N., & Leiserowitz, A. (2014). The role of emotion in global warming policy support and opposition. *Risk Analysis, 34*(5), 937–948. https://doi.org/10.1111/risa.12140

Spears, R., Doosje, B., & Ellemers, N. (1997). Self-stereotyping in the face of threats to group status and distinctiveness: The role of group identification. *Personality and*

Social Psychology Bulletin, 23(5), 538–553. https://doi.org/10.1177/0146167297235009

Spears, R., Jetten, J., & Scheepers, D. (2002). Distinctiveness and the definition of collective self: A tripartite model. In A. Tesser, D. A. Stapel, & J. V. Wood (Eds.), *Self and motivation: Emerging psychological perspectives* (pp. 147–171). American Psychological Association. https://doi.org/10.1037/10448-006

Spears, R., Lea, M., & Lee, S. (1990). De-individuation and group polarization in computer-mediated communication. *British Journal of Social Psychology, 29*(2), 121–134. https://doi.org/10.1111/j.2044-8309.1990.tb00893.x

Stassen, H. M., & Bates, B. R. (2020). Beers, Bros, and Brett: Memes and the visual ideograph of the <Angry White Man>. *Communication Quarterly, 68*(3), 331–354. https://doi.org/10.1080/01463373.2020.1787477

Ștefăniță, O., Corbu, N., & Buturoiu, R. (2018). Fake news and the third-person effect: They are more influenced than me and you. *Journal of Media Research, 11*(3(32)), 5–23. https://doi.org/10.24193/jmr.32.1

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J., & Loker, K. (2019). Who shared it?: Deciding what news to trust on social media. *Digital Journalism, 7*(6), 783–801. https://doi.org/10.1080/21670811.2019.1623702

Su, Y. (2021). It doesn't take a village to fall for misinformation: Social media use, discussion heterogeneity preference, worry of the virus, faith in scientists, and COVID-19-related misinformation beliefs. *Telematics and Informatics, 58*, Article 101547. https://doi.org/10.1016/j.tele.2020.101547

Sumner, E. M., Ruge-Jones, L., & Alcorn, D. (2017). A functional approach to the Facebook Like button: An exploration of meaning, interpersonal functionality, and potential alternative response buttons. *New Media & Society, 20*(4), 1451–1469. https://doi.org/10.1177/1461444817697917

Sun, Y., Chia, S. C., Lu, F., & Oktavianus, J. (2022). The battle is on: factors that motivate people to combat anti-vaccine misinformation. *Health Communication, 37*(3), 327–336. https://doi.org/10.1080/10410236.2020.1838108

Sun, Y., Oktavianus, J., Wang, S., & Lu, F. (2022). The role of influence of presumed influence and anticipated guilt in evoking social correction of COVID-19 misinformation. *Health Communication, 37*(11), 1368–1377. https://doi.org/10.1080/10410236.2021.1888452

Süssenbach, P., Rees, J., & Gollwitzer, M. (2019). When the going gets tough, individualizers get going: On the relationship between moral foundations and

prosociality. *Personality and Individual Differences, 136*, 122–131. https://doi.org/10.1016/j.paid.2018.01.019

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science, 4*(3), Article 160802. https://doi.org/10.1098/rsos.160802

Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. (2020). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology, 41*(1), 21–34. https://doi.org/10.1111/pops.12586

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson Education UK.

Taber, C. S., & Lodge, M. (2012). Motivated skepticism in the evaluation of political beliefs (2006). *Critical Review, 24*(2), 157–184. https://doi.org/10.1080/08913811.2012.711019

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Political psychology: Key readings* (pp. 276–293). Psychology Press. https://doi.org/10.4324/9780203505984-16

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services, 51*, 72–82. https://doi.org/10.1016/j.jretconser.2019.05.026

Tamborini, R., Bowman, N. D., Prabhu, S., Hahn, L., Klebig, B., Grall, C., & Novotny, E. (2018). The effect of moral intuitions on decisions in video game play: The impact of chronic and temporary intuition accessibility. *New Media and Society, 20*(2), 564–580. https://doi.org/10.1177/1461444816664356

Tamborini, R., Hahn, L., Aley, M., Prabhu, S., Baldwin, J., Sethi, N., Novotny, E., Klebig, B., & Hofer, M. (2020). The impact of terrorist attack news on moral intuitions. *Communication Studies, 71*(4), 511–527. https://doi.org/10.1080/10510974.2020.1735467

Tamborini, R., Prabhu, S., Lewis, R. J., Grizzard, M., & Eden, A. (2018). The influence of media exposure on the accessibility of moral intuitions and associated affect. *Journal of Media Psychology, 30*(2), 79–90. https://doi.org/10.1027/1864-1105/a000183

Tandoc, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism, 21*(3), 381–398. https://doi.org/10.1177/1464884919868325

The Electoral Commission. (2019). *Results and turnout at the 2018 May England local elections.* https://www.electoralcommission.org.uk/who-we-are-and-what-we-do/elections-and-referendums/past-elections-and-referendums/england-local-council-elections/results-and-turnout-2018-may-england-local-elections

The Mayor's Office for Policing and Crime. (2019). *What Londoners tell us around knife crime and violence.* https://www.london.gov.uk/moderngovmb/documents/s63353/Appendix B-MOPAC Surveys presentation.pdf

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal, 94*(6), 1395–1415. https://doi.org/10.2307/796133

TikTok. (2020, June 18). *How TikTok recommends videos #ForYou.* https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64,* 231–255. https://doi.org/10.1146/annurev-psych-113011-143812

Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science, 24*(12), 2454–2462. https://doi.org/10.1177/0956797613494848

Tong, C., Gill, H., Li, J., Valenzuela, S., & Rojas, H. (2020). "fake news is anything they say!"– Conceptualization and weaponization of fake news among the American public. *Mass Communication and Society, 23*(5), 755–778. https://doi.org/10.1080/15205436.2020.1789661

Trevors, G., & Duffy, M. C. (2020). Correcting COVID-19 Misconceptions Requires Caution. *Educational Researcher, 49*(7), 538–542. https://doi.org/10.3102/0013189X20953825

Trivedi, N., Krakow, M., Hyatt Hawkins, K., Peterson, E. B., & Chou, W.-Y. S. (2020). "Well, the message is from the institute of something": Exploring source trust of cancer-related messages on simulated Facebook posts. *Frontiers in Communication, 5,* Article 12. https://doi.org/10.3389/fcomm.2020.00012

Tsang, S. J. (2021). Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism and Mass Communication Quarterly, 98*(4), 1059–1077. https://doi.org/10.1177/1077699020952129

Tsang, S. J. (2022). Issue stance and perceived journalistic motives explain divergent audience perceptions of fake news. *Journalism, 23*(4), 823–840. https://doi.org/10.1177/1464884920926002

Turner, J. C., & Reynolds, K. J. (2012). Self-categorization theory. In P. van Lang, E. Higgins, & A. W. Kruglanski (Eds.), *Handbook of theories of social psychology: Vol. 2.* (pp. 399–417). SAGE Publications Ltd. https://doi.org/10.4135/9781446249222

Turner-Zwinkels, F. M., Johnson, B. B., Sibley, C. G., & Brandt, M. J. (2021). Conservatives' moral foundations are more densely connected than liberals' moral foundations. *Personality and Social Psychology Bulletin, 47*(2), 167–184. https://doi.org/10.1177/0146167220916070

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*(6), 479–491. https://doi.org/10.1017/S1930297500004022

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition, 6*(1), 13. https://doi.org/10.5334/joc.259

UK Government. (n.d.). *Share Checklist*. Retrieved 20 June 2023, from https://sharechecklist.gov.uk/

Valdesolo, P., & Desteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science, 18*(8), 689–690. https://doi.org/10.1111/j.1467-9280.2007.01961.x

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology, 44*(5), 1334–1338. https://doi.org/10.1016/j.jesp.2008.03.010

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 de Vyver, J., Houston, D. M., Abrams, D., & Vasiljevic, M. (2016). Boosting belligerence: how the July 7, 2005, London bombings affected liberals' moral foundations and prejudice. *Psychological Science, 27*(2), 169–177. https://doi.org/10.1177/0956797615615584

van der Lee, R., Ellemers, N., & Scheepers, D. (2016). Mastering moral misery: Emotional and coping responses to intragroup morality (vs. competence) evaluations. *Cognition and Emotion, 30*(1), 51–65. https://doi.org/10.1080/02699931.2015.1050357

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: Political bias in perceptions of fake news. *Media, Culture and Society*, *42*(3), 460–470. https://doi.org/10.1177/0163443720906992

Van Nunspeet, F., Ellemers, N., Derks, B., & Nieuwenhuis, S. (2014). Moral concerns increase attention and response monitoring during IAT performance: ERP evidence. *Social Cognitive and Affective Neuroscience, 9*(2), 141–149. https://doi.org/10.1093/scan/nss118

van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin, 134*(4), 504–535. https://doi.org/10.1037/0033-2909.134.4.504

Västfjäll, D., Slovic, P., Burns, W. J., Erlandsson, A., Koppel, L., Asutay, E., & Tinghög, G. (2016). The arithmetic of emotion: Integration of incidental and integral affect in judgments and decisions. *Frontiers in Psychology, 7*, Article 325. https://doi.org/10.3389/fpsyg.2016.00325

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

Voelkel, J. G., & Brandt, M. J. (2019). The effect of ideological identification on the endorsement of moral values depends on the target group. *Personality and Social Psychology Bulletin, 45*(6), 851–863. https://doi.org/10.1177/0146167218798822

Voelkel, J. G., & Feinberg, M. (2018). Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science, 9*(8), 917–924. https://doi.org/10.1177/1948550617729408

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

Walter, A. S., & Redlawsk, D. P. (2019). Voters' partisan responses to politicians' immoral behavior. *Political Psychology, 40*(5), 1075–1097. https://doi.org/10.1111/pops.12582

Wang, R., & Liu, W. (2021). Moral framing and information virality in social movements: A case study of #HongKongPoliceBrutality. *Communication Monographs*, *88*(3), 350–370. https://doi.org/10.1080/03637751.2021.1918735

Wang, S. (2021). Standing up or standing by: Bystander intervention in cyberbullying on social media. *New Media and Society, 23*(6), 1379–1397. https://doi.org/10.1177/1461444820902541

Wang, X., Lei, L., Liu, D., & Hu, H. (2016). Moderating effects of moral reasoning and gender on the relation between moral disengagement and cyberbullying in adolescents. *Personality and Individual Differences, 98*, 244–249. https://doi.org/10.1016/j.paid.2016.04.056

Wang, X., & McClung, S. R. (2012). The immorality of illegal downloading: The role of anticipated guilt and general emotions. *Computers in Human Behavior, 28*(1), 153–159. https://doi.org/10.1016/j.chb.2011.08.021

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011). 'I regretted the minute I pressed share': A qualitative study of regrets on Facebook. *Proceedings of the 7th Symposium on Usable Privacy and Security (SOUPS 2011), 10,* 1-16. https://doi.org/10.1145/2078827.2078841

Wardle, C. (2019). *Understanding information disorder*. First Draft. https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x21167

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology, 49*(6), 1027–1033. https://doi.org/10.1016/j.jesp.2013.07.002

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

Weinzierl, M. A., & Harabagiu, S. M. (2022). From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. *Proceedings of the International AAAI Conference on Web and Social Media, 16*(1), 1087–1097. https://ojs.aaai.org/index.php/ICWSM/article/view/19360/19132

Whiting, A., & Williams, D. (2013). Why people use social media: A uses and gratifications approach. *Qualitative Market Research: An International Journal, 16*(4), 362–369. https://doi.org/10.1108/QMR-06-2013-0041

Wilder, D. A., & Shapiro, P. N. (1984). Role of out-group cues in determining social identity. *Journal of Personality and Social Psychology, 47*(2), 342–348. https://doi.org/10.1037/0022-3514.47.2.342

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research, 47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Williamson, V. (2016). On the ethics of crowdsourced research. *PS - Political Science and Politics, 49*(1), 77–81. https://doi.org/10.1017/S104909651500116X

Willis, H. H. (2005). *Estimating terrorism risk.* RAND. https://doi.org/10.7249/MG388

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In Zanna, M. P. (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 345–411). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(03)01006-2

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology, 65*, 7–19. https://doi.org/10.1016/j.jesp.2016.02.005

Wong, R. Y. M., Cheung, C. M. K., Xiao, B., & Thatcher, J. B. (2021). Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research, 32*(2), 561–581. https://doi.org/10.1287/ISRE.2020.0983

Woong Yun, G., & Park, S. Y. (2011). Selective Posting: Willingness to post a message online. *Journal of Computer-Mediated Communication, 16*(2), 201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x

Wu, T. Y., & Atkin, D. J. (2018). To comment or not to comment: Examining the influences of anonymity and social support on one's willingness to express in online news discussions. *New Media and Society, 20*(12), 4512–4532. https://doi.org/10.1177/1461444818776629

Xia, Y., Lukito, J., Zhang, Y., Wells, C., Kim, S. J., & Tong, C. (2019). Disinformation, performed: Self-presentation of a Russian IRA account on Twitter. *Information Communication and Society, 22*(11), 1646–1664. https://doi.org/10.1080/1369118X.2019.1621921

Xiao, X., Su, Y., & Lee, D. K. L. (2021). Who consumes new media content more wisely? Examining personality factors, SNS use, and new media literacy in the era of

misinformation. *Social Media + Society, 7*(1). https://doi.org/10.1177/2056305121990635

Yang, F., & Horning, M. (2020). Reluctant to share: How third person perceptions of fake news discourage news readers from sharing "real news" on social media. *Social Media + Society, 6*(3). https://doi.org/10.1177/2056305120955173

Yang, J., & Tian, Y. (2021). "Others are more vulnerable to fake news than I Am": Third-person effect of COVID-19 fake news on social media users. *Computers in Human Behavior, 125*. https://doi.org/10.1016/j.chb.2021.106950

Ybarra, O., Chan, E., & Park, D. (2001). Young and old adults' concerns about morality and competence. *Motivation and Emotion, 25*(2), 85–100. https://doi.org/10.1023/A:1010633908298

Yıldırım, M., & Güler, A. (2020). Factor analysis of the COVID-19 Perceived Risk Scale: A preliminary study. *Death Studies, 46*(5) 1065-1072. https://doi.org/10.1080/07481187.2020.1784311

YouGov. (2021). YouGov Coronavirus Handling Confidence Tracker. https://docs.cdn.yougov.com/ixbnz74wqx/YouGov_CoronaConfidence_Tracker_W.pdf

Young, J. C. (2021). Disinformation as the weaponization of cruel optimism: A critical intervention in misinformation studies. *Emotion, Space and Society, 38,* Article 100757. https://doi.org/10.1016/j.emospa.2020.100757

Young, R., Kananovich, V., & Johnson, B. G. (2023). Young adults' folk theories of how social media harms its users. *Mass Communication and Society, 26*(1), 23–46. https://doi.org/10.1080/15205436.2021.1970186

Young, R., Zhang, L., & Prybutok, V. R. (2007). Hacking into the minds of hackers. *Information Systems Management, 24*(4), 281–287. https://doi.org/10.1080/10580530701585823

Zhang, J. (2010). Self-enhancement on a self-categorization leash: Evidence for a dual-process model of first- and third-person perceptions. *Human Communication Research, 36*(2), 190–215. https://doi.org/10.1111/j.1468-2958.2010.01373.x

Appendices

Appendix A

Ethics Application for Study One (Pilot)

Figure A1

Ethics Application Decision Letter for Pilot Study

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Doctoral Research Project

Application ID: ETH2021-0737

Date: 05 Jan 2021

Dear Laura

I am writing to inform you that your application was considered by the Psychology Ethics Committee.

The proposal was approved.

The expiry date for this proposal is 26 Feb 2021.

Yours,

Samuel Evans

Psychology Ethics Committee

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

The desirability of including full details of the consent form in an appendix to your research, and of addressing specifically ethical issues in your methodological discussion.

The requirement to furnish the Research Ethics Committee with details of the conclusion and outcome of the project, and to inform the Research Ethics Committee should the research be discontinued. The Committee would prefer a concise summary of the conclusion and outcome of the project, which would fit no more than one side of A4 paper, please.

Figure A2

Participant Invitation Letter for Pilot Study

Social Media Pilot Study
Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to pre-test materials for future use in research looking at what may lead individuals to interact with false information on social media

Who can take part?

We are looking for adults over the age of 18 who live in England

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide an explanation. You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts. Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete a short questionnaire to provide some basic details about yourself, for example your age and location.

After this you will be asked to rate a selection of social media posts and images. For one set of images, we are interested in understanding how favourable they are towards the UK Government. For the second set of images, we are interested in how much of a risk you feel the content makes COVID-19 appear.

All of the images you will see are drawn from social media. They were not created by the University and we do not endorse the information they contain.

How long will it take?

The whole study should take about 10 minutes

What are the possible disadvantages and risks of taking part?

There are no anticipated disadvantages or risks to your participation

What are the possible benefits of taking part?

Your contribution will help to increase our understanding of why individuals interact with misleading and false information on social media

What if something goes wrong?

This research has been approved by the University of Westminster Psychology Ethics Committee. If you have any questions or concerns about this research you can contact:

Dr Samuel Evans (Chair of the ethics committee)

S.Evans1@westminster.ac.uk or phone: 020 7911 5000

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and / or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner : laura.campbell.joyner@my.westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan : T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili : O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please reach out to the experimenter if you have any questions.

Figure A3

Debrief for Pilot Study

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to test materials that will be used in future studies looking at how individuals interact with false information on social media.

The materials you viewed contained false or misleading information and therefore are not factual

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

NHS COVID-19 Information:

The NHS website provides free and reputable information on COVID-19 symptoms and advice.
For more information visit: https://www.nhs.uk/conditions/coronavirus-covid-19/

FullFact:

FullFact is a team of independent and impartial fact checkers. Their website provides up to date fact checks and will accept information you may need fact checked via their contact page.
For more information visit: https://fullfact.org

SHARE Checklist:

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

First Draft:

First Draft is an independent organization who provide guidance on how to verify content sourced from the internet. They provide a number of free tools, guides and courses
For more information visit: https://firstdraftnews.org

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:
Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:
Chair of Ethics: S.Evans1@westminster.ac.uk

Appendix B

Ethics Application for Studies One (Main) and Three

Figure B1

Ethics Application Decision Letter for Main Study

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Doctoral Research Project

Application ID: ETH2021-0777

Date: 13 Jan 2021

Dear Laura

I am writing to inform you that your application was considered by the Psychology Ethics Committee.

The proposal was approved.

The expiry date for this proposal is 04 Jan 2024.

Yours,

Samuel Evans

Psychology Ethics Committee

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

The desirability of including full details of the consent form in an appendix to your research, and of addressing specifically ethical issues in your methodological discussion.

The requirement to furnish the Research Ethics Committee with details of the conclusion and outcome of the project, and to inform the Research Ethics Committee should the research be discontinued. The Committee would prefer a concise summary of the conclusion and outcome of the project, which would fit no more than one side of A4 paper, please.

Figure B2

Participant Invitation Letter for Study One (Main)

Beliefs and interaction with COVID-19 content on social media

Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to understand whether personal beliefs surrounding the COVID-19 crisis influence engagement with related content on social media

Who can take part?

We are looking for adults over the age of 18 who live in England and have a Facebook account

Do I have to take part?

No your participation is entirely voluntary. You can stop taking part at anytime without having to provide an explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete some demographic items (for example your age, level of education and political affiliations) and to fill in two short questionnaires about political trust and COVID-19 risk perceptions. You will then be shown a short series of images and asked to rate the likelihood of engaging with them on Facebook if they appeared on your feed. Finally you will be asked to rate the moral acceptability of sharing the same images on Facebook.

All of the images you will see are drawn from social media. They were not created by the University and we do not endorse the information they contain.

How long will it take?

The whole study should take about 10 minutes

What are the possible disadvantages and risks of taking part?

There are no anticipated disadvantages or risks to your participation

What are the possible benefits of taking part?

Your contribution will help to increase our understanding of why individuals interact with misleading and false information on social media

What if something goes wrong?

This research has been approved by the University of Westminster Psychology Ethics Committee.
If you have any questions or concerns about this research you can contact:
Dr Samuel Evans (Chair of the ethics committee)
S.Evans1@westminster.ac.uk or phone: 020 7911 5000

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.
If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.
All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.
You will not be personally identifiable in any reports that arise from this study.
Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.
Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and / or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:
Laura Joyner : laura.campbell.joyner@my.westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan : T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili : O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please reach out to the experimenter if you have any questions.

Figure B3

Debrief for Study One (Main)

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to understand whether personal beliefs of trust and risk may influence engagement and moral perceptions of misleading content.

The materials you viewed contained false or misleading information and therefore are not factual

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

NHS COVID-19 Information :

The NHS website provides free and reputable information on COVID-19 symptoms and advice.
For more information visit: https://www.nhs.uk/conditions/coronavirus-covid-19/

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:
Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk
Project Supervisors: T.Buchanan@westminster.ac.uk / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:
Chair of Ethics: S.Evans1@westminster.ac.uk

Please click the arrow below to complete the study and return to Prolific.

Appendix C

Citizen Trust in Government Scale - Grimmelikhuijsen & Knies, 2015

McKnight et al., 2002 items	Our items	In final scale
NO PREFIX	ADDED PREFIX: When it concerns [air quality policy] . . .	*
Overall, LegalAdvice.com is a capable and proficient Internet legal advice provider.	COMP1: [The municipality of XX] is capable.*	*
LegalAdvice.com is competent and effective in providing legal advice.	COMP2: [The municipality of XX] is effective.	
	COMP3: [The municipality of XX] is skilful.	
In general, LegalAdvice.com is very knowledgeable about the law.	COMP4: [The municipality of XX] is expert.*	*
LegalAdvice.com performs its role of giving legal advice very well.	COMP5: [The municipality of XX] carries out its duty very well.*	*
If required help, LegalAdvice.com would do its best to help me.	BEN1: If citizens need help, [the municipality of XX] will do its best to help them.*	*
I believe that LegalAdvice.com would act in my best interest.	BEN2: [The municipality of XX] acts in the interest of citizens.*	*
LegalAdvice.com is interested in my well-being, not just its own.	BEN3: [The municipality of XX] is genuinely interested in the well-being of citizens.*	*
LegalAdvice.com is truthful in its dealings with me.	INT1: [The municipality of XX] approaches citizens in a sincere way.*	*
LegalAdvice.com is sincere and genuine.	INT2: [The municipality of XX] is sincere.*	*
LegalAdvice.com would keep its commitments.	INT3: [The municipality of XX] keeps its commitments.	
I would characterize LegalAdvice.com as honest.	INT4: [The municipality of XX] is honest.*	*

Appendix D

COVID-19 Perceived Risk Scale - Yıldırım & Güler, 2020

Please read each of the following statement and choose an answer that best describes you.

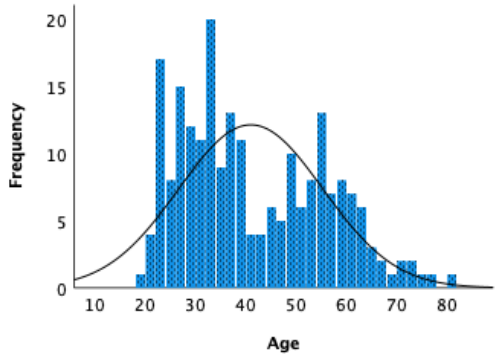
1. What is the likelihood that you would acquire the COVID-19?	1	2	3	4	5
2. What is the likelihood that you would acquire the COVID-19 compared to other persons?	1	2	3	4	5
3. What is the likelihood that you would catch other diseases (e.g., diabetes/asthma).	1	2	3	4	5
4. What is the likelihood that you would die from the COVID-19?	1	2	3	4	5
5. How worried are you about contracting the COVID-19?	1	2	3	4	5
6. How worried are you about a family member contracting the COVID-19?	1	2	3	4	5
7. How worried are you about the COVID-19 occurring in your region?	1	2	3	4	5
8. How worried are you about the COVID-19 emerging as a health issue?	1	2	3	4	5

Note. The first four questions refer to cognitive dimension of perceived risk and the remaining four refer to emotional dimension of perceived risk, with higher scores indicating greater risk associated with COVID-19. A total score can be obtained by summing all items on the scale.

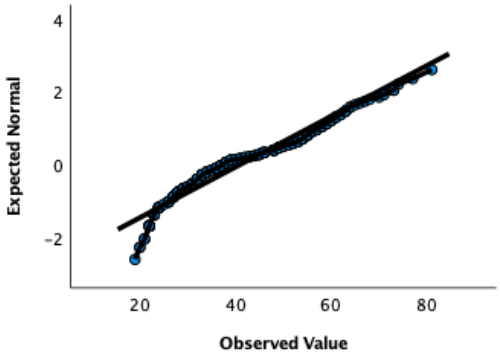
Appendix E

Histograms and Q-Q Plots for Main Variables (Study One)

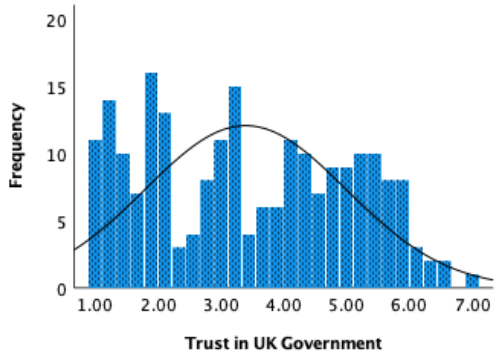
E1



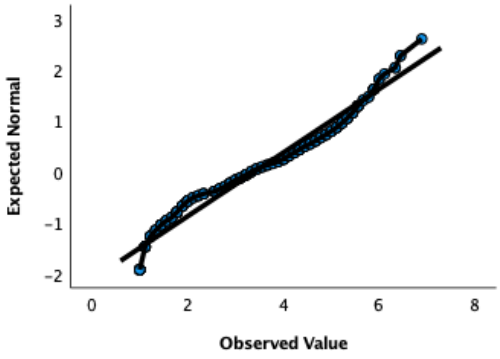
E2



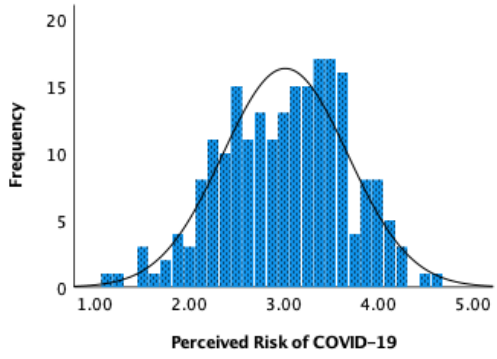
E3



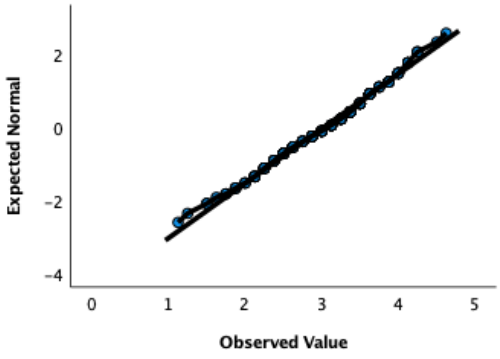
E4



E5



E6

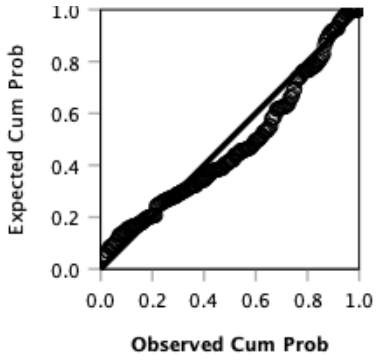


Note. Panel E1. Histogram of Age. Panel E2. Normal Q-Q Plot of Age. Panel E3. Histogram of Trust in UK Government. Panel E4. Normal Q-Q Plot of Trust in UK Government. Panel E5. Histogram of Perceived Risk of COVID-19. Panel E6. Normal Q-Q Plot of Perceived Risk of COVID-19.

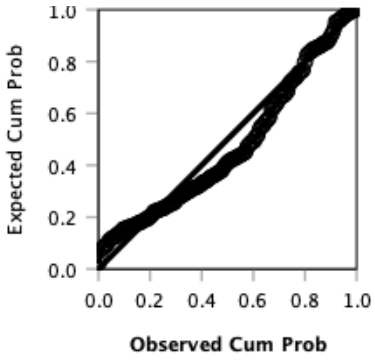
Appendix F

P-P Plots of Residuals for Planned Regressions

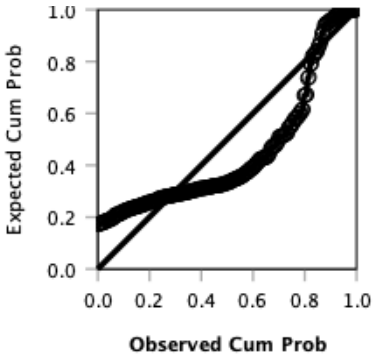
F1



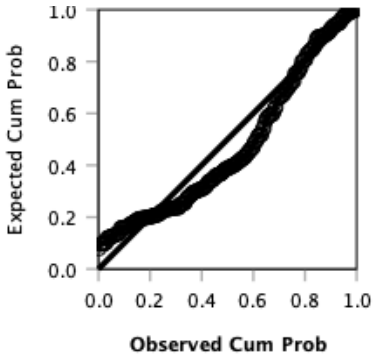
F2



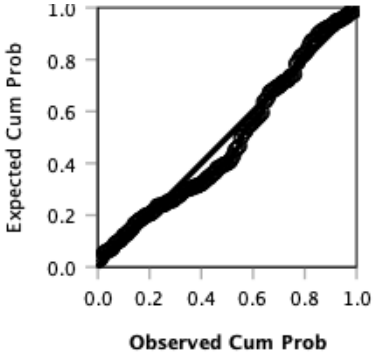
F3



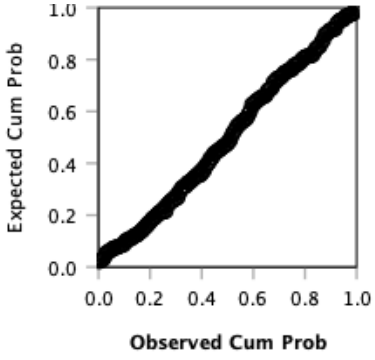
F4



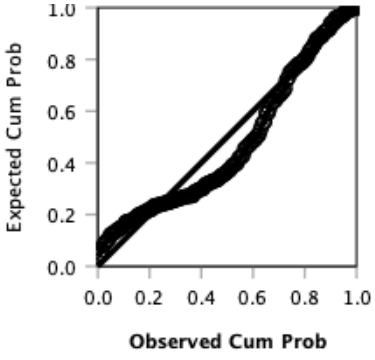
F5



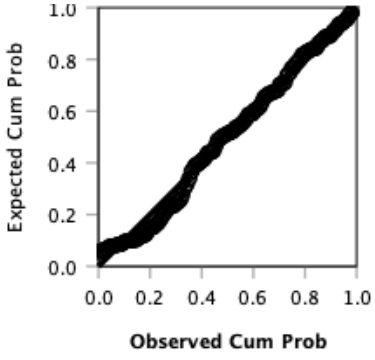
F6



F7



F8



Note. Panels F1-F4. Plots for Interactions with Unfavourable, Favourable, Minimising & Maximising misinformation respectively. Panels F5-F8. Plots for Moral judgements of Unfavourable, Favourable, Minimising & Maximising disinformation respectively.

Appendix G

Ethics Application for Study Two (Pilot & Main)

Figure G1

Ethics Application Conditions Letter

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Doctoral Research Project

Application ID: ETH2021-1792

Date: 22 Apr 2021

Dear Laura

I am writing to inform you that your application was considered by the Psychology Ethics Committee.

The proposal was approved subject to the following:

- Please replace the information in the "what is something goes wrong" section of the information sheet with:

"This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:

Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk"

To reflect recent changes in the ethics procedures.

- Please remove the chair of the psychology ethics committee as a point of contact from the debrief sheet and replace with Dibyesh Anand in case of complaint - as per above.

Please submit the above documentation or clarifications via the VRE no later than 22 May 2021 or at your earliest convenience.

Yours,

Samuel Evans

Psychology Ethics Committee

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

Figure G2

Ethics Application Decision Letter

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Doctoral Research Project

Application ID: ETH2021-1792

Date: 22 Apr 2021

Dear Laura

Thank you for providing your response to the Conditions set by the Committee.

Your response to Conditions has been considered and your proposal is approved.

If your protocol changes significantly in the meantime, please contact me immediately, in case of further ethical requirements.

Yours,

Samuel Evans

Psychology Ethics Committee

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

The desirability of including full details of the consent form in an appendix to your research, and of addressing specifically ethical issues in your methodological discussion.

The requirement to furnish the Research Ethics Committee with details of the conclusion and outcome of the project, and to inform the Research Ethics Committee should the research be discontinued. The Committee would prefer a concise summary of the conclusion and outcome of the project, which would fit no more than one side of A4 paper, please.

Figure G3

Participant Invitation Letter for Study Two Pilot

Social Media Pilot Study

Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to pilot materials for research looking at how we make moral judgements of misinformation on social media.

Who can take part?

We are looking for adults over the age of 18 who live in London, identity as either Labour or Conservative voters and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide and explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete a short questionnaire to provide some basic details about yourself (for example your age and level of education). You will then be shown four simulated images and asked whether you feel they are favourable or unfavourable towards the named party. These are created for the purpose of the experiment and do not reflect the position of any political party or the University.

How long will it take?

The whole study should take around 3 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:
Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner: laura.campbell.joyner@my.westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan: T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili: O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure G4

Debrief for Study Two Pilot

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to test materials that will be used in future studies looking at how individuals make moral judgements of false information on social media.

The materials you viewed were created for the study and contained false information.

This means they are NOT factual and do not represent the performance of any political party or police force.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:

Head of School of Social Sciences: D.Anand@westminster.ac.uk

Figure G5

Participant Invitation Letter for Study Two (Main)

Moral Judgements on Social Media

Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to understand how moral judgements are made in relation to content within social media platforms

Who can take part?

We are looking for adults over the age of 18 who live in London, identity as either Labour or Conservative voters and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide an explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete some demographic items (for example your age and level of education). You will then be shown a simulated image and asked to make several moral judgements about sharing them on social media. These are created for the purpose of the experiment and do not reflect the position of any political party or the University.

How long will it take?

The whole study should take around 3 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:

Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the researcher or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study and may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner: laura.campbell.joyner@my.westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan: T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili: O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure G6

Debrief for Study Two (Main)

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to understand whether political identity influences our moral judgements surrounding the sharing of misinformation.

The materials you viewed were created for the study and contained false information.

This means they are NOT factual and do not represent the actual performance of any political party or police force.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:

Head of School of Social Sciences: D.Anand@westminster.ac.uk

Please click the arrow below to complete the study and return to Prolific.

Appendix H

Summary of Study Two Results with Excluded Participants

Table H1

Two-way ANOVA for Main Dependent Variables

	df	*F*	η_p^2
Moral – Unknown	1, 213	17.89***	.08
Moral – Known	1, 213	4.47*	.02
Reporting	1, 213	2.63	.01

Note. Interaction effect of target and stance. Analysis includes excluded participants

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

Table H2

Differences Between Moral Judgements of Unknown and Known Disinformation

	N	Unknown		Known		*t*	*d*
		M	*SD*	*M*	*SD*		
Ingroup Supporting	54	7.81	2.56	3.20	2.76	−10.51***	1.43
Ingroup Undermining	53	5.40	3.24	2.30	2.28	−6.19***	0.85
Outgroup Supporting	56	5.70	2.91	2.63	2.39	−7.80***	1.01
Outgroup Undermining	54	6.76	3.36	3.20	2.84	−7.42***	1.01

Note. M = Mean, SD = Standard Deviation. Analysis includes excluded participants

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

Table H3

Three-Way ANCOVA Statistics for Moral Acceptability Rating of Unknown Disinformation

	$\bar{x}^2$	F (1, 205)	η_p^2
Age	69.82	7.87**	.04
Gender	42.11	4.74*	.02
Stance	18.27	2.06	.01
Target	9.84	1.12	.00
Party	7.12	0.80	.00
Stance x Target	160.71	18.10***	.08
Stance x Party	1.82	0.21	.00
Target x Party	0.06	0.01	.00
Stance x Target x Party	5.14	0.58	.00

Note. Gender coded as dummy variable, F = 0, M = 1. Analysis includes excluded participants.

*p < .05. ** p < .01. *** p < .001.

Table H4

Three-Way ANCOVA Statistics for Moral Acceptability Rating of Unknown Disinformation

	$\bar{x}^2$	F (1, 205)	η_p^2
Age	9.26	1.53	.01
Gender	10.98	1.81	.01
Stance	2.80	0.46	.00
Target	2.70	0.45	.00
Party	10.86	1.80	.01
Stance x Target	36.36	6.01*	.03
Stance x Party	6.83	1.13	.01
Target x Party	11.75	1.94	.01
Stance x Target x Party	32.59	5.39*	.03

Note. Gender coded as dummy variable, F = 0, M = 1. Analysis includes excluded participants.

*p < .05. ** p < .01. *** p < .001.

Table H5

Three-Way ANCOVA Statistics for Reporting Likelihood of Known Disinformation

	$\bar{x}^2$	F (1, 205)	η_p^2
Age	14.52	1.09	.01
Gender	1.74	0.13	.00
Stance	1.02	0.08	.00
Target	15.04	1.13	.01
Party	51.29	3.85	.02
Stance x Target	39.84	2.99	.01
Stance x Party	0.22	0.02	.00
Target x Party	0.04	0.00	.00
Stance x Target x Party	28.9	2.17	.01

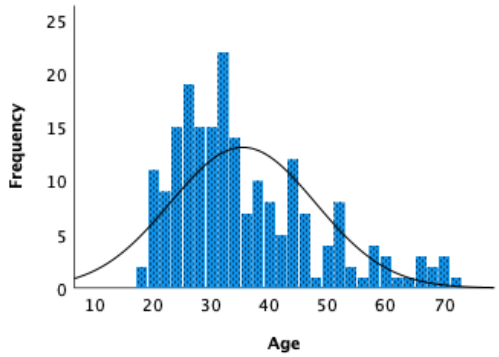
Note. Gender coded as dummy variable, F = 0, M = 1. Analysis includes excluded participants.

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

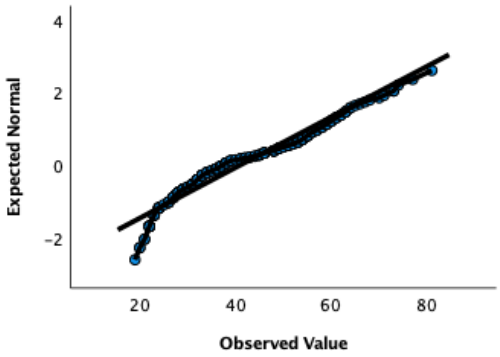
Appendix I

Histograms and Normal Q-Q Plots for Main Variables (Study Two)

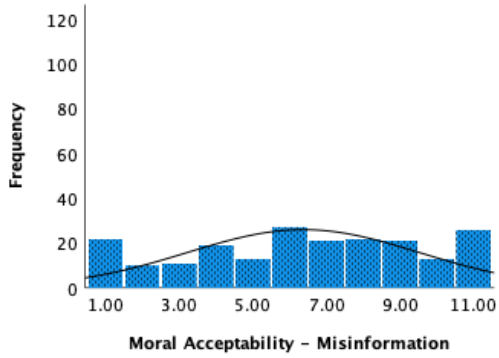
I1



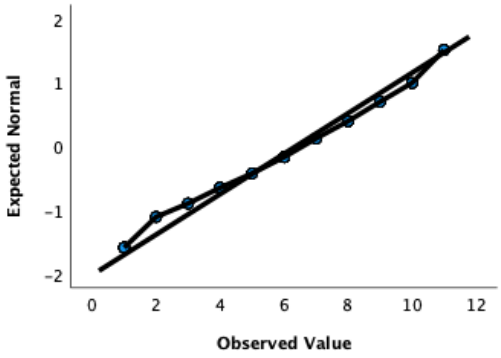
I2



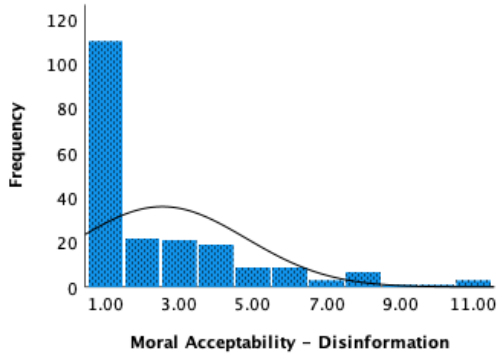
I3



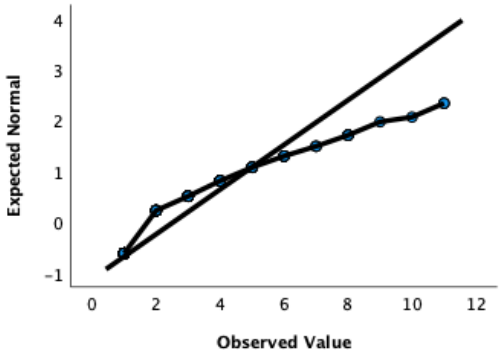
I4



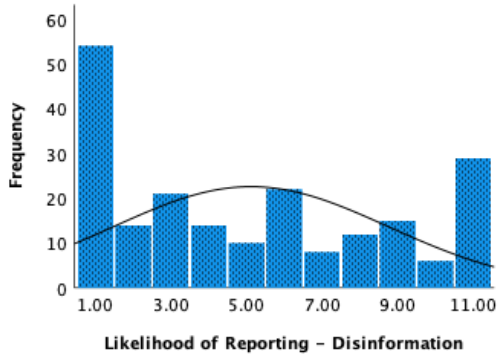
I5



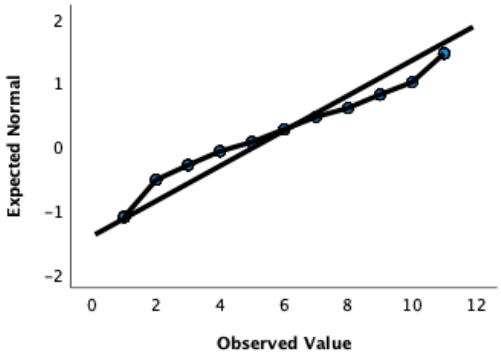
I6



I7



I8

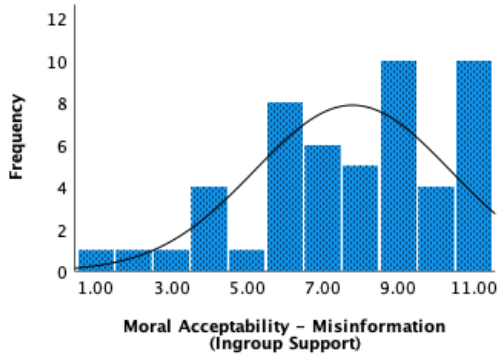


Note. Panel I1-I2. Normality Plots for Age. Panel I3-I4. Normality Plots for Moral Judgements for Misinformation. Panel I5-I6. Normality Plots of Moral Judgements for Disinformation. Panel I7-I8. Normality Plots of Likelihood for Reporting Disinformation.

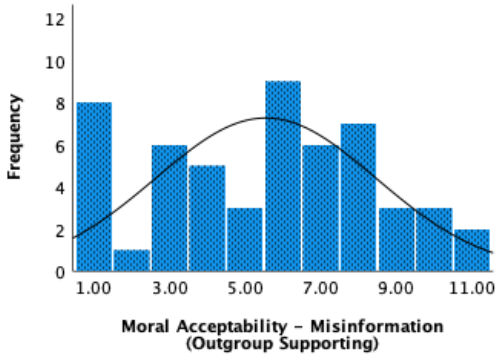
Appendix J

Histograms and Box Plots for Moral Judgements of Misinformation by Condition

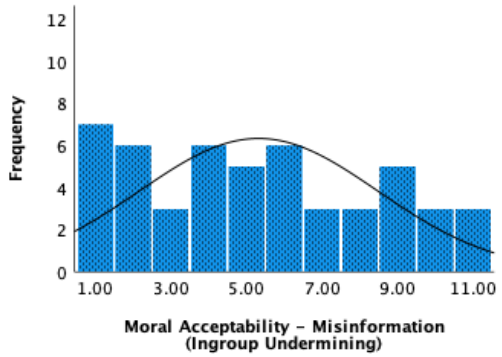
J1



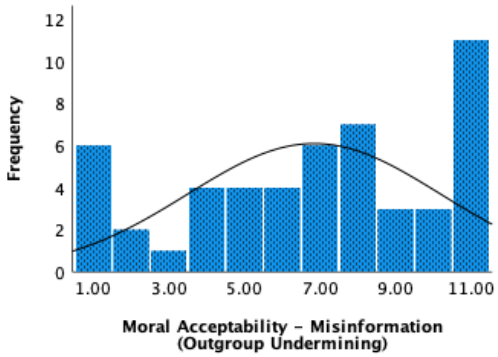
J2



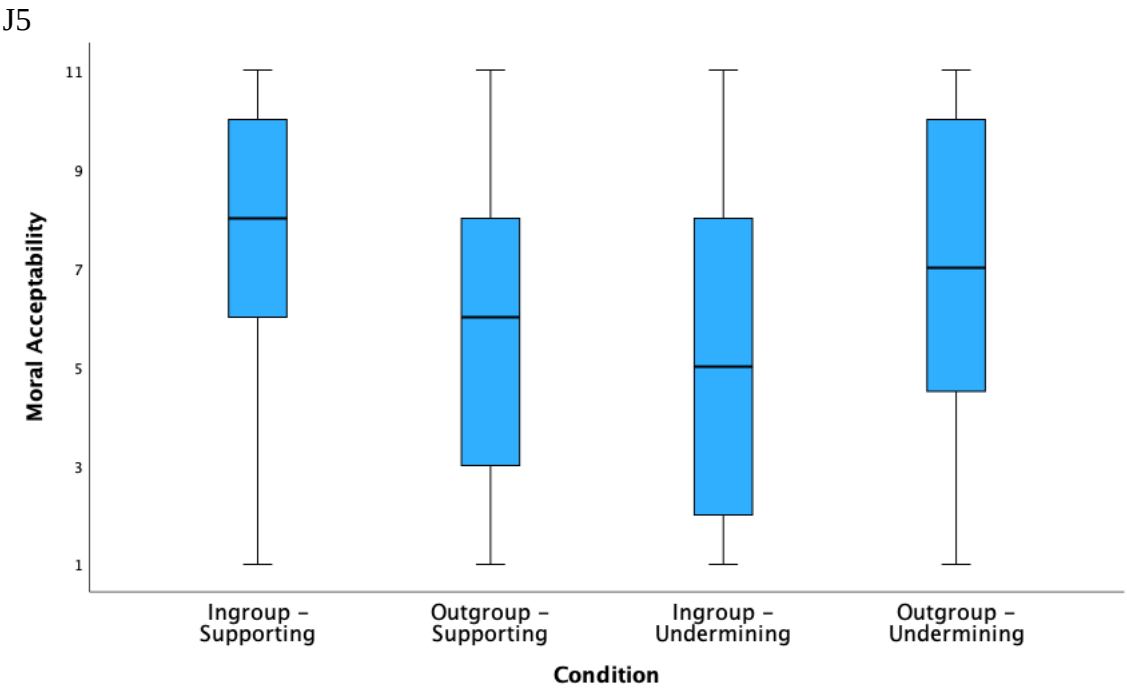
J3



J4



J5

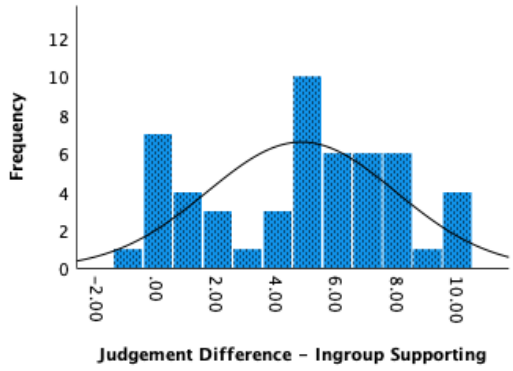


Note. Panel J1. Histogram for Ingroup Supporting Misinformation. Panel J2. Histogram for Outgroup Supporting Misinformation. Panel J3. Histogram for Ingroup Undermining Misinformation. Panel J4. Histogram for Outgroup Undermining Misinformation. Panel J5. Boxplots of Moral Judgements of Misinformation Across Each Condition.

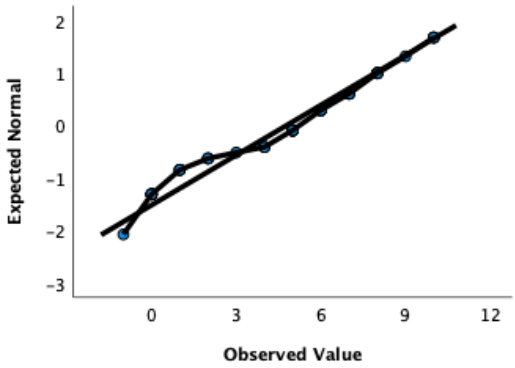
Appendix K

Histograms and Normal Q-Q Plots for Moral Judgement Change

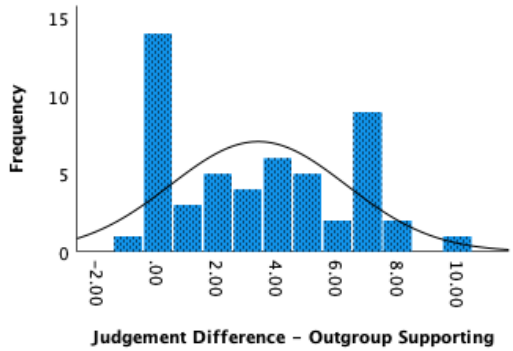
K1



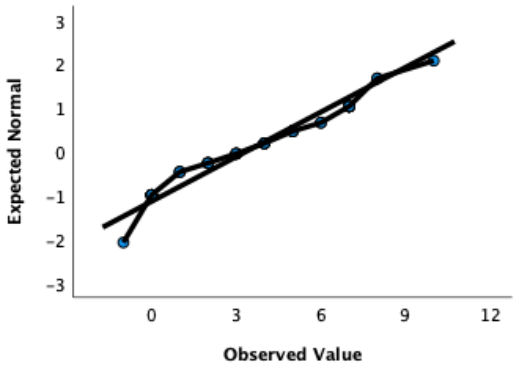
K2



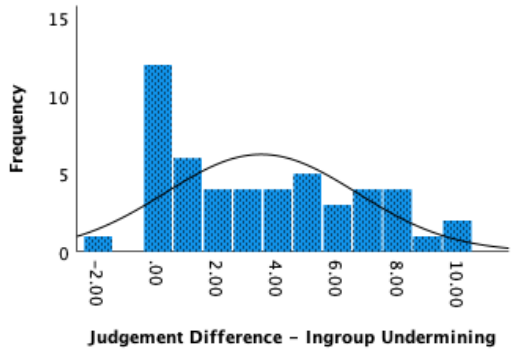
K3



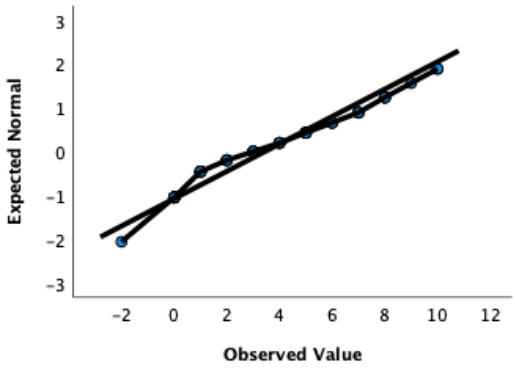
K4



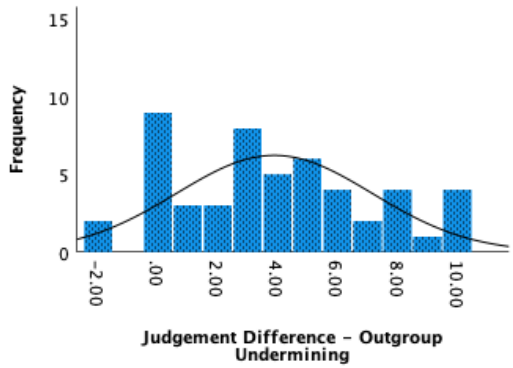
K5



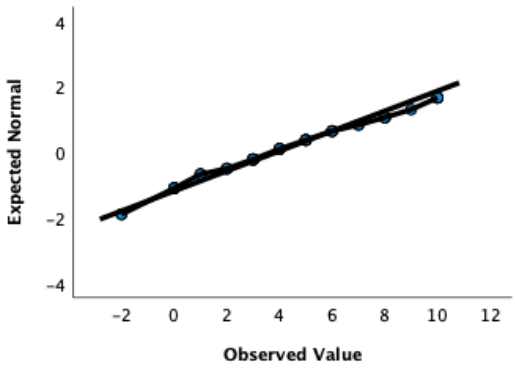
K6



K7



K8



Note. Panel K1-K2. Normality Plots for Ingroup Supporting. Panel K3-K4. Normality Plots for Outgroup Supporting. Panel K5-K6. Normality Plots for Ingroup Undermining. Panel K7-K8. Normality Plots for Outgroup Undermining.

Appendix L

Wilcoxon Signed Rank Tests of Differences Between Moral Judgements of Misinformation and Disinformation

	N	Unknown		Known		*z*	*r*
		M	*IQR*	*M*	*IQR*		
Ingroup Supporting	52	8.00	4.00	2.00	3.00	−5.82***	.81
Ingroup Undermining	50	5.00	6.00	1.00	2.00	−5.25***	.74
Outgroup Supporting	53	6.00	5.00	1.00	2.00	−5.26***	.72
Outgroup Undermining	51	7.00	6.00	2.00	3.00	−5.51***	.77

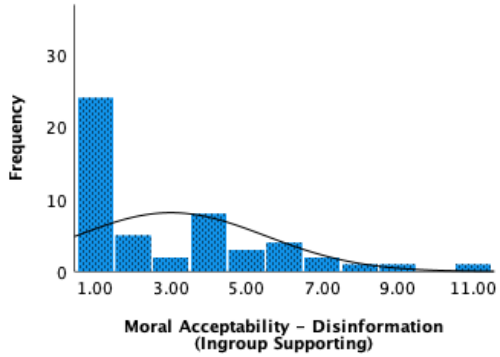
Note. Results of Wilcoxon Signed Rank Test. M = Median, IQR = Interquartile Range

*** $p < .001$.

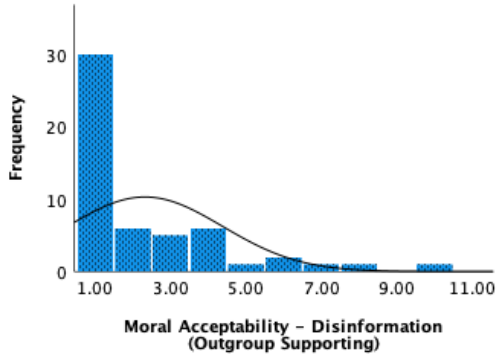
Appendix M

Histograms and Box Plots for Moral Judgements of Disinformation by Condition

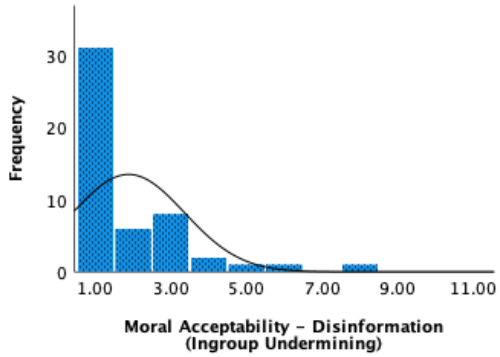
M1



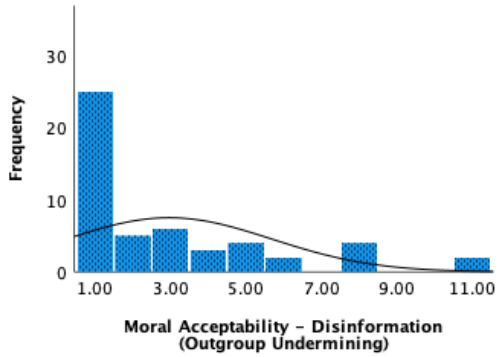
M2



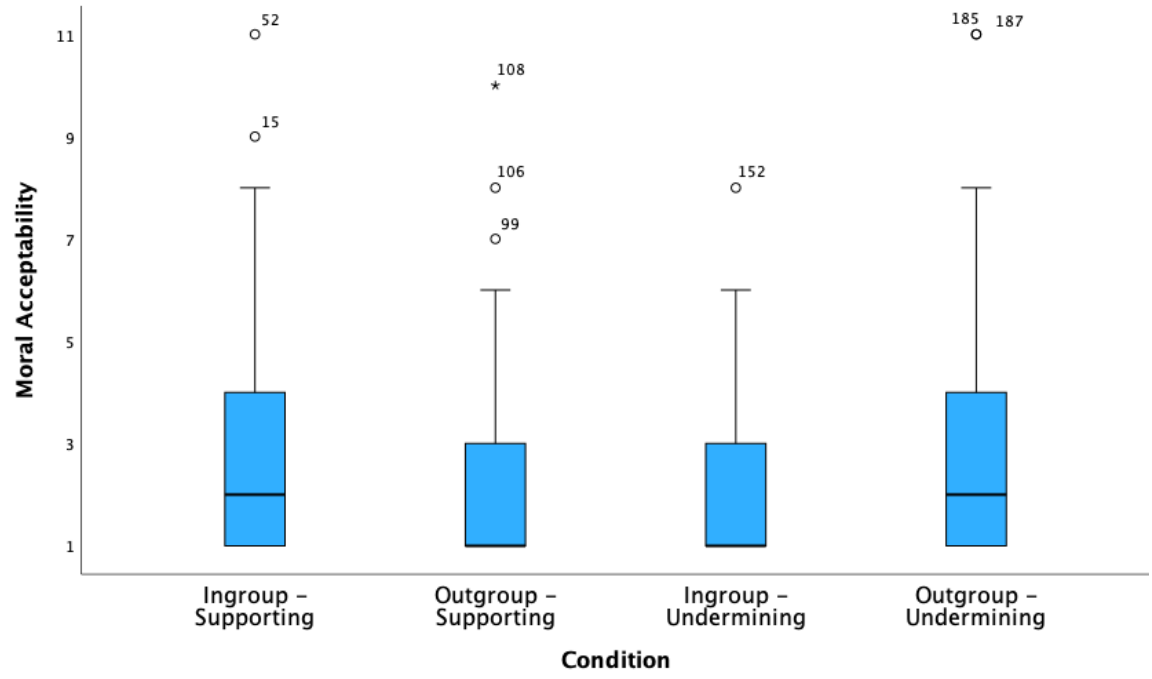
M3



M4



M5

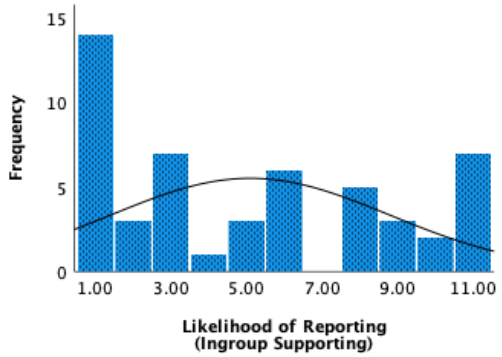


Note. Panel M1. Histogram for Ingroup Supporting Disinformation. Panel M2. Histogram for Outgroup Supporting Disinformation. Panel M3. Histogram for Ingroup Undermining Disinformation. Panel M4. Histogram for Outgroup Undermining Disinformation. Panel M5. Boxplots of Moral Judgements of Disinformation Across Each Condition.

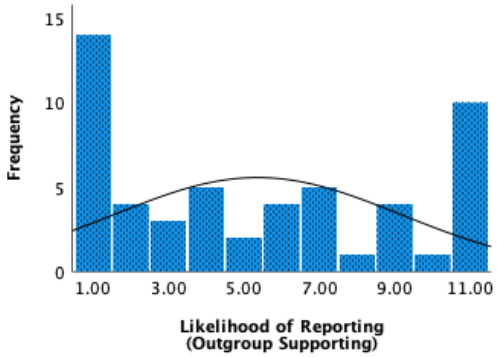
Appendix N

Histograms and Box Plots for Reporting Likelihood of Disinformation by Condition

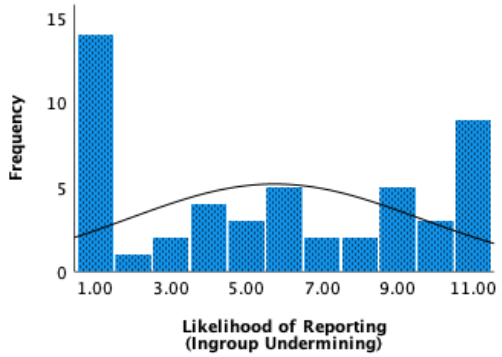
N1



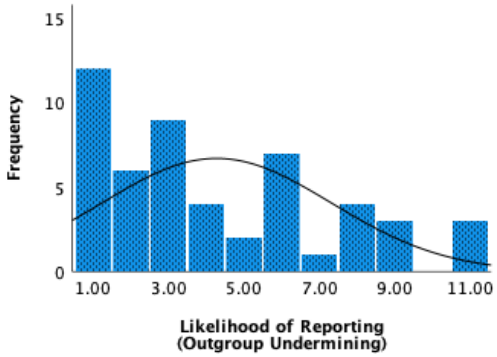
N2



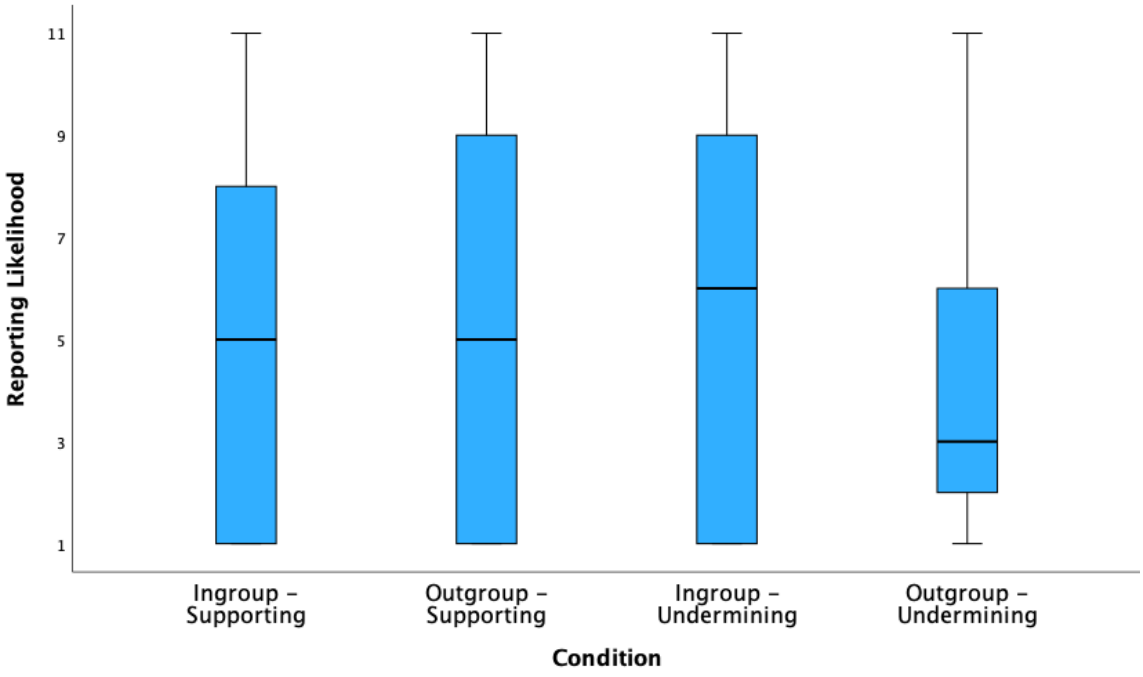
N3



N4



N5

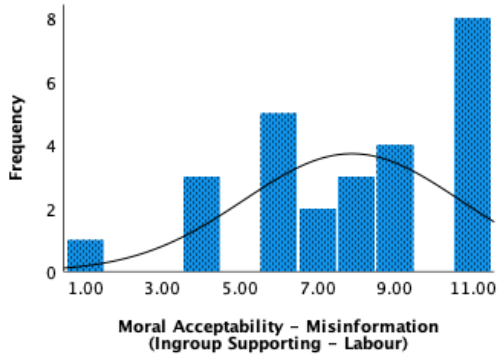


Note. Panel N1. Histogram for Ingroup Supporting Disinformation. Panel N2. Histogram for Outgroup Supporting Disinformation. Panel N3. Histogram for Ingroup Undermining Disinformation. Panel N4. Histogram for Outgroup Undermining Disinformation. Panel N5. Boxplots of Reporting Likelihood of Disinformation Across Each Condition.

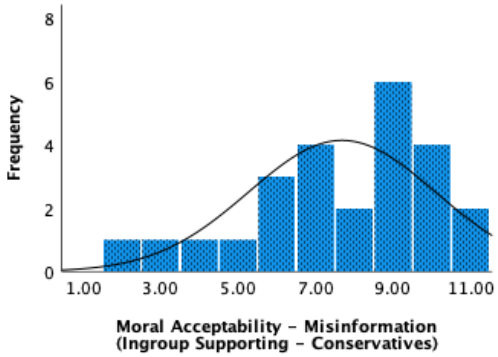
Appendix O

Histograms for Moral Judgements of Misinformation by Condition and Party

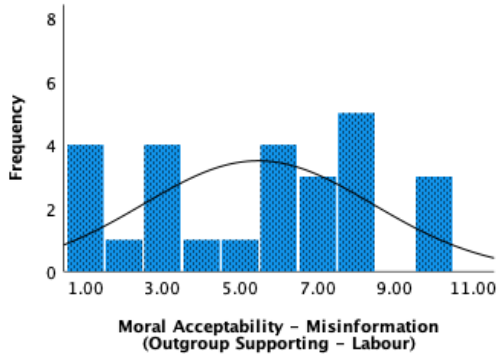
O1



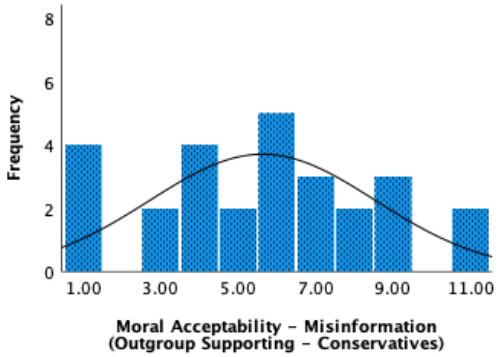
O2



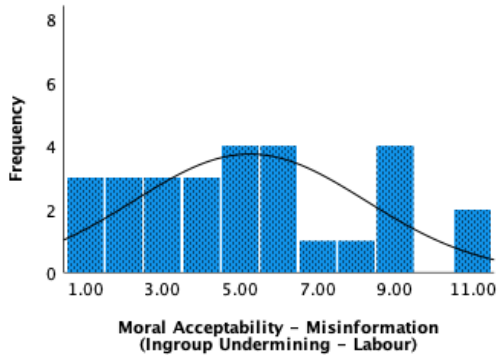
O3



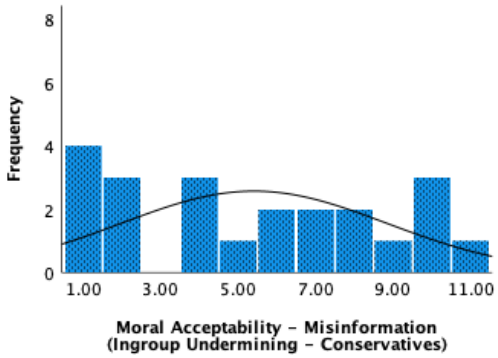
O4



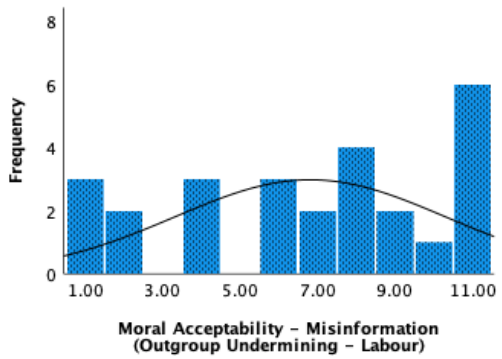
O5



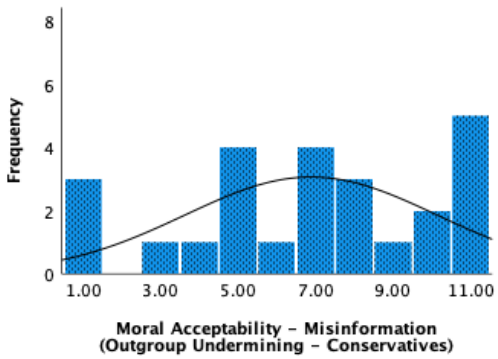
O6



O7



O8

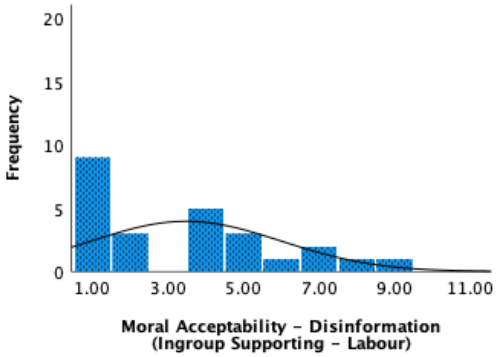


Note. Panel O1-O2. Ingroup Supporting Misinformation (Labour v. Conservative). Panel O3-O4. Outgroup Supporting Misinformation (Labour v. Conservative). Panel O5-O6. Ingroup Undermining Misinformation (Labour v. Conservative). Panel O7-O8. Outgroup Undermining Misinformation (Labour v. Conservative).

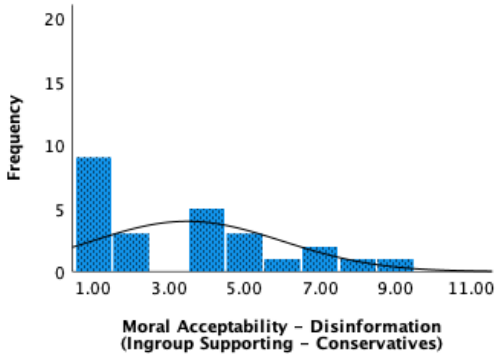
Appendix P

Histograms for Moral Judgements of Disinformation by Condition and Party

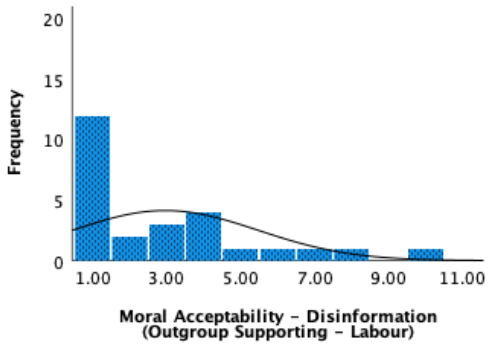
P1



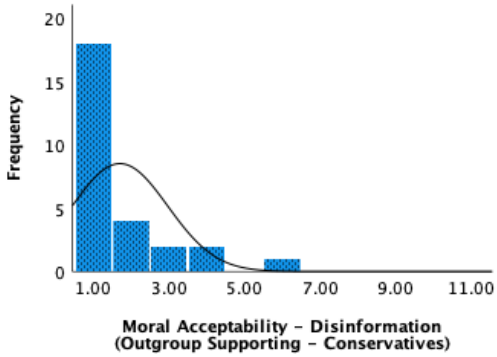
P2



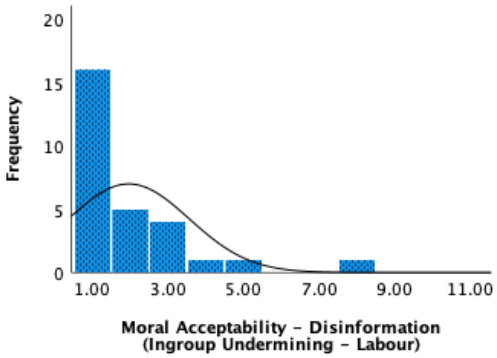
P3



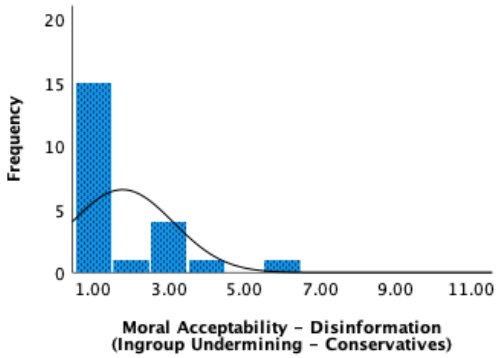
P4



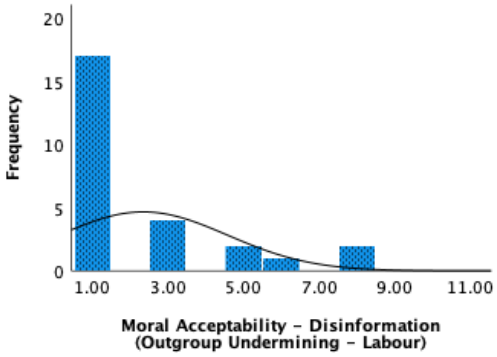
P5



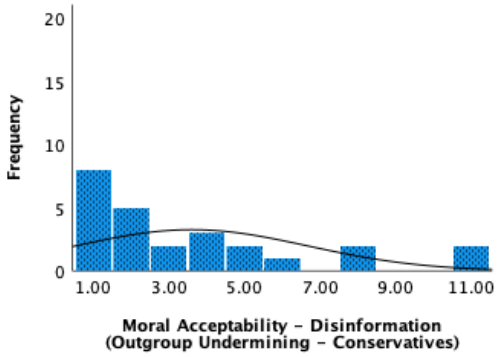
P6



P7



P8

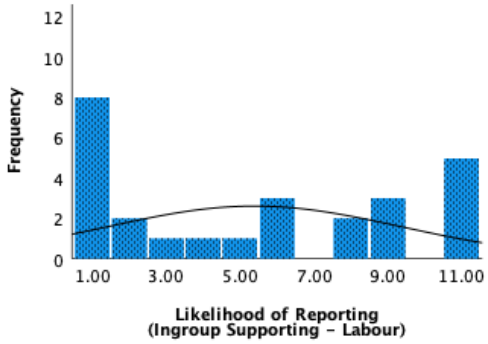


Note. Panel P1-P2. Ingroup Supporting Disinformation (Labour v. Conservative). Panel P3-P4. Outgroup Supporting Disinformation (Labour v. Conservative). Panel P5-P6. Ingroup Undermining Disinformation (Labour v. Conservative). Panel P7-P8. Outgroup Undermining Disinformation (Labour v. Conservative).

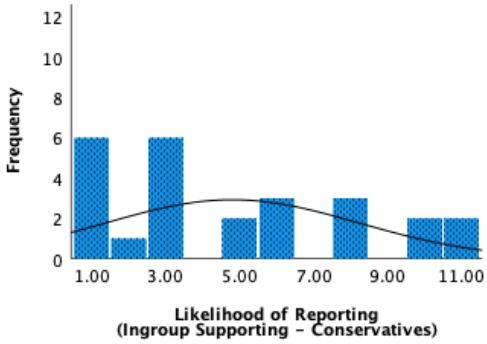
Appendix Q

Histograms for Reporting Likelihood of Disinformation by Condition and Party

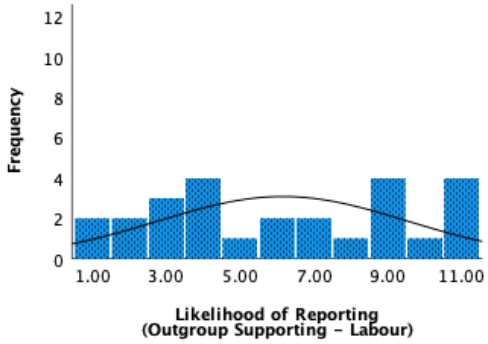
Q1



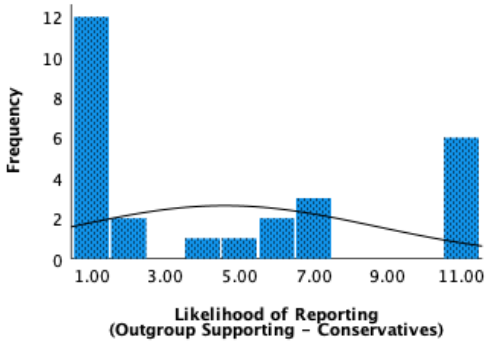
Q2



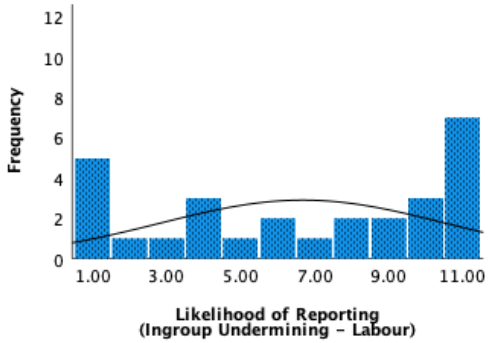
Q3



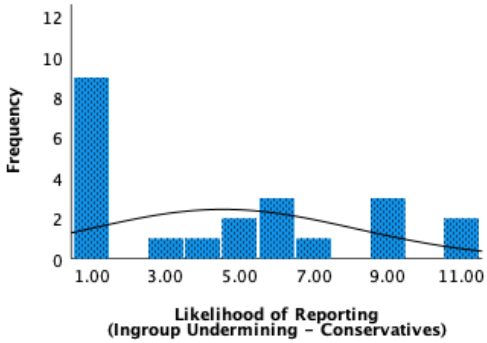
Q4



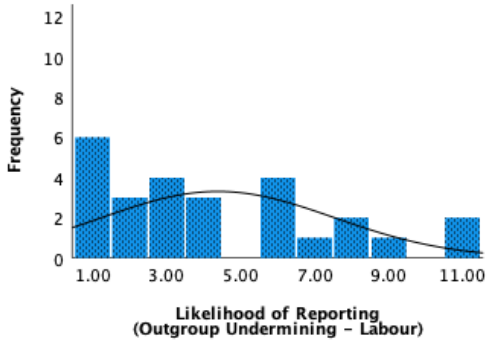
Q5



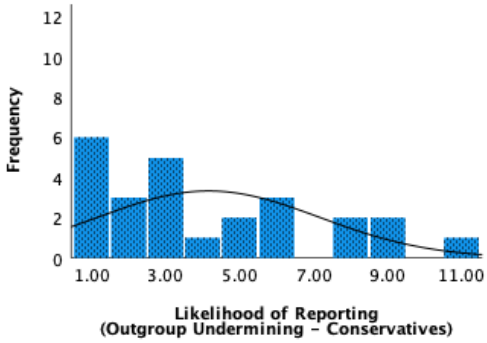
Q6



Q7



Q8



Note. Panel Q1-Q2. Ingroup Supporting Disinformation (Labour v. Conservative). Panel Q3-Q4. Outgroup Supporting Disinformation (Labour v. Conservative). Panel Q5-Q6. Ingroup Undermining Disinformation (Labour v. Conservative). Panel Q7-Q8. Outgroup Undermining Disinformation (Labour v. Conservative).

Appendix R

Pre-registration of Study Three via AsPredicted

'Beliefs and social media spread of disinformation'
(AsPredicted #78270)

Created: 10/28/2021 01:31 AM (PT)
Made Public: 02/03/2023 07:30 AM (PT)

Author(s)
Laura Joyner (University of Westminster) - laura.campbell.joyner@my.westminster.ac.uk
Tom Buchanan (University of Westminster) - t.buchanan@westminster.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
It is predicted that individuals will be more likely to contribute to the spread of disinformation when it supports an issue-related belief. Therefore, Hypotheses 1 and 2 are that:
H1: Individuals who have lower trust in government handling of CV19 will report a greater likelihood of contributing to the spread of disinformation that undermines the government than those with higher trust.
H2: Individuals who have higher trust in government handling of CV19 will report a greater likelihood of contributing to the spread of disinformation that supports the government than those with lower trust.

It is also predicted that individuals will be more likely to judge spreading disinformation that supports an issue-related belief as more morally acceptable. Hypotheses 3 and 4 are therefore that:
H3: Individuals with lower trust in the government will report the sharing of disinformation that undermines government as more morally acceptable than those with higher trust in the government.
H4: Individuals with higher trust in the government will report the sharing of disinformation that supports government as more morally acceptable than those with lower trust in the government.

Finally, it is predicted that moral judgements surrounding the spread of a specific category of disinformation will mediate the relationship between related beliefs and spreading the same category of disinformation. Therefore, Hypotheses 5 and 6 are that:
H5: Moral judgement of sharing 'government undermining' disinformation will mediate the relationship between low trust and increased likelihood of spreading 'undermining' disinformation.
H6: Moral judgement of sharing 'government supporting' disinformation will mediate the relationship between low trust and increased likelihood of spreading 'supporting' disinformation.

3) Describe the key dependent variable(s) specifying how they will be measured.
For H1, H2, H5 and H6 the dependent variable will be the reported likelihood that participants would contribute to the social media spread of two specific categories of disinformation ('favourable' and 'unfavourable' towards the UK government). The present study will be trialling a scale that incorporates methods of contributing to and reducing the spread of disinformation on social media specifically.
The dependent variable for H3 and H4 will be participant's moral judgements of spreading these

two categories of disinformation, measured from 'not at all acceptable' to 'completely morally acceptable'.

4) How many and which conditions will participants be assigned to?
This is a correlational research study. Participants will be asked to rate false or misleading content from two categories of false or misleading content from social media that has previously been pre-tested for allocation purposes. The first contains three items that undermine the UK government in some way, while the second contains three items that support the UK government.

To measure beliefs around trust, participants will also complete the Citizen Trust in Government Organisation scale from Grimmelikhuijen, S. & Knies, E. (2019). Validating a scale for citizen trust in government organisations. https://doi.org/10.1177/0020852315585950

They will also complete the COVID-19 Perceived Risk Scale from Yıldırım, M. & Güler, A. (2020) Factor analysis of the COVID-19 Perceived Risk Scale: A preliminary study, Death Studies, https://doi.org/10.1080/07481187.2020.1784311

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
H1 and H2 will be tested using multiple regressions, with citizen trust, perceived COVID risk, age and gender as predictors of the social media spread of each disinformation 'theme'.
H3 and H4 will also be tested using multiple regressions with citizen trust, perceived COVID risk, age and gender as predictors. However, moral acceptability will be the dependent variable.
H5 and H6 will be tested using mediation analysis, with moral acceptability mediating the relationship between citizen trust and social media spread. Additionally, where age and gender are found to be significant in H1-H4 analyses then they will be included as control variables.
Additional exploratory analysis is also anticipated.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Data will be screened prior to analysis, with the following exclusion criteria applied for removal of responses:
1. Declining consent.
2. Not meeting the recruitment criteria – must be based in England, use social media on a regular basis (e.g. more than once a month) and be over 18 years of age.
3. An implausible completion time, defined by 2SD faster than (below) the mean completion time as would suggest inauthentic responding.
4. Any responses flagged as problematic by Qualtrics' proprietary screening software.

Furthermore, the following criteria will be applied for exclusions during the main analyses:
5. Zero variance between item responses in either the Citizen Trust and COVID-19 Risk Perception scales.
6. If suspicious patterns of responding are detected that may require further removal of participants, then analysis will be reported both with and without said participants.

Any participants who have missing data on the Citizen Trust scale or COVID-19 Risk Perception scale will not be included in analysis where that variable is used.
Where gender is not recorded as either M or F, participants will be excluded only from analyses that specifically involve gender.
For Social Media Spread scale and Moral acceptability responses, if participants have missing data for a specific 'type' of disinformation (e.g. 'Favourable' or 'Unfavourable' towards the government) they will be excluded from that group of analyses only.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
To ensure enough power for the mediation analyses (H5 and H6), sample size planning was conducted using MedPower (Kenny, 2017). A minimum effect size of β = .2 is thought to be the minimum effect size that would be practically significant in social science research (Ferguson, 2009). To detect β = .2 at 80% power, 250 participants would be required. This would also cover the number of participants required to test H1-H4. Again, using Ferguson's (2009) recommended minimum of r2 = .04, 191 participants would be needed to have 80% power. Allowing for data screening exclusions, the target sample size is 280 participants.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532–538. https://doi.org/10.1037/a0015808
Kenny, D. A. (2017, February). MedPower: An interactive tool for the estimation of power in tests of mediation. https://davidakenny.shinyapps.io/MedPower/.

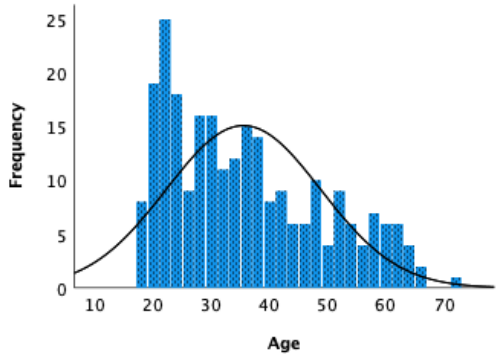
8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
A focus of this study will be to trial a scale to understand how individuals contribute to the digital spread of a social media post. Reliability of this scale will be tested using Cronbach's Alpha.

Additional exploratory analysis will also occur. These may use other demographics that are collected in the study, for example participants' voting intentions.

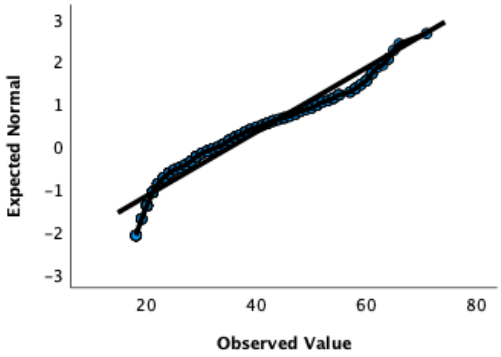
Appendix S

Histograms and Q-Q Plots for Main Variables (Study Three)

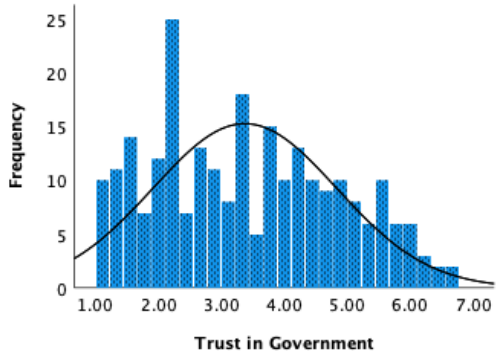
S1



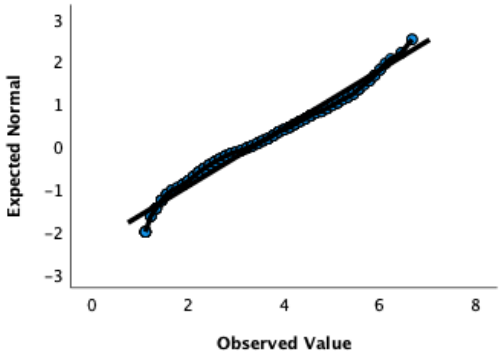
S2



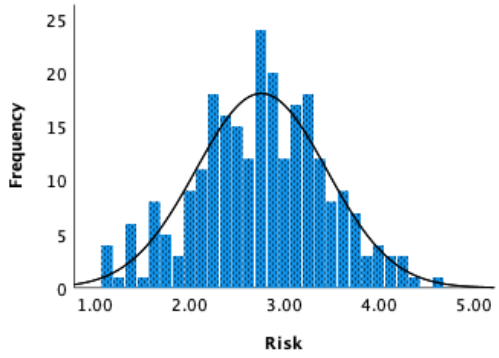
S3



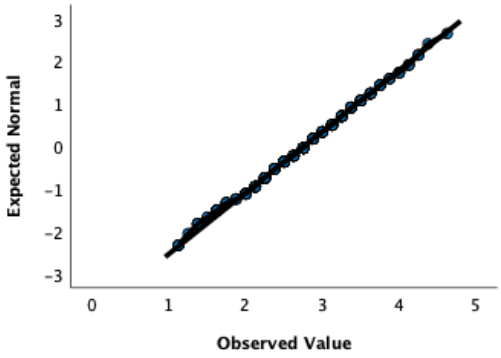
S4



S5



S6

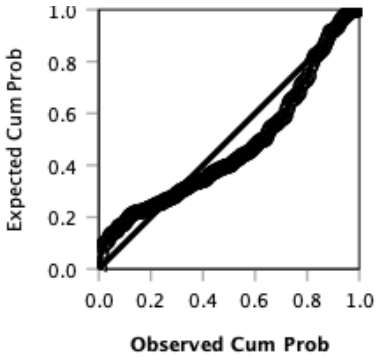


Note. Panel S1. Histogram of Age. Panel S2. Normal Q-Q Plot of Age. Panel S3. Histogram of Trust in UK Government. Panel S4. Normal Q-Q Plot of Trust in UK Government. Panel S5. Histogram of Perceived Risk of COVID-19. Panel S6. Normal Q-Q Plot of Perceived Risk of COVID-19.

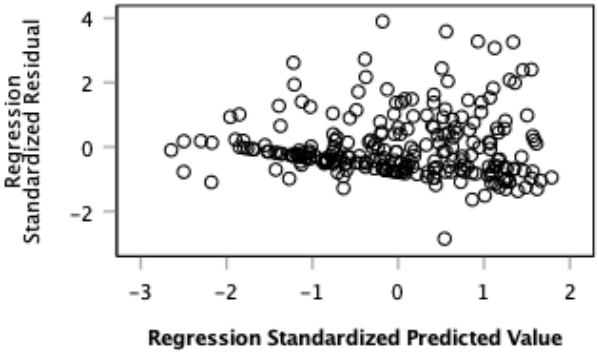
Appendix T

P-P Plots and Scatterplots of Residuals for Planned Regressions

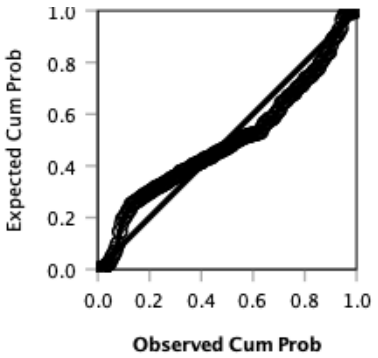
B1



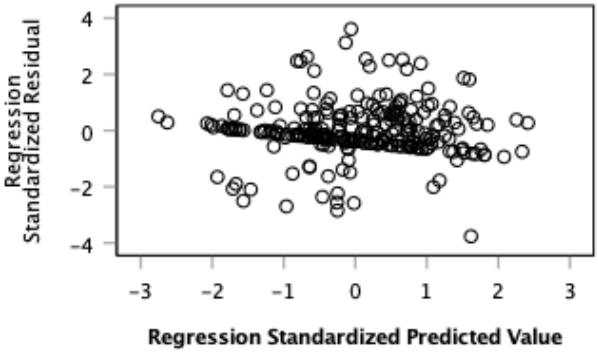
B2



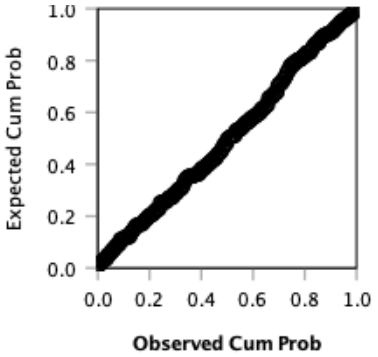
B3



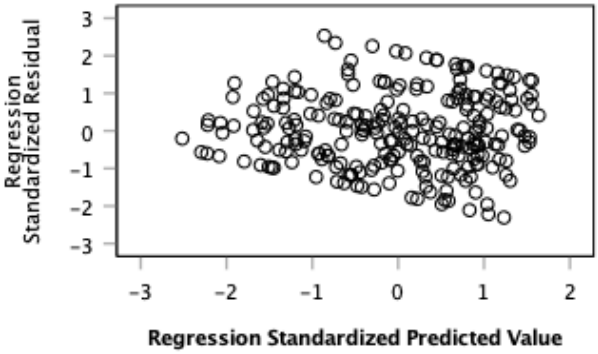
B4



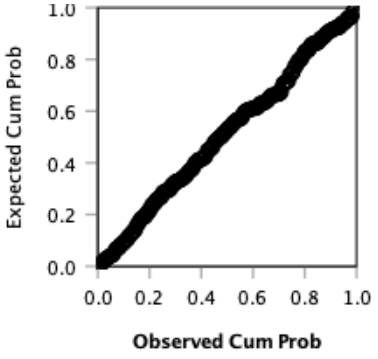
B5



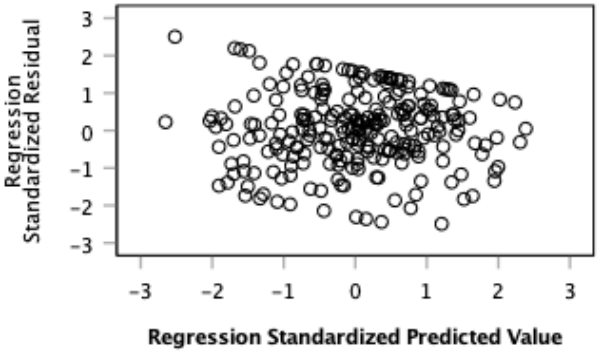
B6



B7



B8



Note. Panels B1-B2. Plots for Spread of Unfavourable misinformation. Panels B3-B4. Plots for Spread of Favourable misinformation. Panels B5-B6. Plots for Moral judgements of Unfavourable misinformation. Panels B7-B8. Plots for Moral judgements of Favourable misinformation.

Appendix U

Ethics Application for Study Four (Pilot & Main)

Figure U1

Ethics Application Decision Letter

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Why do individuals spread disinformation on social media?

Application ID: ETH2122-2442

Date: 09 May 2022

Dear Laura

I am writing to inform you that your application was considered by the Psychology Ethics Committee.

The proposal was approved.

The expiry date for this proposal is 30 Sep 2024.

Yours,

Samuel Evans

Psychology Ethics Committee

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

The desirability of including full details of the consent form in an appendix to your research, and of addressing specifically ethical issues in your methodological discussion.

The requirement to furnish the Research Ethics Committee with details of the conclusion and outcome of the project, and to inform the Research Ethics Committee should the research be discontinued. The Committee would prefer a concise summary of the conclusion and outcome of the project, which would fit no more than one side of A4 paper, please.

Figure U2

Participant Invitation Letter for Study Four Pilot

Social Media Pilot Study
Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to pilot materials for research looking at how we make moral judgements of misinformation on social media.

Who can take part?

We are looking for adults over the age of 18 who live in the United Kingdom, identity as a ***TEAM*** supporter and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide and explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete a short questionnaire to provide some basic details about yourself (for example your age and level of education). You will then be shown four simulated images and asked to rate how favourable you think they are. These images were created for the purpose of the experiment and do not reflect real events or the position of the University.

How long will it take?

The whole study should take around 3 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:

Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner: **laura.campbell.joyner@my.westminster.ac.uk** (Doctoral Researcher)
Professor Tom Buchanan: **T.Buchanan@westminster.ac.uk**
Dr Orkun Yetkili: **O.Yetkili@westminster.ac.uk**

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure U3

Debrief for Study Four Pilot

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to test materials that will be used in future studies looking at how individuals make decisions about misinformation on social media.

The materials you viewed were created for the study and contained false information.

This means they are NOT factual and do not represent the actions of any individuals or teams.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify the information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:

Head of School of Social Sciences: D.Anand@westminster.ac.uk

Please click the arrow below to complete the study and return to Prolific.

Figure U4

Participant Invitation Letter for Study Four (Main)

Why do people interact with social media content
Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to understand why people might interact with content within social media platforms.

Who can take part?

We are looking for adults over the age of 18 who live in the United Kingdom, identity as a ***TEAM*** supporter and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide and explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete some demographic items (for example your age and level of education). You will then be asked to rate your agreement with / relevance of a number of short statements. This will be followed by some questions about interacting with a specific image on social media, including a short writing task. This image was created for the purpose of the experiment and do not reflect real events or the position of the University. Finally, you will be asked to indicate your political orientation.

How long will it take?

The whole study should take around 9 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:

Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner: laura.campbell.joyner@my.westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan: T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili: O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure U5

Debrief for Study Four (Main)

Debrief Sheet

Thank you very much for taking part in this study

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to understand how our identity might influence the decisions we make about misinformation on social media. In particular, we want to see how people's moral values affect their views on how acceptable it is to share material that may not be true.

During the study, some participants will have seen a version of the image with a label stating it was not true. Others were shown the same information but without this label.

The materials you viewed were created for the study and contained false information.

This means they are NOT factual and do not represent the actions of any individuals or teams.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify the information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.

For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: laura.campbell.joyner@my.westminster.ac.uk

Project Supervisor: T.Buchanan@westminster.ac.uk / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:

Head of School of Social Sciences: D.Anand@westminster.ac.uk

Please click the arrow below to complete the study and return to Prolific.

Appendix V

Moral Foundations Questionnaire – Graham et al., 2011

Part 1. When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale:

[0] = not at all relevant (This consideration has nothing to do with my judgments of right and wrong)
[1] = not very relevant
[2] = slightly relevant
[3] = somewhat relevant
[4] = very relevant
[5] = extremely relevant (This is one of the most important factors when I judge right and wrong)

_____1. Whether or not someone suffered emotionally
_____2. Whether or not some people were treated differently than others
_____3. Whether or not someone's action showed love for his or her country
_____4. Whether or not someone showed a lack of respect for authority
_____5. Whether or not someone violated standards of purity and decency
_____6. Whether or not someone was good at math
_____7. Whether or not someone cared for someone weak or vulnerable
_____8. Whether or not someone acted unfairly
_____9. Whether or not someone did something to betray his or her group
_____10. Whether or not someone conformed to the traditions of society
_____11. Whether or not someone did something disgusting
_____12. Whether or not someone was cruel
_____13. Whether or not someone was denied his or her rights
_____14. Whether or not someone showed a lack of loyalty
_____15. Whether or not an action caused chaos or disorder
_____16. Whether or not someone acted in a way that God would approve of

Part 2. Please read the following sentences and indicate your agreement or disagreement:

[0]	[1]	[2]	[3]	[4]	[5]
Strongly disagree	Moderately disagree	Slightly disagree	Slightly agree	Moderately agree	Strongly agree

_____17. Compassion for those who are suffering is the most crucial virtue.
_____18. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.
_____19. I am proud of my country's history.
_____20. Respect for authority is something all children need to learn.
_____21. People should not do things that are disgusting, even if no one is harmed.
_____22. It is better to do good than to do bad.
_____23. One of the worst things a person could do is hurt a defenseless animal.
_____24. Justice is the most important requirement for a society.
_____25. People should be loyal to their family members, even when they have done something wrong.
_____26. Men and women each have different roles to play in society.
_____27. I would call some acts wrong on the grounds that they are unnatural.
_____28. It can never be right to kill a human being.

_____29. I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.
_____30. It is more important to be a team player than to express oneself.
_____31. If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.
_____32. Chastity is an important and valuable virtue.

To score the MFQ yourself, you can copy your answers into the grid below. Then add up the 6 numbers in each of the five columns and write each total in the box at the bottom of the column. The box then shows your score on each of 5 psychological "foundations" of morality. Scores run from 0-30 for each foundation. (Questions 6 and 22 are just used to catch people who are not paying attention. They don't count toward your scores).

Question #	Your Response	Question #	Your Response	Question #	Your Response	Question #	Your Response	Question #	Your Response	Question #	Your Response
1		2		3		4		5		6	
7		8		9		10		11			
12		13		14		15		16			
17		18		19		20		21		22	
23		24		25		26		27			
28		29		30		31		32			
Harm / Care		Fairness / Reciprocit		In-group/ Loyalty		Authority / Respect		Purity / Sanctity			

The average politically moderate American's scores are: 20.2, 20.5, 16.0, 16.5, and 12.6.

Liberals generally score a bit higher than that on Harm/care and Fairness/reciprocity, and much lower than that on the other three foundations. Conservatives generally show the opposite pattern.

The Moral Foundations Questionnaire (MFQ-30, July 2008) by Jesse Graham, Jonathan Haidt, and Brian Nosek.
For more information about Moral Foundations Theory, scoring this form, or interpreting your scores, see: www.MoralFoundations.org. To take this scale online and see how you compare to others, go to www.YourMorals.org

Appendix W

Moral Foundations Questionnaire (Liberty Items) – R. Iyer et al., 2012

Economic/Government Liberty:

*Whether or not private property was respected (**relevance rating**)*

People who are successful in business have a right to enjoy their wealth as they see fit

Society works best when it lets individuals take responsibility for their own lives without telling them what to do.

The government interferes far too much in our everyday lives.

*The government should do more to advance the common good, even if that means limiting the freedom and choices of individuals. (**Reverse scored**)*

Property owners should be allowed to develop their land or build their homes in any way they choose, as long as they don't endanger their neighbors.

Lifestyle Liberty:

*Whether or not everyone was free to do as they wanted. (**Relevance rating**)*

I think everyone should be free to do as they choose, so long as they don't infringe upon the equal freedom of others.

People should be free to decide what group norms or traditions they themselves want to follow.

Appendix X

Planned Tests with Excluded Participants (Study Four)

Table X1

Two-Way ANCOVA Statistics for Likelihood of Spreading Misinformation

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	4.35	1	4.35	1.67	.01
Gender [a]	3.78	1	3.78	1.45	.01
Valence	245.95	1	245.95	94.52***	.27
Tag	72.07	1	72.07	27.70***	.10
Valence * Tag	4.60	1	4.60	1.77	.01
Residuals	663.54	255	2.60		

[a] Gender coded as dummy variable, F = 0, M = 1.

*p < .05. ** p < .01. *** p < .001.

Table X2

Two-Way ANCOVA Statistics for Moral Acceptability of Spreading Misinformation

	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Age	99.84	1	99.84	13.10***	.05
Gender [a]	11.07	1	11.07	1.45	.01
Valence	1280.17	1	1280.17	167.93***	.40
Tag	472.23	1	472.23	61.95***	.20
Valence * Tag	163.33	1	163.33	21.42***	.08
Residuals	1936.36	254	7.62		

[a] Gender coded as dummy variable, F = 0, M = 1.

*p < .05. ** p < .01. *** p < .001.

Table X3

Ordinary Least Squares Regression Coefficients (with Standard Errors) From a First Stage Moderated Mediation Model Predicting Likelihood of Contributing to Spread

		Outcome		
		M: Moral Judgement		*Y*: Level of Spread Contribution
Constant		5.89*** (0.64)		4.37***(0.38)
X: Valence	$a_1 \rightarrow$	6.01*** (0.44)	$c' \rightarrow$	0.55* (0.28)
W: Tag	$a_2 \rightarrow$	−1.11* (0.52)	$b_2 \rightarrow$	−0.48 (0.25)
XW: Valence x Tag	$a_3 \rightarrow$	−3.17*** (0.68)	$b_3 \rightarrow$	0.35 (0.37)
Age		−0.04*** (0.01)		0.0001 (0.01)
M: Moral Judgement			$b_1 \rightarrow$	0.28*** (0.03)
	R	0.72		0.70
	R^2	0.51		0.49
			Index	95% bootstrap CI[a]
Moderated mediation			−0.89	−1.32, −0.49

Note. Valence (0 = negative, 1 = positive) and Tag (0 = no tag, 1 = fact-check tag) coded as dummy variables

[a] Percentile bootstrap CI based on 5,000 bootstrap samples.

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

Appendix Y

Pre-registration of Study Four via AsPredicted

'Identity-based moral judgements & interactions with social media disinformation'

(AsPredicted #96907)

Created: 05/12/2022 04:09 AM (PT)

Author(s)
Laura Joyner (University of Westminster) - laura.campbell.joyner@my.westminster.ac.uk
Tom Buchanan (University of Westminster) - t.buchanan@westminster.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
It is predicted that people are less likely to contribute to the wider spread of disinformation on social media when the content undermines the ingroup. Therefore, Hypothesis 1 is that:
H1: Individuals will be more likely to spread disinformation that is positive about their ingroup than disinformation that is negative about their ingroup.

When content contains additional information stating it is false, it is predicted people will be less likely to spread it further. People may also judge the content as being less acceptable to spread when this 'fact check' tag is attached. It is therefore predicted that:
H2: Individuals will be less likely to spread content that displays a 'fact check' tag compared to content with no tag
H3: Individuals will judge it to be less morally acceptable to spread content that displays a 'fact check' tag compared to content with no tag

It is also predicted that the relationship between the valence of the content (e.g. positive or negative about the ingroup) and the likelihood of spread is partially explained by moral evaluations. Both the relationships between valence and moral evaluation, and valence and spread will also be weakened by the inclusion of a fact check tag:
H4: The relationship between content valence and spread will be mediated by moral acceptability and moderated by the inclusion of a fact check tag

3) Describe the key dependent variable(s) specifying how they will be measured.
The dependent variable for H1, H2 and H4 will be the likelihood of contributing to the spread of disinformation content on social media. This is measured by a social media spread scale which incorporates actions contributing to or reducing the onward spread of content on social media. The dependent variable for H3 will be participant's moral judgements of spreading the content, measured from 'not at all acceptable' to 'completely morally acceptable'.

4) How many and which conditions will participants be assigned to?
This is a 2x2 between-groups design. Participants will be randomly allocated to one of four conditions where they will be presented with a single item of disinformation throughout (one for each condition). These items differ by valence of the content (e.g. 'positive' or 'negative' for the

participants ingroup) and whether it is tagged with fact check information (e.g. 'tag' or 'no-tag')

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
H1-H3 will be tested using two 2x2 ANCOVAs with the independent variables 'valence' and 'fact check', while controlling for age and gender. The first ANCOVA will have a DV of 'spread' (H1 & H2) and the second ANCOVA will have a DV of 'moral acceptability' (H3).
H2 will be tested using mediation analysis, with moral acceptability mediating the relationship between stance and social media spread.
H4 will be tested using a moderated mediation analysis. 'Moral acceptability' will be included as the mediator (M) between 'valence' (X) and 'spread' (Y). 'Fact check' will be included as the moderator variable (W).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Data will be screened prior to analysis, with the following exclusion criteria applied for removal of responses:
1. Declining consent.
2. Not meeting the recruitment criteria – must be based in UK, use social media on a regular basis (e.g. more than once a month), supports one of the premier league teams that the stimuli was created for (e.g. Arsenal, Chelsea, Liverpool, Manchester United and Tottenham) and be over 18.
3. An implausible completion time, defined by 2SD faster than (below) the mean completion time as would suggest inauthentic responding.
4. Any responses flagged as problematic by Qualtrics' proprietary screening software.

Furthermore, the following criteria will be applied for exclusions during the main analyses:
5. Zero variance between item responses in the Moral Foundations Questionnaire
6. Fail the 'catch' items on the Moral Foundations Questionnaire
7. If suspicious patterns of responding are detected that may require further removal of participants, then analysis will be reported both with and without said participants.

Any participants who have missing data on the Moral Foundations Questionnaire, Strength of Identity Measure and Political Alignment questions will not be included in analysis where that variable is used.
Where gender is not recorded as either M or F, participants will be excluded only from analyses that specifically involve gender.
For text-analysis, participants who do not fill in the textbox will be excluded from analysis that relate to these scores

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.
To ensure enough power for the moderated mediation analyses (H4), sample size planning was first conducted using MedPower (Kenny, 2017). A minimum effect size of β = .2 is thought to be the minimum effect size that would be practically significant in social science research (Ferguson, 2009). To detect β = .2 at 80% power, 250 participants would be required. However, this tool is designed for mediation analysis planning specifically. Therefore, to accommodate the inclusion of moderator variables, this proposed sample size was confirmed using G*Power by planning for a linear multiple regression. To reach a minimum of r2 = .04, 191 participants would be needed for 80% power, suggesting a sample of 250 would be acceptable for a moderated mediation analysis. For H1-H3, to detect ηp2 = .04, 191 participants would also be needed to have 80% power. Allowing for data screening exclusions, the target sample size is 280 participants.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532–538. https://doi.org/10.1037/a0015808

Kenny, D. A. (2017, February). MedPower: An interactive tool for the estimation of power in tests of mediation. https://davidakenny.shinyapps.io/MedPower/.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Additional exploratory analysis will also occur. These may use other demographics that are collected in the study, for example participants' political orientation and strength of identity, as well as scores from the Moral Foundation Question and text analysis.

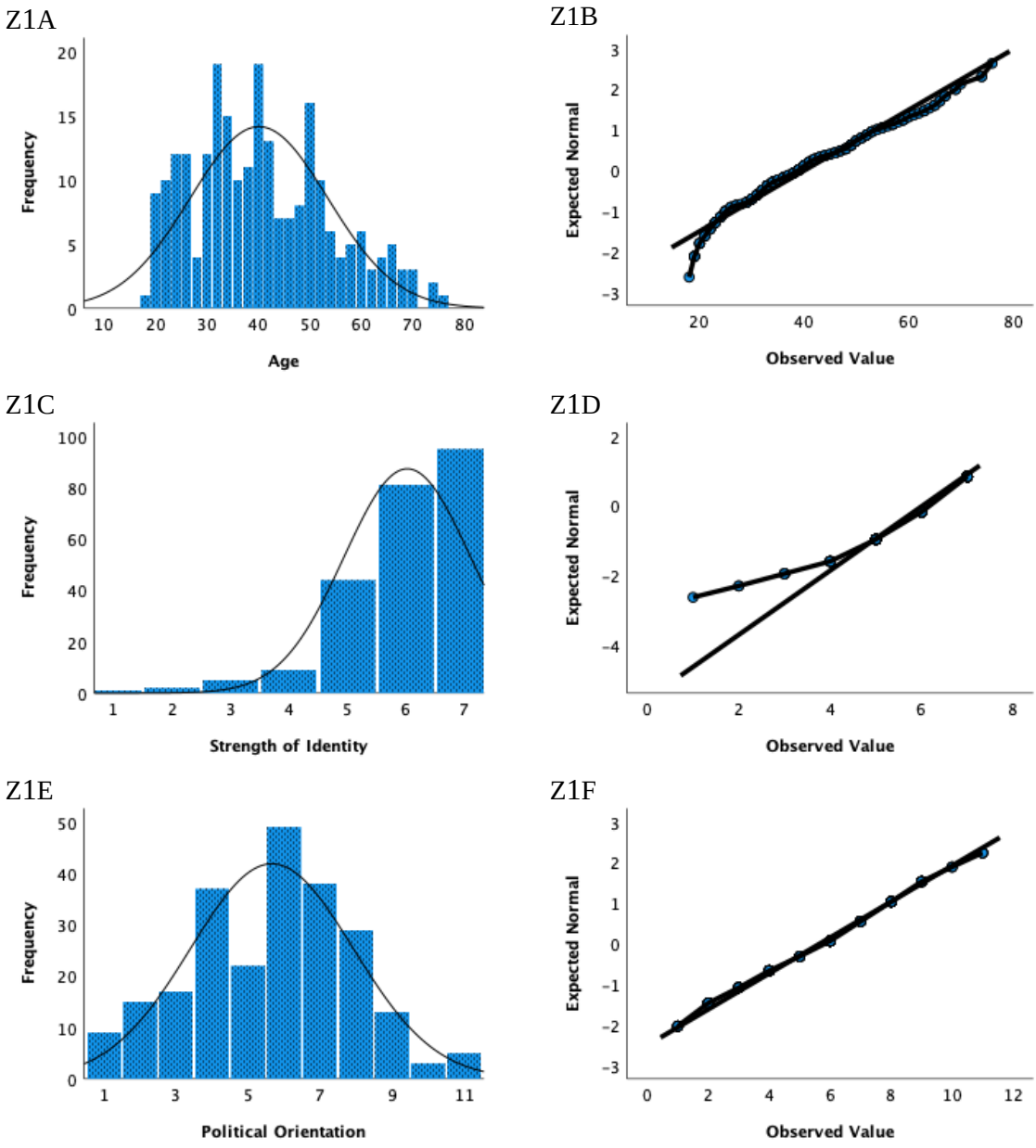
Data collection will be paused after the first 20 responses to check data quality and whether participants are responding to the free-text question in the expected way. No analysis will be carried out at this point, but question wording may be amended if it appears participants are having trouble responding.

Appendix Z

Histograms and Q-Q Plots for Main Variables (Study Four)

Figure Z1

Histograms and Q-Q Plots for Participant Variables

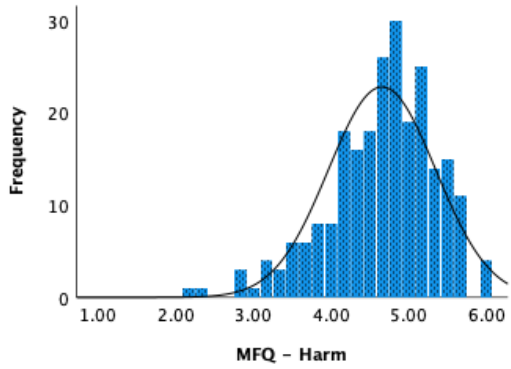


Note. Panels Z1A-B. Normality Plots for Age. Panels Z1C-D. Normality Plots for Strength of Identity. Panels Z1E-F. Normality Plots for Political Orientation.

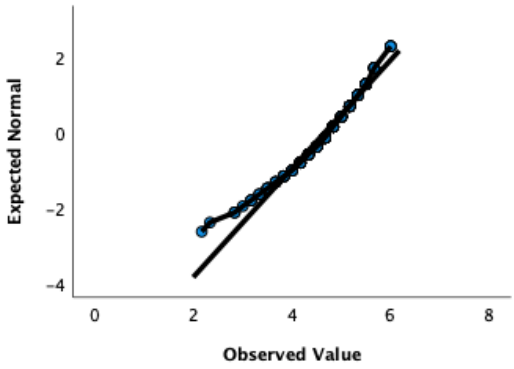
Figure Z2

Histograms and Q-Q Plots for Moral Foundations Questionnaire

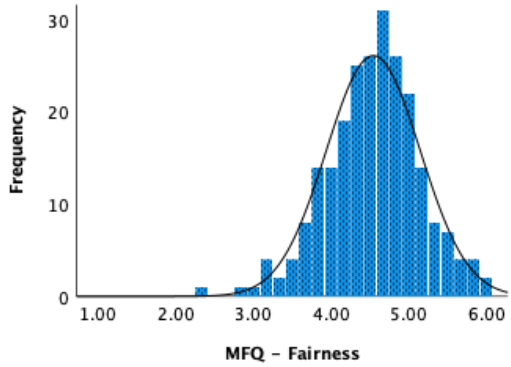
Z2A



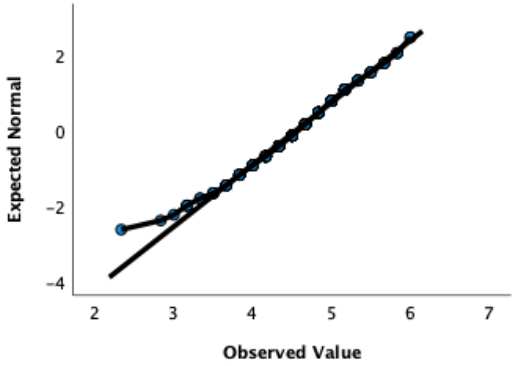
Z2B



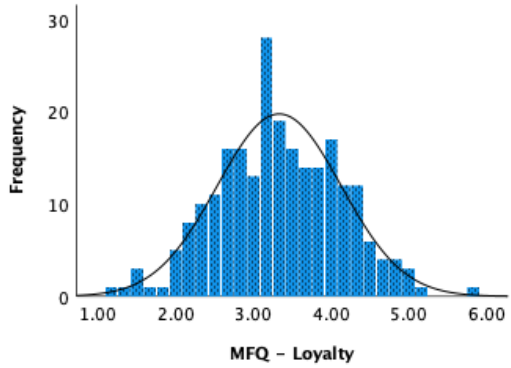
Z2C



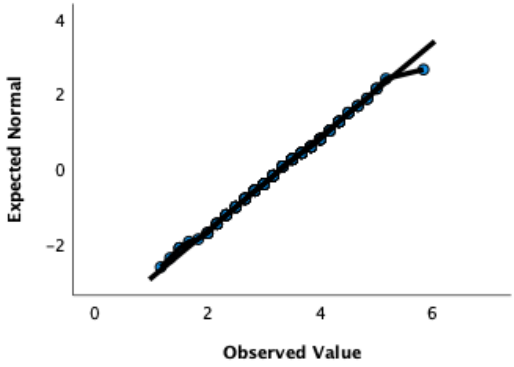
Z2D



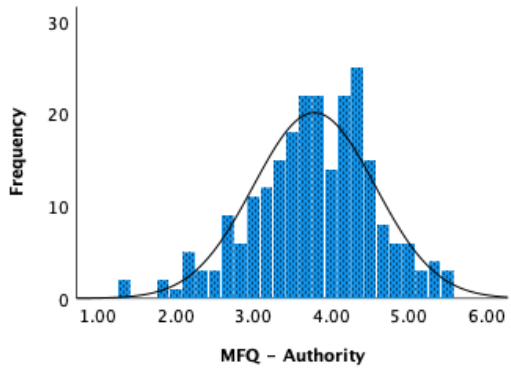
Z2E



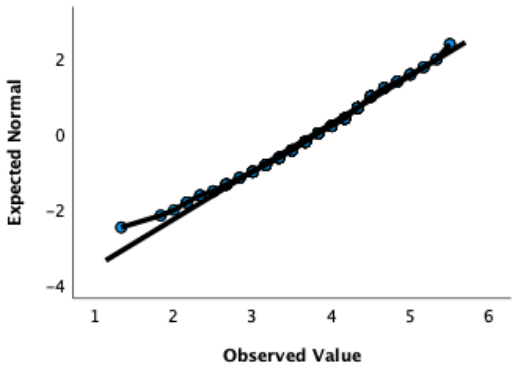
Z2F



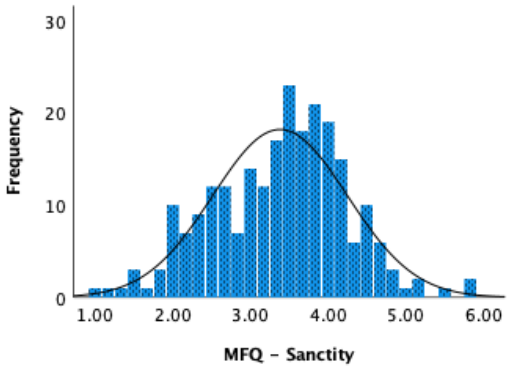
Z2G



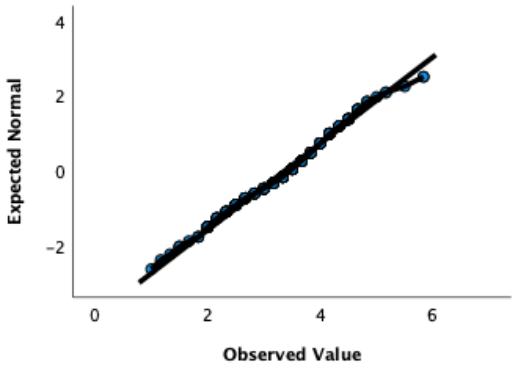
Z2H



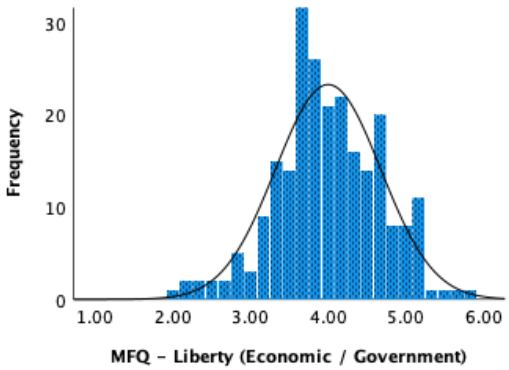
Z2I



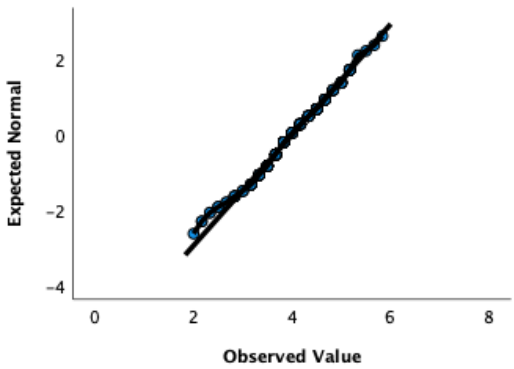
Z2J



Z2K



Z2L

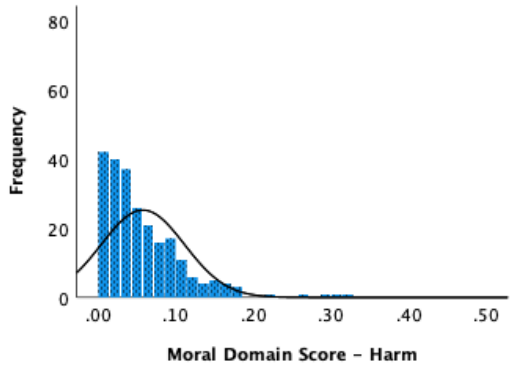


Note. Panels Z2A-B. Normality Plots for MFQ Harm Score. Panels Z2C-D. Normality Plots for MFQ Fairness Score. Panels Z2E-F. Normality Plots for MFQ Loyalty Score. Panels Z2G-H. Normality Plots for MFQ Authority Score. Panels Z2I-J. Normality Plots for MFQ Sanctity Score. Panels Z2K-L. Normality Plots for MFQ Liberty (Economic / Government) Score.

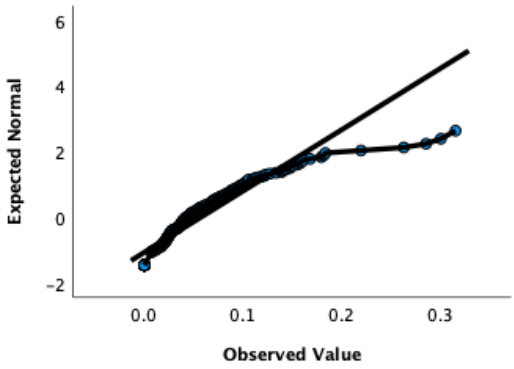
Figure Z3

Histograms and Q-Q Plots for Moral Domain Scores

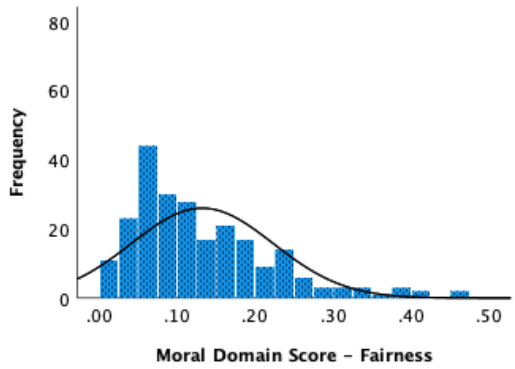
Z3A



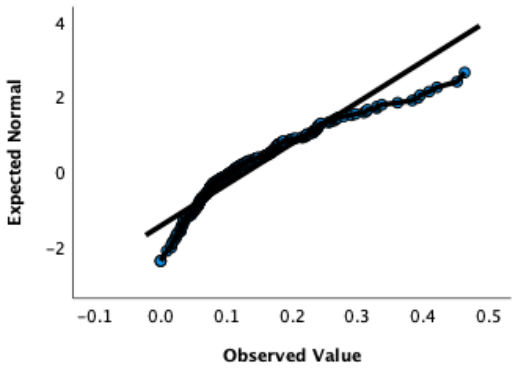
Z3B



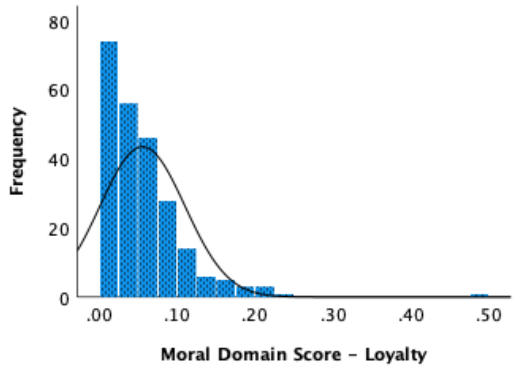
Z3C



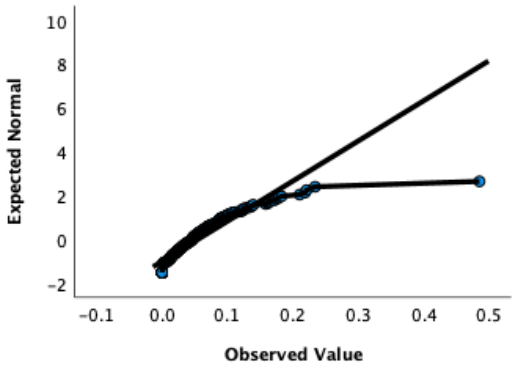
Z3D



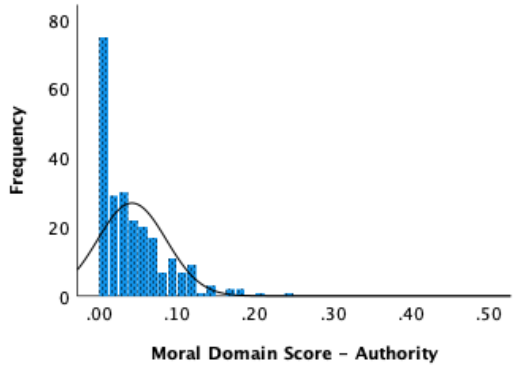
Z3E



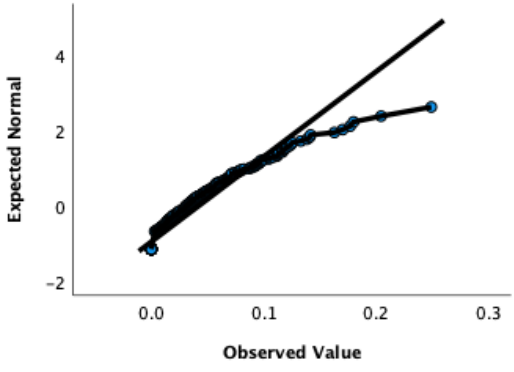
Z3F



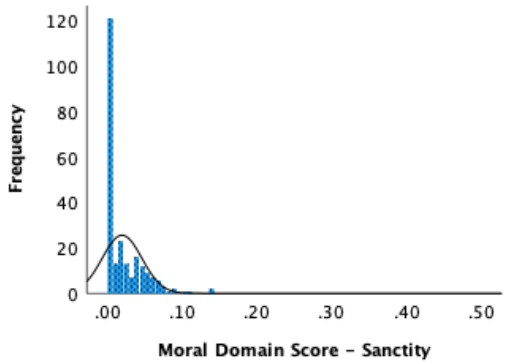
Z3G



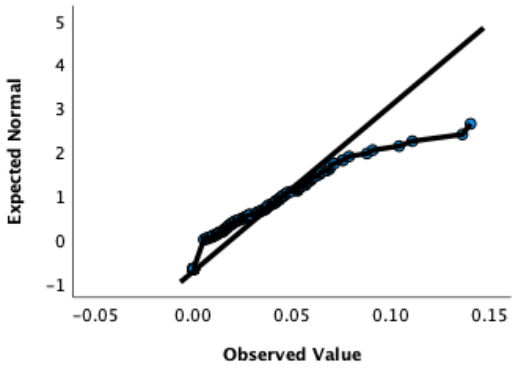
Z3H



Z3I



Z3J

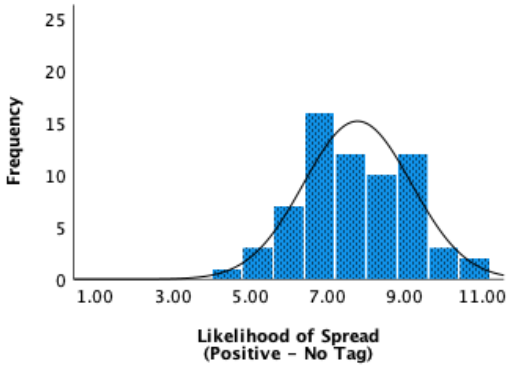


Note. Panels Z3A-B. Normality Plots for Harm Moral Domain Score. Panels Z3C-D. Normality Plots for Fairness Moral Domain Score. Panels Z3E-F. Normality Plots for Loyalty Moral Domain Score. Panels Z3G-H. Normality Plots for Authority Moral Domain Score. Panels Z3I-J. Normality Plots for Sanctity Moral Domain Score.

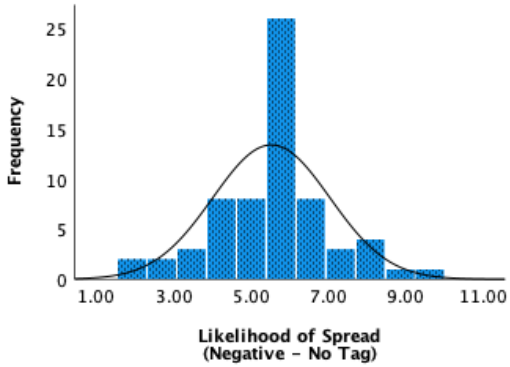
Appendix AA

Histograms and Box Plots for Likelihood of Spread by Condition

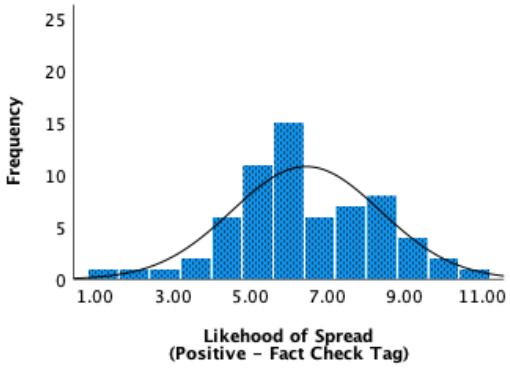
AA1



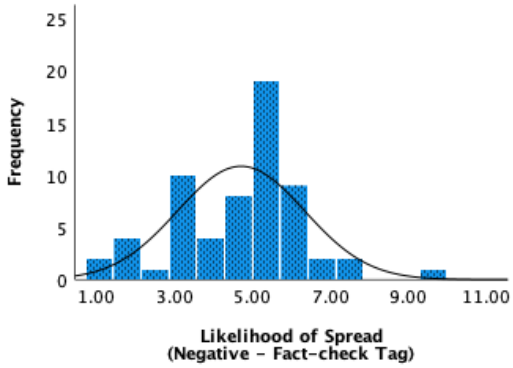
AA2



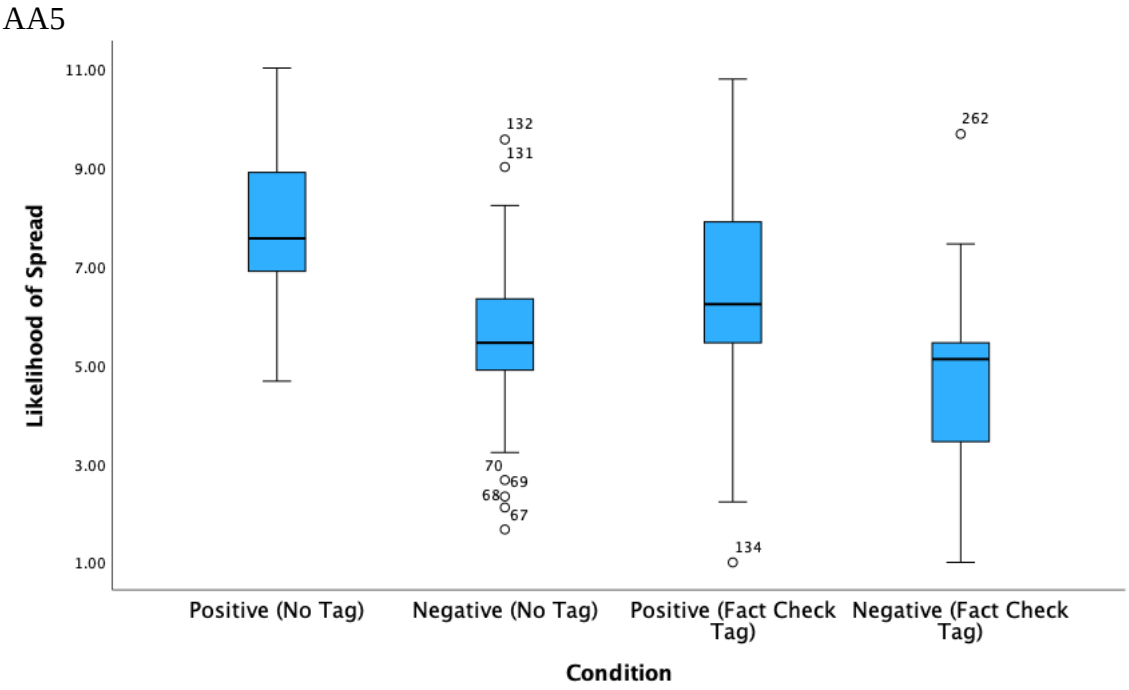
AA3



AA4



AA5



Note. Panel AA1. Histogram for Positive Misinformation about Ingroup. Panel AA2. Histogram for Negative Misinformation about Ingroup. Panel AA3. Histogram for Positive Disinformation (with Fact-Check) about Ingroup. Panel AA4. Histogram for Negative Disinformation (with Fact-Check about Ingroup. Panel AA5. Boxplots of Likelihood of Spread Across Each Condition.

Appendix BB

Pairwise Comparisons for Likelihood of Spread

Table BB1

Pairwise Comparisons (Fact-Check vs No Fact-Check) of Likelihood of Spread Scores

				95% Confidence Interval for Difference	
Valence of Post	Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Positive	1.31	.28	<.001	.76	1.87
Negative	0.81	.29	.01	.25	1.37

Table BB2

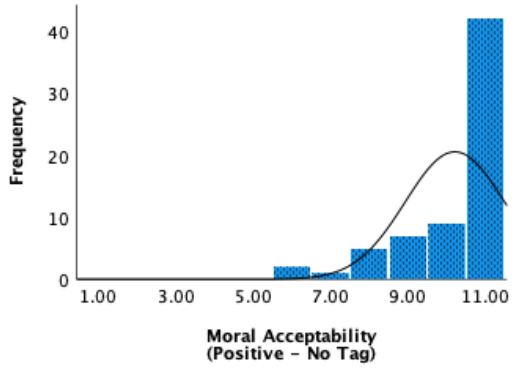
Pairwise Comparisons (Positive vs Negative) of Likelihood of Spread Scores

				95% Confidence Interval for Difference	
Fact-check Tag	Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
None	2.21	.28	<.001	1.66	2.77
Includes Tag	1.71	.29	<.001	1.14	2.27

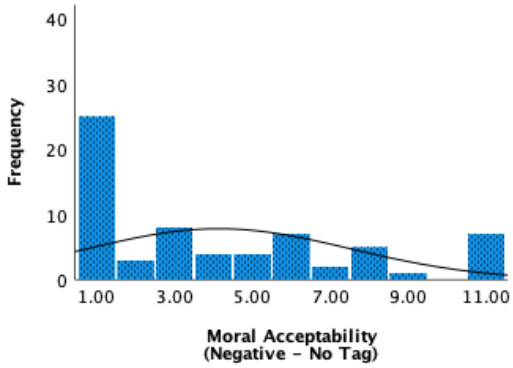
Appendix CC

Histograms and Box Plots for Moral Judgement by Condition

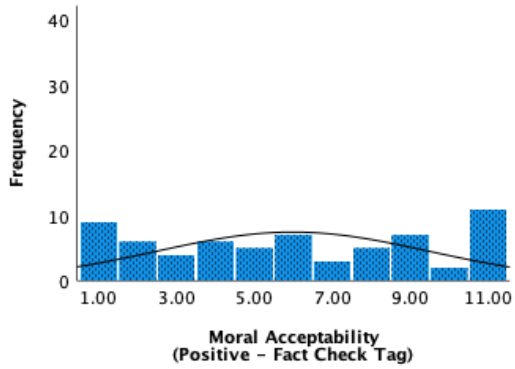
CC1



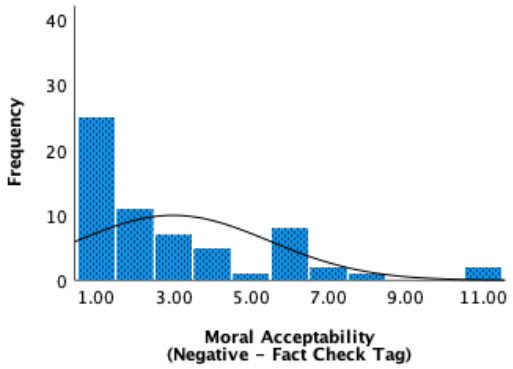
CC2



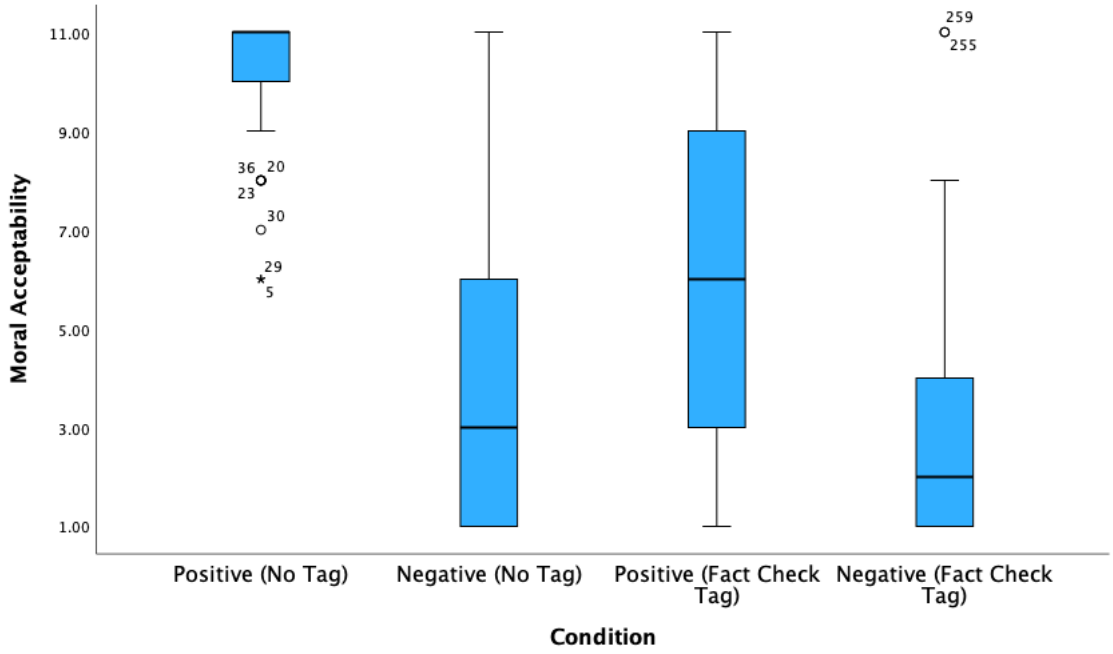
CC3



CC4



CC5



Note. Panel CC1. Histogram for Positive Misinformation about Ingroup. Panel CC2. Histogram for Negative Misinformation about Ingroup. Panel CC3. Histogram for Positive Disinformation (with Fact-Check) about Ingroup. Panel CC4. Histogram for Negative Disinformation (with Fact-Check about Ingroup. Panel CC5. Boxplots of Moral Judgements Across Each Condition.

Appendix DD

Two-Way ANCOVA Statistics for Moral Acceptability of Spreading Misinformation

Source	Sum of Squares	*df*	Mean Square	*F*	η^2_p
Gender	1.37	1	1.37	.17	.001
Valence	1299.66	1	1299.66	162.76***	.39
Tag	473.62	1	473.62	59.31***	.19
Valence * Tag	165.09	1	165.09	20.68***	.08
Residuals	2036.19	255	7.99		

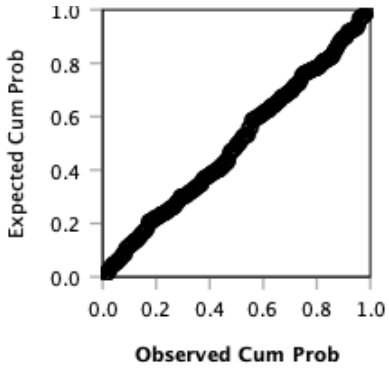
[a] Gender coded as dummy variable, F = 0, M = 1.

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

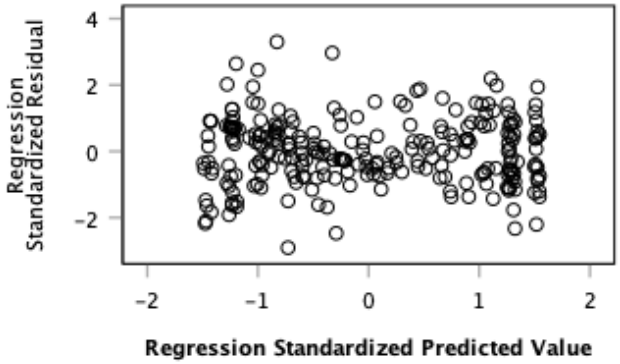
Appendix EE

P-P Plots and Scatterplots of Residuals for Planned Conditional Process Analysis

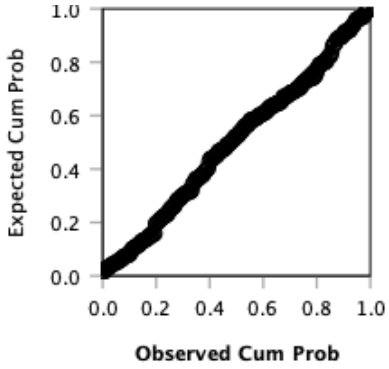
EE1



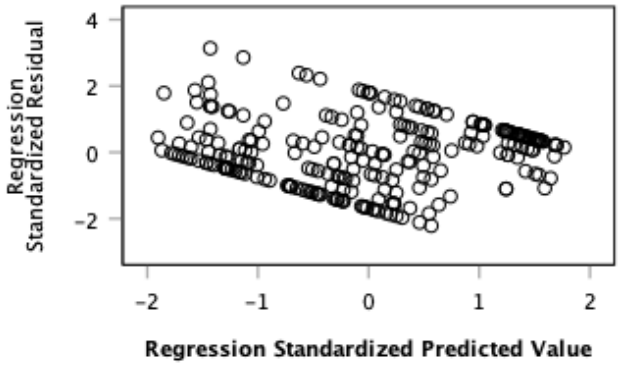
EE2



EE3



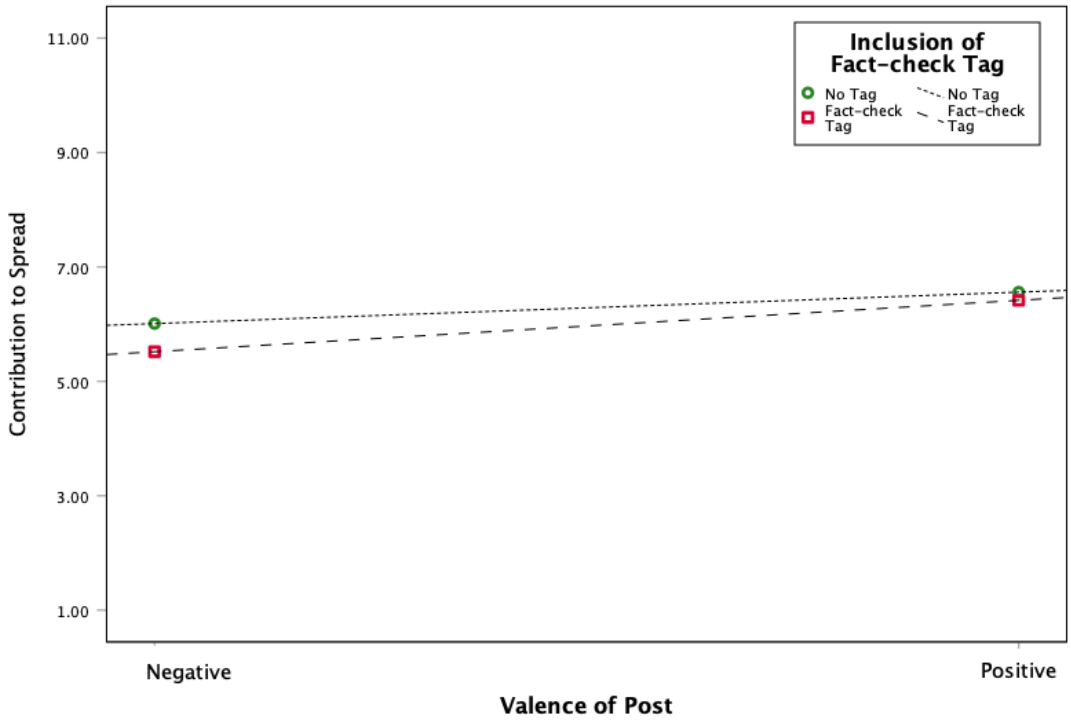
EE4



Note. Panels EE1-EE2. Plots for Likelihood of Spread. Panels EE3-EE4. Plots for Moral Judgements.

Appendix FF

Conditional Direct Effect of Valence & Fact-Check Tags on Spread Contributions



Note. Controlled for age and indirect effects (e.g. moral judgement)

Appendix GG

Holms Bonferroni Corrections for Study Four

Ranking	Applied correction
1	$p < .0125$
2	$p < .0167$
3	$p < .025$
4	$p < .05$

Note. Significance value ranked from smallest to largest.

Appendix HH

Ethics Application for Study Five (Pilot & Main)

Figure HH1

Ethics Application Significant Amendments Letter

UNIVERSITY OF
FORWARD
THINKING
WESTMINSTER

Project title: Why do individuals spread disinformation on social media?

Application ID: ETH2223-0568

Date: 25 Oct 2022

Dear Laura

I am writing to inform you that your significant amendments to protocol were considered by the Psychology Research and Knowledge Exchange Ethics Working Group .

The proposal was approved.

Yours,

Anna Cheshire

Psychology Research and Knowledge Exchange Ethics Working Group

I am advised by the Committee to remind you of the following points:

Your responsibility to notify the Research Ethics Committee immediately of any information received by you, or of which you become aware, which would cast doubt upon, or alter, any information contained in the original application, or a later amendment, submitted to the Research Ethics Committee and/or which would raise questions about the safety and/or continued conduct of the research.

The need to comply with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) 2018.

The need to comply, throughout the conduct of the study, with good research practice standards.

The need to refer proposed amendments to the protocol to the Research Ethics Committee for further review and to obtain Research Ethics Committee approval thereto prior to implementation (except only in cases of emergency when the welfare of the subject is paramount).

The desirability of including full details of the consent form in an appendix to your research, and of addressing specifically ethical issues in your methodological discussion.

The requirement to furnish the Research Ethics Committee with details of the conclusion and outcome of the project, and to inform the Research Ethics Committee should the research be discontinued. The Committee would prefer a concise summary of the conclusion and outcome of the project, which would fit no more than one side of A4 paper, please.

Figure HH2

Participant Invitation Letter for Study Five Pilot

Social Media Pilot Study

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to pilot materials for research looking at whether the way individuals make moral judgements could influence the spread of false information on social media.

Who can take part?

We are looking for adults over the age of 18 who live in the USA, identity as either Democrat or Republican voters and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide and explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete a short questionnaire to provide some basic details about yourself (for example your age and level of education). You will then be shown two simulated images and asked to rate the extent to which they represent certain moral values. These are created for the purpose of the experiment and do not reflect the position of the University.

How long will it take?

The whole study should take around 3 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:
Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?
This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics. If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm. All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018. You will not be personally identifiable in any reports that arise from this study. Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor. Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?
This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?
This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:
Laura Joyner: L.Joyner1@westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan: T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili: O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure HH3

Debrief for Study Five Pilot

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to test materials that will be used in future studies looking at whether appeals to different moral values could influence the spread of false information on social media.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist : The SHARE Checklist has been created by the UK Government to help the public identify misleading information online. For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: L.Joyner1@westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:
Head of School of Social Sciences: D.Anand@westminster.ac.uk

Please click the arrow below to complete the study and be returned to Prolific

Figure HH4

Participant Invitation Letter for Study Five (Main)

Moral Judgements and Social Media Interactions

Study Invitation

You are being invited to take part in a research project. Before you decide whether to take part, it is important for you to understand why the research is being conducted and what participation will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part

What is the purpose of the project?

This research is part of a PhD study being conducted at the University of Westminster by Laura Joyner. The purpose of this study is to understand how moral judgements influence people's interactions with content on social media platforms

Who can take part?

We are looking for adults over the age of 18 who live in the USA, identity as either Democrat or Republican voters and currently have an active social media account.

Do I have to take part?

No, your participation is entirely voluntary. You can stop taking part at any time without having to provide and explanation.

You can choose to decline answering any question or undertaking any task that is asked of you, even after the study starts.

Please note that once you have completed the study you will be unable to withdraw, because the data is being collected anonymously and there is no way to identify individuals' responses

What will happen to me if I take part?

This is an online study. When you have read the information sheet and given consent to take part you will be asked to complete some demographic items (for example your age and level of education). You will be asked about whether you would interact with a simulated image on social media and to make a moral judgement about the image. The image was created for the purpose of the experiment and do not reflect the position of any political party or the University.

How long will it take?

The whole study should take around 3 minutes

What are the possible disadvantages and risks of taking part?

There is no anticipated disadvantages or risks to your participation.

What are the possible benefits of taking part?

Your contribution will help with developing our understanding of why individuals interact with misinformation on social media

What if something goes wrong?

This research has been approved by the Psychology Research Ethics Working Group at the University of Westminster.

If you would like to make a complaint about this research, please contact:

Professor Dibyesh Anand (Head of School of Social Sciences) - D.Anand@westminster.ac.uk

What will happen to my data?

This research is being conducted in accordance with the University of Westminster Code of Ethical Conduct and the British Psychological Society (BPS) Code of ethics.

If you provide any personally identifiable data it will be treated confidentially and in accordance with the University of Westminster ethical guidelines and British Psychological Society code of human research ethics. Note in exceptional circumstances, the duty of confidentiality may be over ridden by more compelling duties such as to protect the individual from harm.

All data will be securely stored and managed in accordance with the Data Protection Regulation 2018 and the General Data Protection Act 2018.

You will not be personally identifiable in any reports that arise from this study.

Your data may be shared with other members of the research team including the supervisor of the research or those working closely with the supervisor.

Your anonymised data may be used for future research and may undergo secondary analysis. This future research may be unrelated to the goals of this study may be conducted by researchers unrelated to this research project.

What will happen to the results of the research project?

This research will be written up and submitted for assessment as part of Laura Joyner's PhD submission at the University of Westminster

Who is organising and/or funding this project?

This project is not funded by a research agency

Here are the names and contact details of the researcher conducting this study and their supervisors:

Laura Joyner: l.joyner@westminster.ac.uk (Doctoral Researcher)
Professor Tom Buchanan: T.Buchanan@westminster.ac.uk
Dr Orkun Yetkili: O.Yetkili@westminster.ac.uk

Thank you for considering taking part, please ask the experimenter if you have any questions.

Figure HH5

Debrief for Study Five (Main)

Debrief Sheet

Thank you very much for taking part in this study.

What was the study about?

Now that the experiment is over, we can tell you more about it. The aim of the study was to understand whether appeals to different moral values could influence the spread of false information on social media.

The materials you viewed were created for the study and contained false information.

This means they are NOT factual and do not represent the actual performance of any political party or police force.

What can I do to find out more information or if I would like further support?

We hope that this study has not raised any uncomfortable feelings. However, if you would like to know more about how to verify information you see online the following resources may be able to offer some help. The list is not exhaustive, but designed to provide helpful avenues should you feel you need them:

SHARE Checklist :

The SHARE Checklist has been created by the UK Government to help the public identify misleading information online.
For more information visit: https://sharechecklist.gov.uk

If you have any questions about the research and wish to discuss them with the researchers please use the following contact email:

Doctoral Researcher: l.joyner@westminster.ac.uk
Project Supervisor: T.Buchanan@westminster.ac.uk / / O.Yetkili@westminster.ac.uk

If you have questions or concerns that cannot be answered by the researchers please contact:

Head of School of Social Sciences: D.Anand@westminster.ac.uk

Appendix II

Pre-registration of Study Five via AsPredicted

'Moral reframing intervention to reduce user-spread of misinformation'
(AsPredicted #110905)

Created: 10/27/2022 06:27 AM (PT)

Author(s)
Laura Joyner (University of Westminster) - laura.campbell.joyner@my.westminster.ac.uk
Tom Buchanan (University of Westminster) - t.buchanan@westminster.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Viewing a moral appeal (binding or individualising) may increase carefulness about spreading potential misinformation:
H1a: Participants exposed to either binding or individualising moral appeals will judge misinformation as less morally acceptable to spread than participants who are not.
H1b: Participants exposed to either binding or individualising moral appeal condition will be less likely to contribute to the onward spread of misinformation than participants who are not.

Moral appeals may be more effective when consistent with their moral values. It is predicted people's evaluations of misinformation will be more negative when the appeal is consistent with (vs. opposes) moral values associated with their political orientation:
H2a: The effect of a moral appeal on moral judgements of misinformation will be stronger when the appeal is consistent with participants' moral values.
H2b: The effect of a moral appeal on intentions to spread misinformation will be stronger when the appeal is consistent with participants' moral values.

Prior research has found that re-framing a moral appeal so it relates to binding values can make them more effective for political conservatives:
H3a: Conservative participants who read a binding moral appeal will judge misinformation as less morally acceptable to spread than other conservatives.
H3b: Conservative participants who read a binding moral appeal will be less likely to spread misinformation than other conservatives.

Associating 'accuracy' with a potential moral violation could help improve the efficacy of accuracy interventions:
H4a: The effect of an accuracy intervention on lowering moral judgements of misinformation will be stronger for participants who read a moral appeal.
H4b: The effect of an accuracy intervention on lowering intentions to spread misinformation will be stronger for participants who read a moral appeal.
H5a: The effect of an accuracy intervention on lowering moral judgements of misinformation will be strongest for participants who read a value-consistent moral appeal.
H5b: The effect of an accuracy intervention on lowering intentions to spread misinformation will be strongest for participants who read a value-consistent moral appeal.

Several studies have suggested prompting social media users to consider accuracy may help

reduce intentions to spread misinformation:
H6a: Participants presented with an accuracy intervention will judge misinformation as less morally acceptable to spread than participants who are not.
H6b: Participants presented with an accuracy intervention will be less likely to spread misinformation than participants who are not.

3) Describe the key dependent variable(s) specifying how they will be measured.
The dependent variable for H1a-H4a will be participant's moral judgements of spreading the content, measured from 'not at all acceptable' to 'completely morally acceptable'.

The dependent variable for H1b-H4b will be the likelihood of contributing to the spread of disinformation content on social media. This is measured by a social media spread scale which incorporates actions contributing to or reducing the onward spread of content on social media.

4) How many and which conditions will participants be assigned to?
This is a 3x2x2 between-groups design (Moral Frame x Partisanship x Accuracy Prompt). Participants will be randomly allocated to one of six conditions. First, participants will either be shown a moral appeal that is framed to appeal to one type of moral values (e.g. 'binding' or 'individualising') or will not be shown an appeal ('no framing').

All participants will be shown a piece of misinformation. The sentiment within the misinformation will be the same, but the party which is favourably presented will be participants' own party (e.g. Democrat or Republican). Additionally, the misinformation may or may not be accompanied by an accuracy intervention (e.g. 'accuracy tag' or 'no tag').

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
H1, H2, H4-H6 will be tested using two 3x2x2 ANCOVAs with the independent variables 'moral frame', 'partisanship' and 'accuracy prompt'. The first ANCOVA (a) will have a DV of 'moral acceptability' and the second DV (b) will have a DV of 'spread'.

H3 will be tested using two 2x2 ANCOVAs based on Republican voter data only. The independent variables will be 'moral frame' and 'accuracy prompt'. Again, the first ANCOVA (a) will have a DV of 'moral acceptability' and the second DV (b) will have a DV of 'spread'.

Age and gender (dummy coded for male and female) will be included as covariates in all analyses.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Data will be screened prior to analysis, with the following exclusion criteria applied for removal of responses:
1. Declining consent.
2. Not meeting the recruitment criteria – must be based in US, use social media on a regular basis (e.g. more than once a month), support either the Democrat or Republican party and be over 18.
3. An implausible completion time, defined by 2SD faster than (below) the mean completion time as would suggest inauthentic responding.
4. Any responses flagged as problematic by Qualtrics' proprietary screening software.

Furthermore, the following criteria will be applied for exclusions during the main analyses:
5. If suspicious patterns of responding are detected that may require further removal of participants, then analysis will be reported both with and without said participants.

Any participants who have missing data on the Strength of Identity Measure, partisanship and political alignment questions will not be included in analysis where that variable is used. Where gender is not recorded as either M or F, participants will be excluded only from analyses

that specifically involve gender.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
To ensure enough power for a three-way ANCOVA, a power analysis has been conducted using G*Power. Prior studies using the accuracy intervention have reported small effect sizes. Therefore, to detect eta^2 = .02 at 80% power, a minimum of 476 participants would be required. Allowing for data screening exclusions, the target sample size is 520 participants.
This would also allow for enough power to test H3, where a minimum of 235 participants who vote Republican would be needed eta^2 = .04 at 80% power. An effect size of eta^2 = .04 is thought to be the minimum effect size that would be practically significant in social science research (Ferguson, 2009).

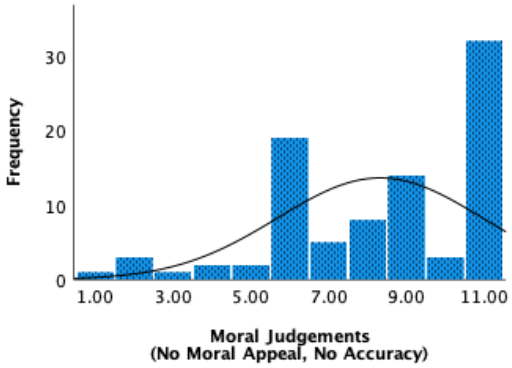
Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532–538. https://doi.org/10.1037/a0015808

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Additional exploratory analysis will also occur. These may use other demographics that are collected in the study, for example participants' political orientation and strength of identity.

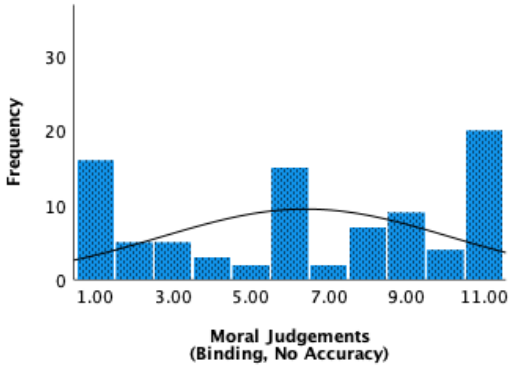
Appendix JJ

Histograms and Box Plots for Moral Judgements by Condition

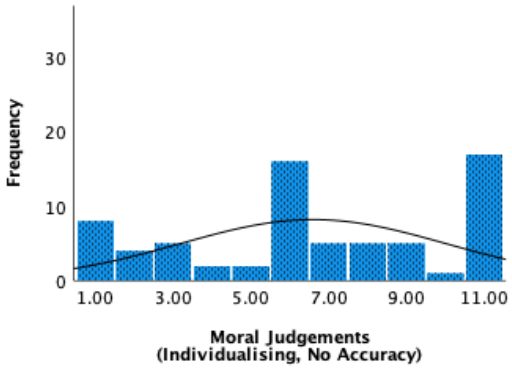
JJ1



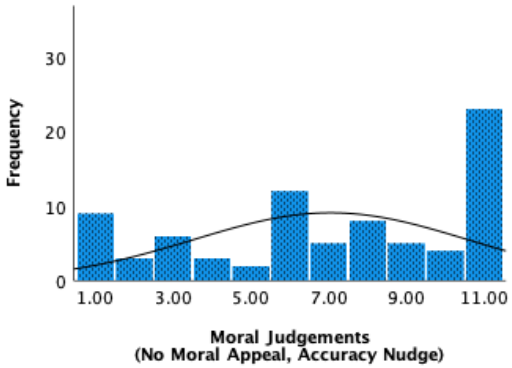
JJ2



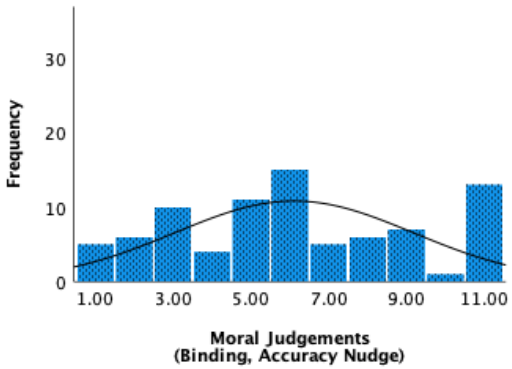
JJ3



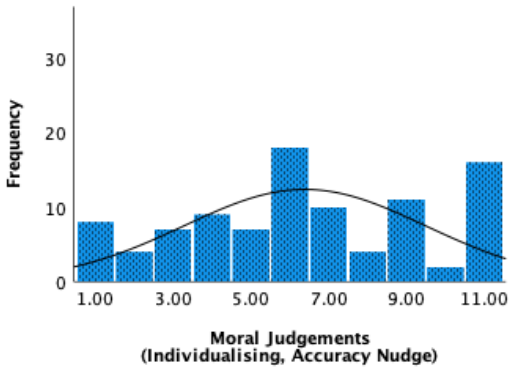
JJ4



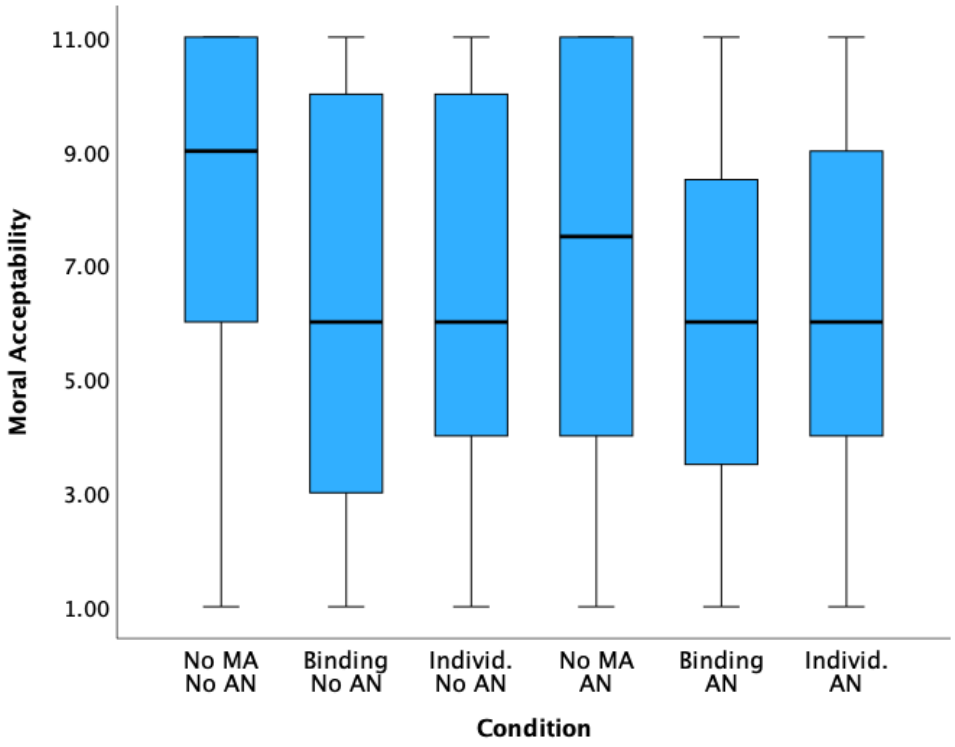
JJ5



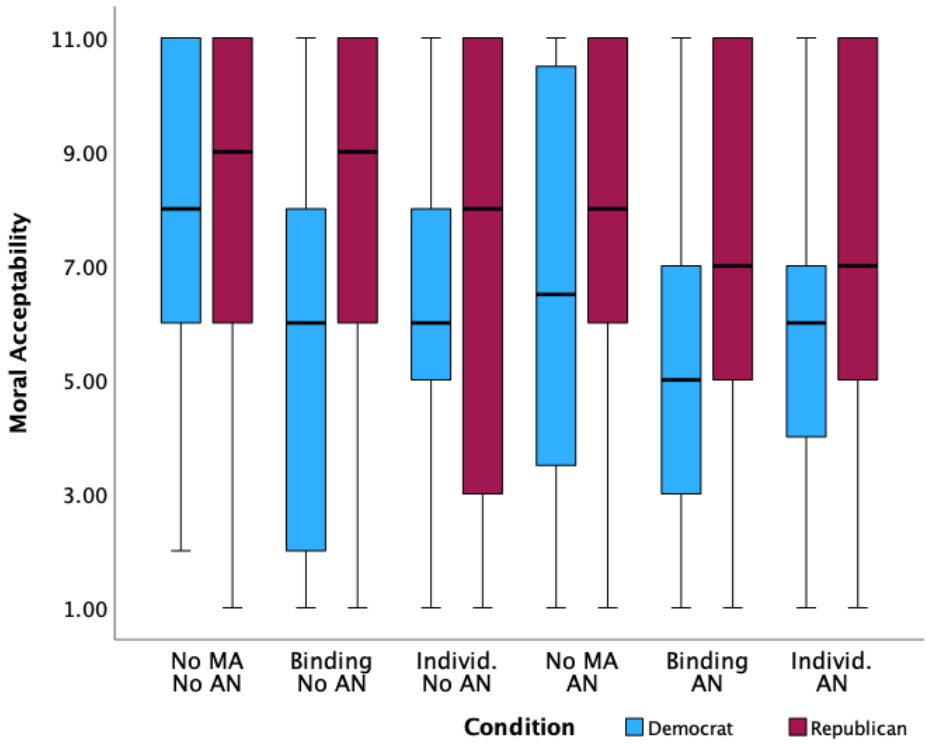
JJ6



JJ7



JJ8



Note. Panel JJ1. Histogram for No Intervention Condition. Panel JJ2. Histogram for Binding Appeal, No Accuracy Nudge Condition. Panel JJ3. Histogram for Histogram for Individualising Appeal, No Accuracy Nudge Condition. Panel JJ4. Histogram for Accuracy Nudge, No Moral Appeal Condition. Panel JJ5. Histogram for Accuracy Nudge, Binding Appeal Condition. Panel JJ6. Histogram for Accuracy Nudge, Individualising Appeal Condition. Panel JJ7. Boxplots of Moral Judgements Across Each Condition. Panel JJ8. Boxplots of Moral Judgements Across Each Condition Split by Political Affiliation.

Appendix KK

Post-hoc Comparisons of Moral Appeal on Moral Judgements

Comparison		Mean					95% *CI*	
MA 1	MA 2	Difference	*SE*	*df*	*t*	*d*	Lower	Upper
Binding	Individ.	-0.22	0.35	489	-0.62	-0.07	-0.29	0.15
Binding	No MA	-1.45	0.35	489	-4.20***	-0.46	-0.67	-0.24
Individ.	No MA	-1.23	0.35	489	-3.53***	-0.39	-0.61	-0.17

Note. Comparisons are based on estimated marginal means

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

Appendix LL

Robust ANOVA statistics for Moral Acceptability of Spreading Misinformation

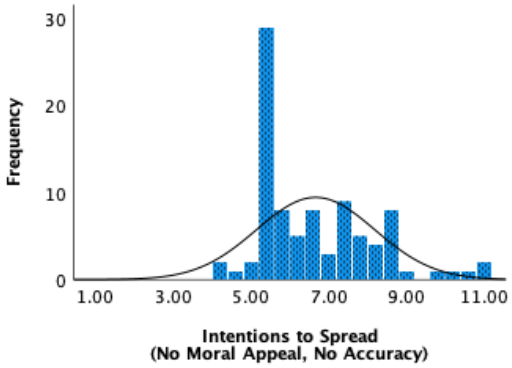
	Q	*p*
Moral Appeal (MA)	14.97	< .001
Accuracy Nudge (AN)	3.10	.08
Political Affiliation (PA)	28.51	.001
MA x AN	1.13	.57
MA x PA	3.23	.20
AN x PA	0.03	.87
MA x AN x PA	1.72	.43

Note. Method of trimmed means, trim level 0.2

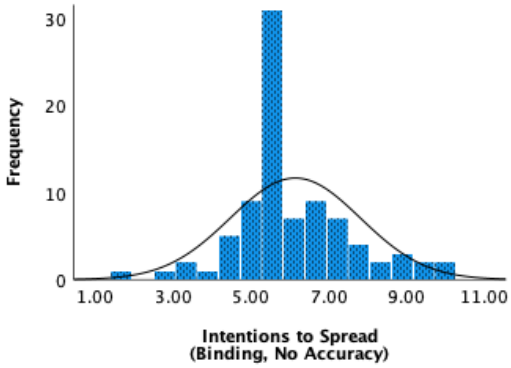
Appendix MM

Histograms and Box Plots for Intentions to Spread by Condition

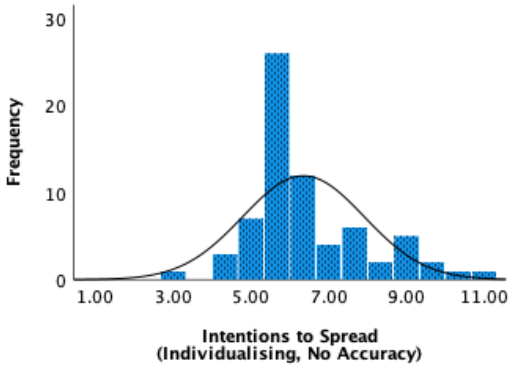
MM1



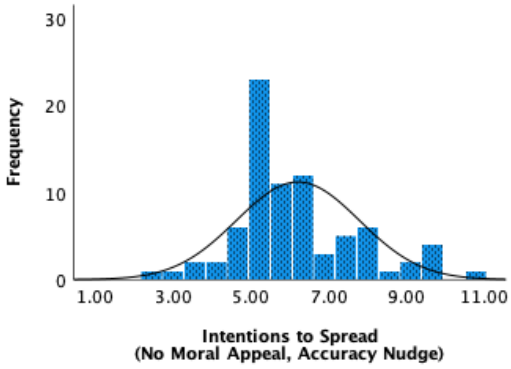
MM2



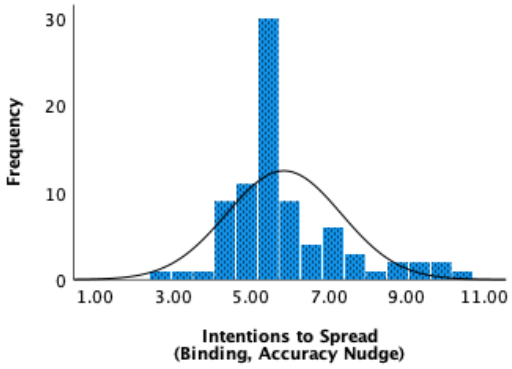
MM3



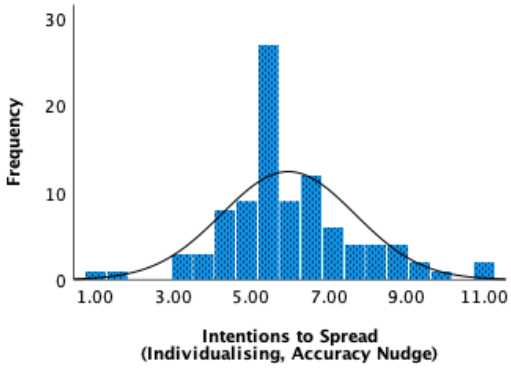
MM4



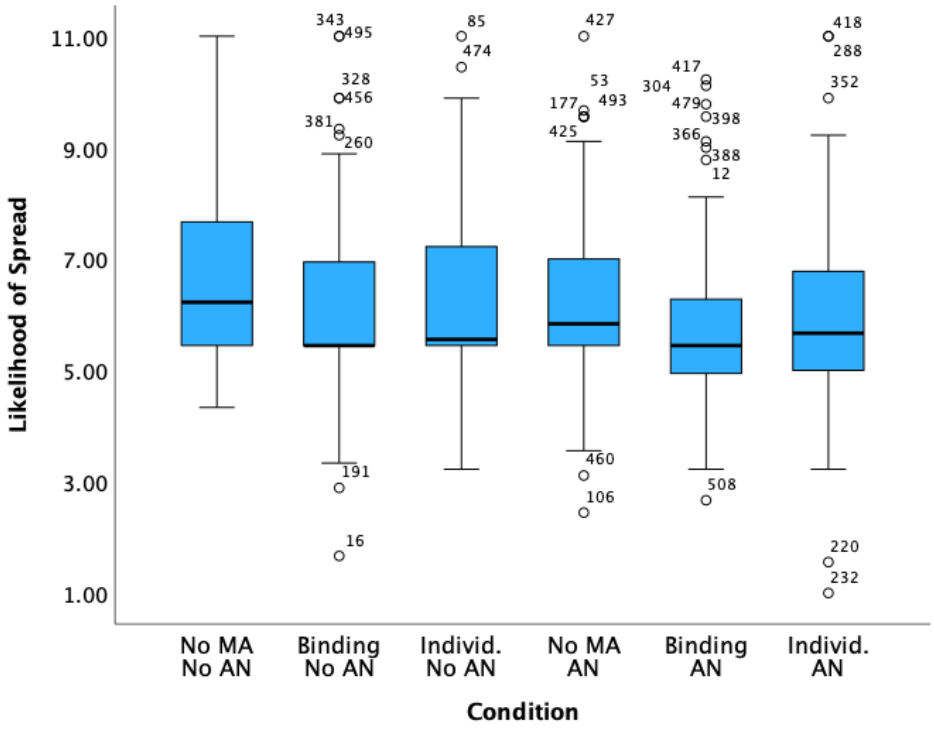
MM5



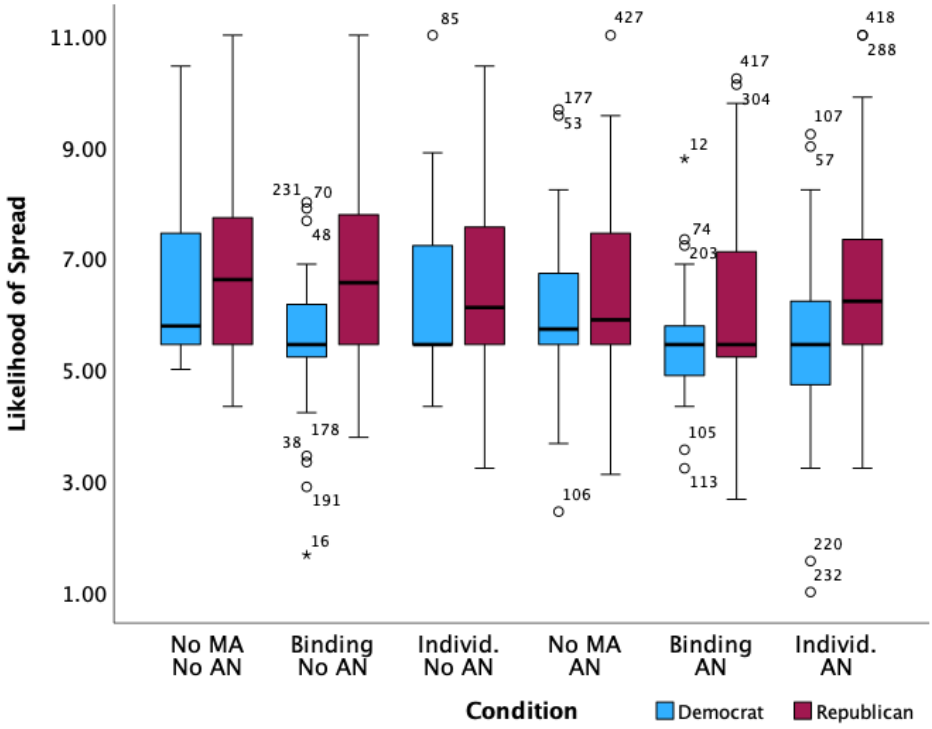
MM6



MM7



MM8



Note. Panel MM1. Histogram for No Intervention Condition. Panel MM2. Histogram for Binding Appeal, No Accuracy Nudge Condition. Panel MM3. Histogram for Histogram for Individualising Appeal, No Accuracy Nudge Condition. Panel MM4. Histogram for Accuracy Nudge, No Moral Appeal Condition. Panel MM5. Histogram for Accuracy Nudge, Binding Appeal Condition. Panel MM6. Histogram for Accuracy Nudge, Individualising Appeal Condition. Panel MM7. Boxplots of Likelihood of Spread Across Each Condition. Panel MM8. Boxplots of Likelihood of Spread Across Each Condition Split by Political Affiliation.

Appendix NN

Robust ANOVA Statistics for Intentions to Spread Misinformation

	Q	*p*
Moral Appeal (MA)	5.04	.09
Accuracy Nudge (AN)	5.77	.02
Political Affiliation (PA)	14.70	.001
MA x AN	0.63	.73
MA x PA	1.30	.53
AN x PA	0.15	.70
MA x AN x PA	2.37	.31

Note. Method of trimmed means, trim level 0.2