Examining the Impact of a Reasoning Aid to Help People Evaluate the Evidentiary Weight of Consensus

Hannah Le Leu

This thesis is submitted in partial fulfillment of the Honours Degree of Bachelor of Psychological Science

School of Psychology

The University of Adelaide

September 2021

Word Count: 9490

Table of Contents

List of Figures

List of Tables

Abstract

Social media is a vortex of information and people may see distorted views of consensus, where the independence of information and sources is unclear. A tool that summarises consensus information might help people to navigate these important cues. This study examined whether a reasoning aid (in the form of a diagram) visually illustrating both the number of independent people supporting/disagreeing with a claim and the diversity of arguments would persuade people to change their original beliefs. Participants (n=605) were recruited through Amazon's Mechanical Turk to evaluate 24 claims on a mock Twitter interface. Participants were randomly assigned to conditions with either tweets only, diagram only or tweets with a diagram. Participants rated their initial agreement level (0-100) with each claim and then saw the diagram and/or set of tweets, then were able to update their agreement level if their original opinion had now changed. The findings of this study show that without assistance, people mostly rely on cues of argument quantity, such as the number of tweets for a given stance. However, when presented with a diagram, people were able to utilise cues of argument quality, such as when there were different sources providing the information and when multiple arguments were used.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge, this thesis contains no material previously published except where due reference is made. I give permission for the digital version of this thesis to be made available on the web, via the University of Adelaide's digital thesis repository, the library search and through web search engines, unless permission has been granted by the School to restrict access for a period of time.

Signed

September 2021

Acknowledgments

Firstly, I want to thank my amazing supervisors Dr Rachel Stephens and Dr Keith Ransom for making this research available to me and dedicating hours of their time to the success of this study. I want to thank them both for all of their time they put aside to meet with me and provide helpful critiques to improve my writing and critical thinking abilities. I would also like to thank my partner, Mitch, and my family, especially my Dad, for their unwavering support over the year.

Contributions

In writing this thesis, my supervisors and I collaborated to develop the research aims of interest, design the appropriate methodology and pre-register the study on AsPredicted. I conducted the literature search and wrote additional tweets required for the stimulus set expansion. All materials were adopted from Ransom et al. (2021) for the control group (TWEETS ONLY), along with the expanded stimuli and diagrams. My supervisors programmed the mock Twitter interface to be used for the study and created the diagram I helped design. My supervisors helped me code the R script required for the statistical analyses, but I ran the script and interpreted the analyses. The study was funded by a grant supported by the *AI for Decision Making Program*, Department of Defence and the Office of National Intelligence, delivered in partnership with the Defence Innovation Partnership in South Australia (Project ID: 167650398).

1. Introduction

Consider a claim such as “police should wear body cameras”, or “phone calls create stronger bonds than texts or emails”. What is the basis of your degree of belief in each claim? We do not have direct access to the relevant evidence for many of the claims we come across — for example, we probably have not personally studied police officers with or without body cameras, and although we may have personal experience with phone calls and social bonds, we probably have not systemically collected the relevant observations. Instead, we must rely upon indirect access to evidence through information from other people, which often includes arguments from people on social media. Reasoning based on information from social media adds a new level of complexity as there is an abundance of different quality information, but it is unclear how we weight argument quality. It is also unclear how sensitive we are to cues that suggest information is good quality, such as when there are different sources supporting the information or a wide variety of arguments reinforcing it.

1.1 Category-Based Induction and Sampling Assumptions

In relatively simple reasoning tasks, people determine the strength of arguments by utilising properties of the arguments themselves (Dowden, 2017). In category-based induction tasks, people view an argument in the form of a list of sentences, with any number of premises and one conclusion (Ranganath et al., 2010). The task usually requires a person

to utilise the premises in order to consider some kind of generalisation to the conclusion. For example, people may be asked to rate the strength of the argument, “horses have a merocrine gland and monkeys have a merocrine gland, so all mammals must have a merocrine gland” (Heit, 2000). Stronger arguments are considered arguments where the premises are more likely to result in people believing the conclusion. A larger number of premises that support the conclusion generally lead to more compelling arguments. Consider a scenario that discusses which animals possess “sesamoid bones”, which is referred to as property *P.* Giving three examples of animals that have *P* rather than only two examples of animals that have *P* has been shown to increase argument strength. However, it is important that the additional examples increase stimulus diversity, as similar examples add limited new information to guide generalisations (Osherton et al., 1990). For example, stating that both cows and sheep share *P* is not enough to conclude that all mammals share *P* because they are too similar. Cows and sheep are too typical and similar in nature, so both of them sharing *P* is not enough to generalise *P* to all mammals. In contrast, stating that horses and monkeys share *P* seems to add more strength to the conclusion that all mammals share *P* (Spellman et al., 1999). Furthermore, any additional arguments should be strong otherwise this could lead to the weak evidence effect. This occurs when weak arguments are added to strong arguments, thereby weakening the effect overall (Friedrich & Smith, 1988). For example, if a reference letter for a job candidate focuses on the candidate’s neat business attire this may weaken their overall application; even though dressing neatly is positive evidence for a good candidate, it implies a lack of stronger reasons.

When reasoning, people are also known to take into account assumptions of how the examples were generated to decide whether the conclusion is believable. For example, consider two arguments: 1) “German Shepherds have *P*, therefore all dogs have *P*”; 2) “German Shepherds have *P,* Dobermans have *P,* Rottweilers have *P*, therefore all dogs have

P." Interestingly, people tend to rate the first argument as more believable, even though it only has one example, perhaps due to the belief that the experimenter has purposefully provided maximally informative examples in each argument. This means that for the second argument, people believe *P* must only belong to larger, stereotypically aggressive dogs (Medlin et al., 2003). People may also assume that the examples are sampled at random from dogs that possess *P.* Since the three examples randomly selected all happen to be larger, stereotypically aggressive dogs, this could imply that it is statistically unlikely that all types of dogs possess *P* (Kemp & Tenenbaum, 2009; see also Hendrickson et al., 2019; Perfors et al., 2014).

However, real-life everyday reasoning is often far more complex. For example, imagine you came across a "trending" claim that "police should wear body cameras" on your Twitter feed. When deciding whether you agree with the claim, would you rely on the tweets that you saw, and/or how you assume the tweets were generated? Considering the tweets alone, compared to the simple category-based induction tasks, the structure of the conceptual space where this problem would be represented is far more complex (Ransom et al., 2021). The space is high-dimensional and there are fewer prior data to rely on when determining your stance on the issue. Assumptions of how the arguments in the tweets were generated also extend far beyond informative or random sampling (Ransom et al., 2021). It can be extremely difficult to determine which assumption applies, such as whether a person positing that police should wear cameras has any stake in this issue; perhaps they own a camera technology company. Sampling assumptions become even more complicated when considering the independence of individual arguments. If one person posts multiple times to Twitter, it is not clear whether they acquired and verified these arguments independently, or if they were derived post-hoc from the conclusion. In real-life scenarios it is also difficult to know the quality of an apparent consensus. If hundreds of people all have the same opinion,

did they generate this opinion independently, or did they all generate this same opinion because they viewed the same source (Yousif et al., 2019)?

1.2 Perceived Consensus

A perceived consensus can be highly — and sometimes dangerously — influential; the number of people in favour of a claim is used as an important cue to believability. Classic research on conformity emphasises how people over-rely on the perceived consensus, even when this consensus is clearly wrong (Asch, 1956). Asch (1956) examined how people reasoned when all other participants in the room were confederates and gave an obviously wrong answer. People were shown a range of different lines, along with a target line, and were asked to select which line most closely matched the target line, answering in front of everyone (the real participant always chose last). The study found that 75% of people were willing to select an obviously incorrect answer at least once if the rest of the people in their group selected this wrong answer before them.

1.3 Sensitivity to the Quality of a Consensus

Although it is clear that a perceived consensus is a powerful cue that a claim should be believed, there is little research on people's sensitivity to the quality of a consensus or whether people distinguish between different types of consensus, based on independent or dependent evidence (Yousif et al., 2019). To illustrate, imagine that you were unsure about a newly proposed policy and went around your office at work to find out your colleagues' opinions. If every single co-worker you approached gave the same answer in support of the

policy, that might seem like a pretty convincing consensus. However, what if you found out that all of your colleagues received their information from the exact same source, rather than independently coming to this conclusion? Would this change your perspective on the apparent "consensus", and if so, why? This scenario illustrates a "false" consensus, as even though it appears that there is a majority supporting the policy, there is actually only one known original source that supports it, and then a lot of repetition of this source (Yousif et al., 2019). Interestingly, there are contradicting findings regarding whether people are able to utilise cues of consensus quality, or if they are persuaded by "false" consensuses (Harkins & Petty, 1981; Ransom et al., 2021; Yousif et al., 2019).

Yousif et al. (2019) explored whether people were sensitive to the difference between a "true" consensus (based on independent primary sources) and a "false" consensus (based on a shared primary source) over five different experiments. In Experiment 1, participants were shown a range of news articles about a claim and were asked to rate their belief in the claim. The number of secondary sources (news articles) that contributed and the number of primary sources that were relied upon in the articles were varied. The results showed that people were not at all sensitive to the difference between a "true" or "false" consensus. This finding was repeated in Experiments 2, 3, 4 and 5, with this lack of sensitivity to consensus quality persisting through multiple manipulations. The lack of sensitivity occurred regardless of expertise, when participants were explicitly told to attend to the source(s), even when people made prior ratings that they preferred a true consensus, and for directly perceivable events (i.e., eyewitness accounts of a bear sighting).

On the other hand, Harkins and Petty (1981) found that having three different people give different arguments in support of a claim was far more persuasive than one person giving three different arguments. Even though the exact same content and amount of information was presented, people were more sensitive to the number of different people

presenting the arguments. This shows that people were more compelled by diverse authors (equivalent to multiple "secondary" sources in Yousif et al., 2019), as a cue of a higher-quality consensus. This supports an "argument of the pool" theory, whereby knowing that several different people generated a range of different arguments may have led reasoners to infer that there must be a large pool of reasonable arguments in support of the claim (Harkins & Petty, 1981). One person using different arguments does not imply this as the individual may have just exhausted the pool of reasonable arguments (Harkins & Petty, 1981).

Ransom et al. (2021) also recently investigated whether people were more sensitive to the quantity of evidence, or the quality of evidence for a claim. Simply because there is a large quantity of evidence, such as an abundance of social media posts supporting a claim, does not mean this evidence is good quality. The researchers explored what cues to evidence quality people were sensitive to on social media; specifically, whether people were more sensitive to the *number of posts* made in support of a claim (quantity), or to cues of quality such as the *number of people* supporting a claim and whether they used *diverse* or *repeated* arguments to support a claim. Participants rated their initial agreement with a range of claims (e.g., "charitable giving will increase in the next three years") on a scale of 1-100 and were then shown a sample of tweets arguing against or in favour of the claim. The level of support for each claim depended on the randomly assigned condition: full consensus (4 vs. 0 tweets), majority consensus (4 vs. 1 tweets) and contested consensus (4 vs. 4 tweets). Each trial, the tweets for a target stance (for or against) differed in whether they were written by different people or the same person and whether there were different arguments, or the same general argument repeated in each tweet. After viewing the tweets, participants had the chance to update their agreement rating. The study found that reasoners were most sensitive to quantity cues (number of tweets for vs. against the claim) and showed limited sensitivity to cues regarding the quality of information (people and argument diversity). In fact, reasoners were

slightly more persuaded when the tweets repeated the *same* argument than when there were different arguments, even when it was the same person tweeting the same argument multiple times.

Concerningly, Weaver (2007) also found that people can be persuaded by the repetition of a single opinion. The study found that participants preferred when different people stated the same opinion; however, participants were more persuaded when one person repeated the same point multiple times than when a person stated the same point once (Weaver, 2007). This finding shows that people often infer that a repeated opinion is a prevalent one, even when this repetition comes solely from one group member and participants are indeed aware that it is only from one person (Weaver, 2007). If people treat repetition from one source as a cue suggesting that the opinion is widespread, this could have huge implications on social media, where people have the ability to post repetitively.

1.4 Reasoning Tools on Social Media

Together, the studies by Harkins and Petty (1981), Ransom et al. (2021), Weaver (2007) and Yousif et al. (2019) highlight that people do not seem to reliably attend to important cues about argument strength and consensus quality, and need help to make these cues more salient. This is critical on social media platforms where people encounter more information than they can evaluate systematically (Gunaratne et al., 2020). People do not have the cognitive capacity, motivation, or time to evaluate every piece of information they view online. The fact that low-credibility information is able to spread rapidly and easily suggests people are vulnerable to manipulation and in need of some type of reasoning aid or intervention to reason more effectively (Shu et al., 2017). Various social media platforms have attempted to address this issue through implementing warning labels on social media content that a third-party fact-checker has disputed (Koch et al., 2021). Fact-checkers have

the potential to significantly reduce the proliferation and impact of misinformation, through debunking false claims and influencing the likelihood users see misinformation (Allen et al., 2021).

However, there are doubts about the effectiveness of current fact-checking warnings as a third-party must examine every new piece of information to either verify or dispute it (Pennycook et al., 2020). This is problematic as it is significantly easier to create misinformation than it is to check its accuracy, meaning there will only be a limited amount of misinformation that is successfully labelled with warnings. Troublingly, when a warning is absent, it might create an "implied truth effect" and users may assume the information has been verified, even if it is inaccurate (Allen et al., 2021). Even when fact-checking warnings are successfully implemented, there is a lack of public trust in the objectivity of the warnings (Allen et al., 2021). However, Shu et al. (2019) found that fact-checker warnings would be significantly more trustworthy and effective if they offered explanations for their recommendations; specifically, if they included a sample of user posts that guided the refutation. Another limitation is that warning labels are generally only used on headlines that a third-party assesses to be blatant misinformation, which whilst is a serious type of misinformation, is far from the only form (Pennycook et al., 2020). These warnings do not consider more ambiguous types of information such as conspiracy theories that associate real events with nonsensical conclusions. Some information may also seem questionable but is not exactly debunkable; for example, claiming that "it will be impossible to find good quality avocados in 10 years".

1.5 Crowd-Sourced Reasoning

The limitations of third-party fact-checkers have led to research into using the "wisdom of the crowds" effect to help people evaluate the veracity of information on social

media (Collins et al., 2021). This is a well-documented effect whereby independent judgements are aggregated to create a combined judgement with high accuracy (Simoiu et al., 2019). The wisdom of the crowds effect has been reported in a wide variety of contexts and domains, such as answering general knowledge questions (Rauhut & Lorenz, 2011), identifying online phishing and scams (Moore & Clayton, 2008; Liu et al., 2012), and predicting COVID-19 mortality rates by region months in advance (Turiel & Aste, 2020). Pennycook and Rand (2019) explored whether the wisdom of the crowds effect could be utilised to judge online news sources. They found that crowd-sourced judgements accurately identified the reliable and unreliable news sources, and are more effective than fact-checkers as they are scalable (Pennycook & Rand, 2019). Utilising the collective intelligence of online communities to discern questionable information could solve the problem of limited time and resources, whilst also avoiding issues with public trust of third-party fact-checkers (Collins et al., 2021). Researchers are currently working on automated fact-checking algorithms that uses a wisdom of crowds approach by utilising user posts (Shu et al., 2017). However, these algorithms will still struggle to establish ground truths for more ambiguous types of misinformation.

1.6 A Novel Reasoning Aid

The current study proposes a new type of tool to assist people to reason in a way that is better calibrated with the available evidence. This reasoning aid is based on the concept and design of multidimensional scaling (MDS) used (for example) to illustrate the similarity of categories in psychological space (Hout et al., 2015) and topic space visualisation diagrams (Ajjour et al., 2017). It is challenging to attempt to quantify heterogeneity (diversity) of arguments and reasoners. In MDS, the more diverse items are from each other, the more distance there is between them in space (see Appendix A). Topic space visualisation

has also been used to represent the diversity of items (arguments), as seen in Ajjour et al. (2017). Ajjour et al. (2017) depicted the topic space (a controversial claim) through a regular polygon shape, with one vertex for each represented argument topic (see Appendix B). Coloured dots represent specific arguments (green dots for pro arguments and red dots for con arguments) and are plotted in the pentagon, where arguments that are closely related to an argument topic are plotted closer to the relative vertex(s). The current study uses distance to capture the diversity of arguments — similar/repeated arguments are represented by closely grouped icons, while dissimilar argument are more widely distributed.

1.7 The Current Study

This thesis aims to investigate the effectiveness of a new reasoning aid, designed to help people navigate cues to the quantity (number of posts) and quality (diversity of sources and arguments) of consensus information in a social media context. The reasoning aid draws on the concept of crowd-sourced judgements where lay-people have provided their agreement or disagreement with a claim and an argument or justification, so that a summary of this information potentially resembles a judgement of high accuracy. The reasoning aid created is in the form of a diagram that illustrates the number of different people supporting/refuting a claim, the number of Twitter posts these people made, and whether they used diverse arguments or repeated an argument multiple times in their posts. If shown to be effective, this kind of reasoning aid could be used alongside complementary tools such as automated fact-checking services.

The current experiment extends on the paradigm explored by Ransom et al. (2021) and investigates whether belief revision is affected by the number of tweets, whether there are different people or the same person tweeting multiple times and whether diverse or repeated arguments were used. This experiment also includes the addition of a diagram to explore how

making these factors salient through a visual representation affects, or more specifically, improves reasoning with the available evidence. The first aim is to replicate the key effects that without an aid, people are more sensitive to quantity than quality of consensus evidence. Some methodological improvements were also made to Ransom et al. (2021), such as controlling the total number of tweets across *Pro* and *Con* arguments. The second aim is to compare people's reasoning when the reasoning aid is added to the posts. It is expected that the aid may improve sensitivity to the quality of a consensus. A third aim and condition explores the effect of the diagram alone, with no example tweets. This condition is expected to have less of an overall effect than the condition that includes the diagram with tweets, as summary statistics on consensus have been shown to be less effective than displaying example opinions (Harris et al., 2019). Overall, it is hypothesised that the reasoning aid will reduce the influence of the number of tweets relative to the effects of whether there are different tweeters and diverse arguments. The aid is also hypothesised to increase the difference in the relative persuasiveness of diverse versus repeated arguments.

2. Method

To examine the potential usefulness of a novel reasoning tool on social media, I conducted an online experiment in which participants viewed arguments supporting and/or refuting claims on a mock Twitter interface. I explored whether a reasoning aid in the form of a diagram would draw people's attention to argument quality through visually representing the number of different tweeters and their tweets for/against each claim, and the number of different arguments used to support the target and opposing stances.

2.1 Pre-Registration

The study's variables, hypotheses and planned analyses were pre-registered on AsPredicted before any data were collected (https://aspredicted.org/blind.php?x=/6T9_769).

2.2 Design

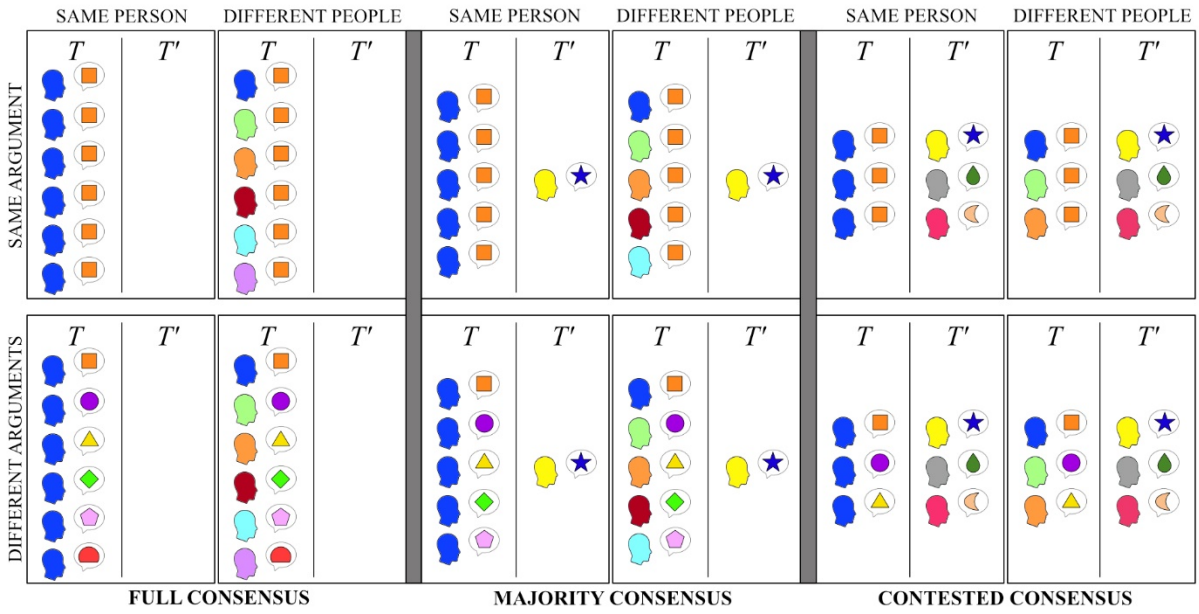
The study used a 3 (INFORMATION LEVEL: TWEETS ONLY vs. DIAGRAM ONLY vs. TWEETS WITH DIAGRAM) × 3 (NUMBER OF TWEETS: FULL vs. MAJORITY vs. CONTESTED CONSENSUS) × 2 (ARGUMENT DIVERSITY: DIVERSE vs. NON-DIVERSE) × 2 (AUTHOR DIVERSITY: SAME TWEETER vs. DIFFERENT TWEETERS) factorial design, with the last three factors illustrated in Figure 1. INFORMATION LEVEL was varied between subjects, whilst NUMBER OF TWEETS, ARGUMENT DIVERSITY and AUTHOR DIVERSITY were manipulated within subjects.

2.2.1 Information Level

Participants assigned to the TWEETS ONLY condition saw the claims and a variety of tweets supporting/refuting each claim (see Figure 2). Participants assigned to the DIAGRAM ONLY condition saw the claims and the diagrams summarising the number of tweets,

Figure 1

Experiment Design



Note. Participants were randomly assigned to an INFORMATION LEVEL group where NUMBER OF TWEETS, ARGUMENT DIVERSITY and AUTHOR DIVERSITY were manipulated within-subjects. The number of reply *Target Tweets* (*T*) versus *Opposing Tweets* (*T'*) varied by the NUMBER OF TWEETS, with either a FULL, MAJORITY or CONTESTED CONSENSUS. For *T* there was either one person tweeting multiple times or different people tweeting. They either repeated the same argument or used different arguments in their tweets.

argument diversity and the number of different tweeters for each claim. The corresponding tweets were displayed on the screen (including tweeter information), but the tweet contents

were blurred out (see Figure 3). Participants in the TWEETS WITH DIAGRAM condition saw the claims, the diagram, and the reply tweets supporting/refuting the claim (see Figure 4).

2.2.2 NUMBER OF TWEETS

The number of reply tweets supporting either side of the claim was varied. There were three within-subjects levels: 6:0 tweets (FULL CONSENSUS); 5:1 tweets (MAJORITY CONSENSUS); and 3:3 tweets (CONTESTED CONSENSUS). In the FULL CONSENSUS condition, the reply tweets only consisted of *Target Tweets* with no *Opposing Tweets* (i.e., six tweets arguing in favour or against the target claim and no tweets opposing them). For the MAJORITY CONSENSUS there were a majority of *Target Tweets* (5) with one *Opposing Tweet*. There was no numerical advantage for the CONTESTED CONSENSUS as there were an equal number of *Target Tweets* (3) as there were *Opposing Tweets* (3). The current study added a methodological improvement to Ransom et al. (2021) by holding constant the total number of tweets across the FULL, MAJORITY and CONTESTED CONSENSUS conditions. Another methodological improvement was the manipulation of the NUMBER OF TWEETS within subjects, unlike Ransom et al. (2021).

Figure 2

TWEETS ONLY Condition

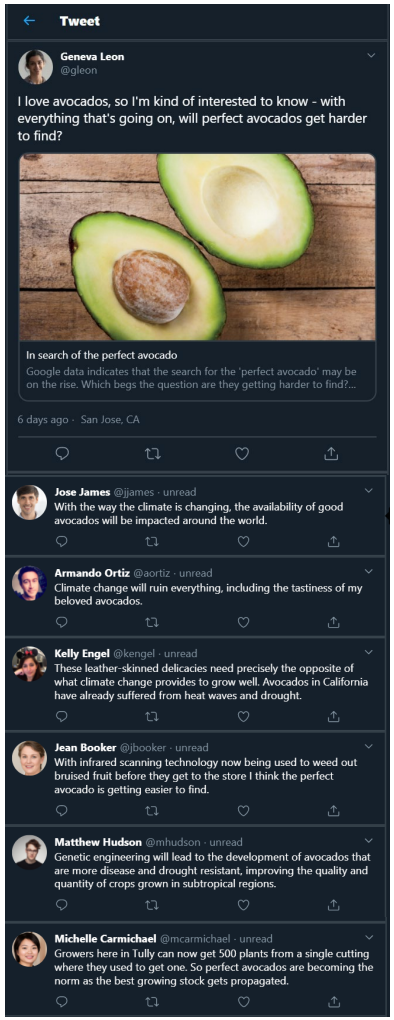


Figure 3

DIAGRAM ONLY Condition

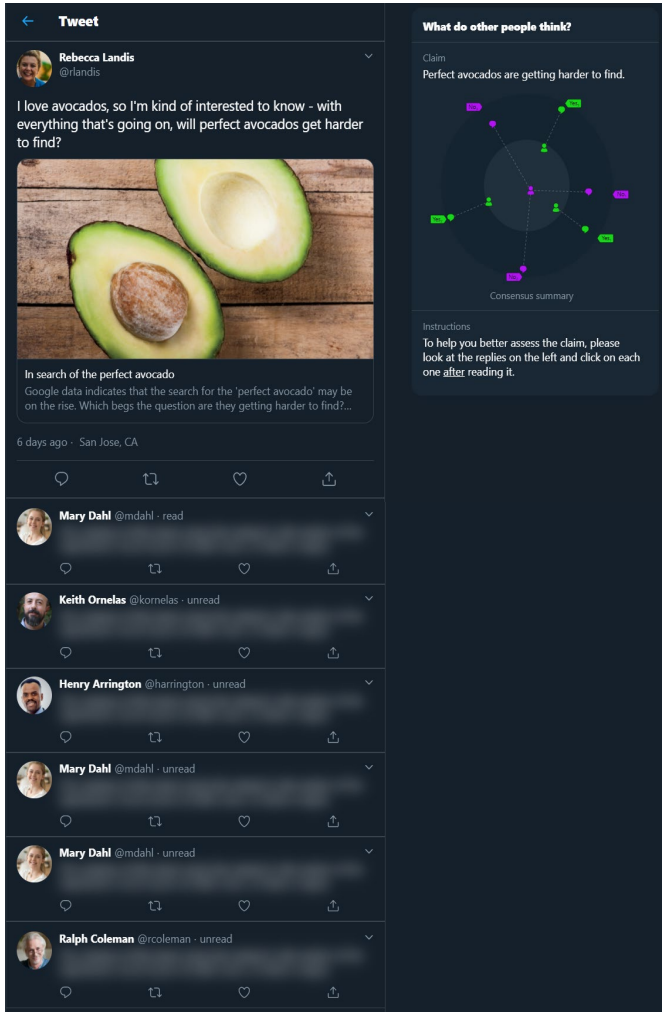
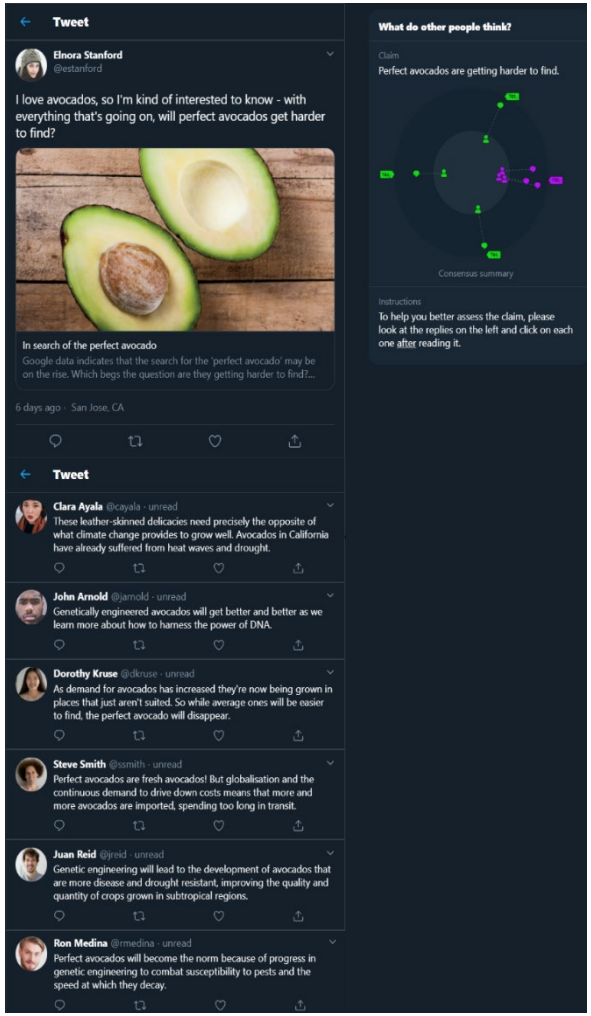


Figure 4

TWEETS WITH DIAGRAM Condition



2.2.3 ARGUMENT DIVERSITY

Within-subjects, the tweets were manipulated to present either a range of different arguments or reword the same argument multiple times. The DIVERSE tweets clearly advanced different arguments (see Table 1), whilst the NON-DIVERSE tweets conveyed the same core message, with shared key-words to enhance similarity (see Table 2).

Table 1

DIVERSE Tweets for the Claim, "Perfect Avocados are Getting Harder to Find."

#	Pro	Con
1.	"As demand for avocados has increased they're now being grown in places that just aren't suited. So while average ones will be easier to find, the perfect avocado will disappear."	"Growers here in Tully can now get 500 plants from a single cutting where they used to get one. So perfect avocados are becoming the norm as the best growing stock gets propagated."
2.	"Perfect avocados are fresh avocados! But globalisation and the continuous demand to drive down costs means that more and more avocados are imported, spending too long in transit."	"With infrared scanning technology now being used to weed out bruised fruit before they get to the store I think the perfect avocado is getting easier to find."
3.	"Monoculture crops like avocados gradually kill off the soil that supports them. Given that they can't grow everywhere it seems that high quality is not sustainable in the long run."	"The big supermarket chains have done a lot to improve the avocado by rewarding farmers for consistency and quality rather than just volume."
4.	"These leather-skinned delicacies need precisely the opposite of what climate change provides to grow well. Avocados in California have already suffered from heat waves and drought."	"Genetic engineering will lead to the development of avocados that are more disease and drought resistant, improving the quality and quantity of crops grown in subtropical regions."
5.	"Great avocados require a lot of resources and care, so economic cutbacks due to the	"With social media hype increasing the demand for perfect avocados, market

	pandemic will lead to lower quality produce."	forces will ensure that suppliers are motivated to deliver the goods."
6.	"For goods like fruit and veg, there is always more profit in speed and high volume rather than quality, so the perfect avocado will gradually be replaced by tasteless rubbish."	"With the mild summer we just experienced, avocados are growing in abundance! There's more avocados so there's a lot more chance to find a perfect one!"

Table 2

NON-DIVERSE Tweets for the Claim, "Perfect Avocados are Getting Harder to Find."

#	Pro	Con
1.	"These leather-skinned delicacies need precisely the opposite of what climate change provides to grow well. Avocados in California have already suffered from heat waves and drought."	"Genetic engineering will lead to the development of avocados that are more disease and drought resistant, improving the quality and quantity of crops grown in subtropical regions."
2.	"Climate change will ruin everything, including the tastiness of my beloved avocados."	"Genetically engineered avocados will get better and better as we learn more about how to harness the power of DNA."
3.	"With the way the climate is changing, the availability of good avocados will be impacted around the world."	"Perfect avocados will become the norm because of progress in genetic engineering to combat susceptibility to pests and the speed at which they decay."
4.	"Climate change means that Mexico (the world's largest producer of avocados) stands to lose half its workable farms in 10 years..."	"The quality of avocados will only increase as farmers gain access to genetically engineered plants."
5.	"The extreme weather events caused by climate change is making it harder to grow enough avocados to keep up with demand, let alone decent avocados."	"Support for genetically modified crops is on the rise, so we are sure to see an increase in the quality of avocados."

6.	"We're having more extreme weather events everywhere - climate change! Since avocados require a stable climate, perfect ones are going to get increasingly rare."	"Recent advances in genetic engineering means that great avocados will become more plentiful."

2.2.4 AUTHOR DIVERSITY

AUTHOR DIVERSITY for each claim was also manipulated within-subjects. The tweets would either come from a variety of DIFFERENT TWEETERS or the SAME TWEETER. A distinct user icon and name was randomly allocated and displayed alongside each tweet to convey this information (see Figures 2-4).

2.3 Materials

2.3.1 Claims and Tweets

Participants were shown 24 Twitter posts containing claims about a range of topics (Table 3). The same 20 topics were used as Ransom et al. (2021) with four additional topics created by using arguments found online, including on debate sites and social media. The claims were chosen to vary in plausibility and nature; for example, opinion based, technical topics, or eyewitness accounts. For each participant the claims were randomly allocated to the conditions of AUTHOR DIVERSITY, ARGUMENT DIVERSITY and NUMBER OF TWEETS, such that there were two trials per cell. The stimulus set was expanded from that of Ransom et al. (2021) so that there were six diverse *Pro* tweets, six diverse *Con* tweets, six repeated *Pro* tweets and six repeated *Con* tweets for every claim (see Tables 1-2). Minor edits were made to some tweets to update them (e.g., to "Children learn more effectively by handwriting than by typing.").

Table 3

List of Claims

#	Claims
1.	"Police should wear body cameras."
2.	"Golf is a sport."
3.	"Phone calls create stronger bonds than text or emails."
4.	"Children learn more effectively by handwriting than by typing."
5.	"People are sleeping more during lockdown."
6.	"Working from home is more productive."
7.	"School uniforms are a good idea."
8.	"Narcissists are more politically engaged."
9.	"Genetically modified crops are a good idea."
10.	"Manchester City fans started the fight."
11.	"Perfect avocados are getting harder to find."
12.	"Investment in clean coal technology will help the environment."
13.	"Medical marijuana should not be used for pets."
14.	"Charitable giving will increase over the next three years."
15.	"People are likely to be more tolerant toward other racial groups and nationalities having experienced a pandemic."
16.	"The movie 'Lofty Heights' will be popular."
17.	"Britain's economy will improve as a result of Brexit."
18.	"Standardised testing should be used more widely in schools."
19.	"Hydraulic fracturing for gas production should be encouraged."
20.	"Lockdowns should be abandoned."
21.	"Capital punishment should be abolished."
22.	"Advancing AI will do more harm than good."
23.	"A college degree is worth it."
24.	"It's time we became a cashless society."

2.3.2 Diagrams

Figures 5, 6, 7 and 8 step through the features of an example diagram. The diagrams summarise information about both the *Pro* and *Con* sides of the claim through green and purple colours, respectively. AUTHOR DIVERSITY is depicted through the number of green or

purple people icons for *Pro* and *Con* (see Figures 5 and 6). The NUMBER OF TWEETS is illustrated through the number of green (*Pro*) and purple (*Con*) speech icons (see Figures 7 and 8). ARGUMENT DIVERSITY is shown via the number of *Pro* and *Con* labels and how spread out or clustered the speech icons are (see Figures 7 and 8). There were 12 different diagram types in total according to the three factors in Figures 5-8 and each diagram type was randomly jittered for each trial.

2.4 Procedure

Before the experiment, participants were given instructions on what the experiment involved (i.e., reading claims with reply tweets and making ratings) and were asked to complete three verification questions to 100% accuracy to make sure they read and understood the instructions. Participants who failed the verification questions were shown the instructions again. Participants also completed demographic information. Participants in the conditions with diagrams were then shown detailed instructions of how to interpret the diagrams and asked to complete three multiple choice questions to 100% accuracy before beginning the study. Participants who failed these verification questions were shown the diagram instructions again.

There were 24 trials, each with a different claim such as “Perfect avocados are getting harder to find.” Each trial began by presenting a Twitter post with one claim, including a photo and some brief contextual information (see Figures 2-4). After reading the post, participants rated their agreement with the claim on a slider (0 = do not agree at all; 100 = fully agree). Participants saw a range of reply tweets to each claim, presented in random order (except in the DIAGRAM ONLY condition where the tweet contents were blurred). The reply tweets consisted of *Target Tweets* (*T*) and *Opposing Tweets* (*T’*), which were tweets that opposed the *Target Tweets*. All *Target Tweets* were randomly set either in favour (*Pro*) or

against (*Con*) the target claim for each trial and the *Opposing Tweets* took the relative opposing stance. All participants were then given the chance to update their agreement rating. Participants were unable to update or confirm their agreement level until they marked every tweet as "read" by clicking on each, which helped to ensure participants were actually viewing the tweets — this design feature was another methodological improvement to Ransom et al. (2021). All claims were presented in random order, except the last four claims in Table 3, which were new additions to the stimulus set (to permit more direct comparisons with Ransom et al. 2021, although such analyses are beyond the scope of this thesis). There were four follow up questions after each of the final four trials, which questioned the participants on what they had just seen in the previous trial. The questions examined participants' understanding of the number of tweets, the number of people and the similarity of the tweets for a particular stance ("did the tweets seem to raise points in favour of the claim or against it?"; "did the people involved seem to agree or disagree with the claim?"; and "how similar were the tweets which argued against the claim?").

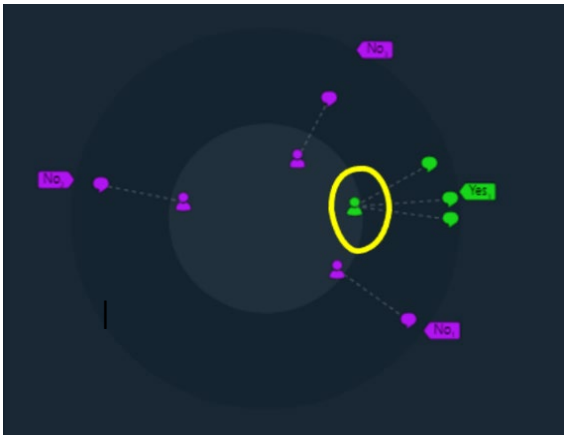
2.5 Participants

Data for the current study were collected in August 2021 using Amazon's Mechanical Turk platform. There were 605 participants that participated in the study and received $5 USD as compensation. Participants were randomly allocated to the DIAGRAM ONLY (n = 199), TWEETS ONLY (n = 208) and DIAGRAM WITH TWEETS (n =198). Participants' ages ranged between 19 and 78 (mean age 38.6), included 52.5% males and were drawn mainly from the U.S and Brazil (87.77%). The ethnic backgrounds of participants varied, with the majority of participants identifying as either White (70.58%), Asian (9.26%), Latinx (6.94%) or Black (6.78%). Most participants identified as native English speakers (79.01%), but all participants were previously screened for understanding the English language. The majority of

participants had high-school equivalent education or higher (99.17%) and identified politically as moderately liberal (23.97%). Participants were also asked about their social media use, specifically how often they use Facebook or Twitter. The most common response was daily use of Facebook (52.89%) and daily use of Twitter (39.67%).

Figure 5

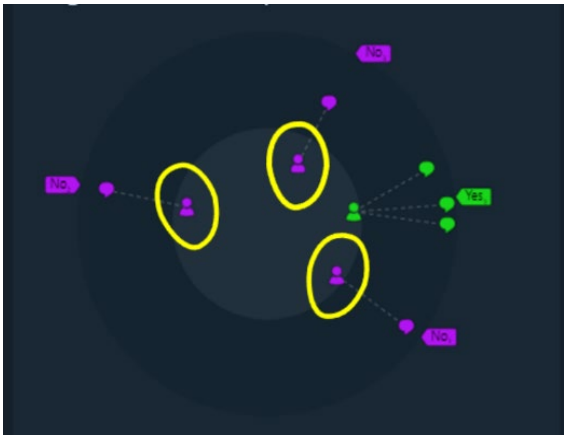
Number of Tweeters for the Claim



Note. There is one tweeter that supports this claim, circled in yellow.

Figure 6

Number of Tweeters Against the Claim



Note. There are three tweeters that disagree with this claim, circled in yellow.

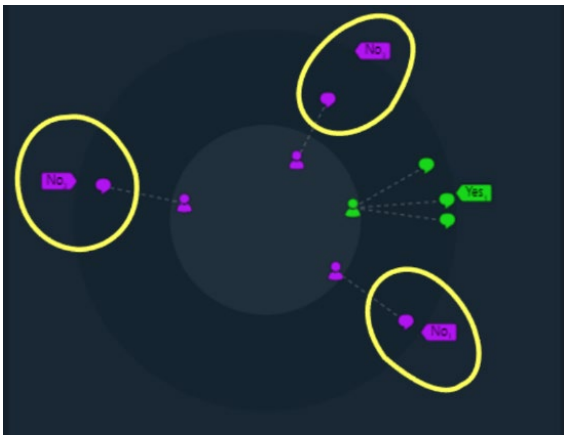
Figure 7

Number of, and Diversity of, Posts Supporting the Claim

Note. Circled in yellow, there are three posts that have been made in support of the claim and every post has repeated the same argument.

Figure 8

Number of, and Diversity of, Posts Against the Claim



Note. Circled in yellow, there are three posts that have been made against the claim and every post has used a different argument.

3.0 Results

3.1 Prior Analyses

An important assumption of the experimental design was that the topics selected for the claims would elicit a variety of prior distributions; people's prior agreement levels would vary across claims and show a variety of distributions of belief. To verify this assumption, the prior ratings were plotted by topic (Figure 9). Prior beliefs varied widely across topics and people, allowing the study's hypotheses to be tested under diverse conditions.

To check for patterns of peculiar or random responding, people's updated ratings were plotted as a function of their initial ratings for each INFORMATION LEVEL condition (Figure 10). As expected, overall there was a positive correlation between people's initial and updated ratings, $r(1422) = .82$, $p < .001$. Furthermore, the plots also show clustering of the ratings from people who viewed *Pro* target tweets in the top left corner and the ratings from the *Con* target tweets in the bottom right corner, indicating that people generally understood and were affected by the stance (*Pro*/*Con*) of the tweets.

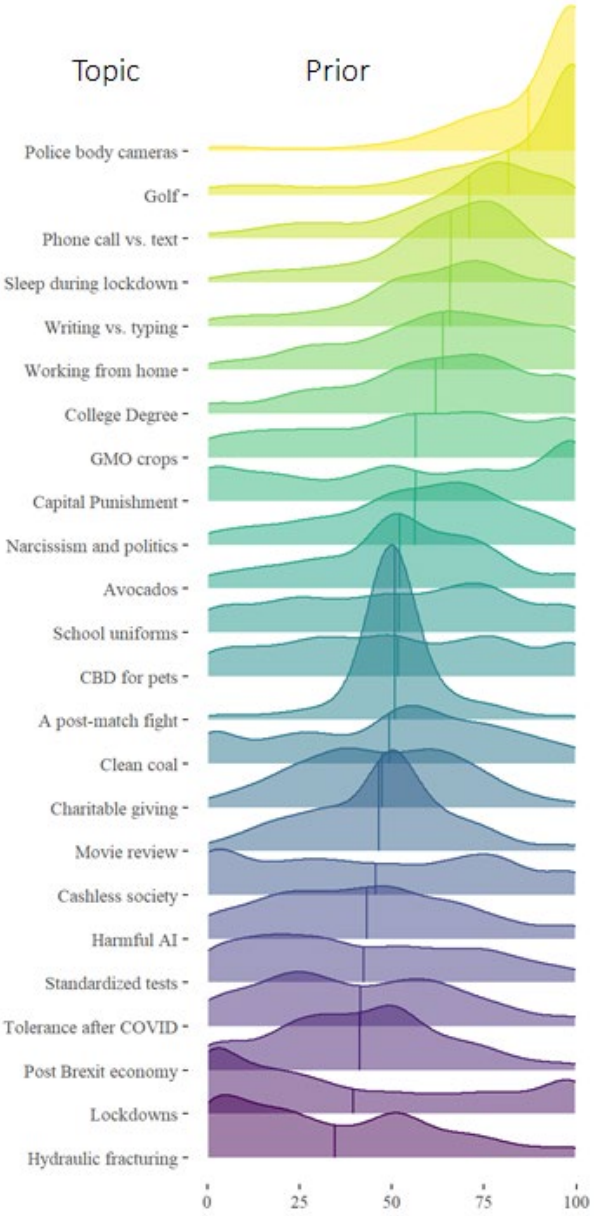
3.2 Trial Duration

The duration (seconds) for each participant to complete each trial was examined for each INFORMATION LEVEL condition to help assess whether participants were actually reading the tweets for each trial. An ANOVA showed that participants varied in the time it took to complete each trial, depending on which condition they were in. As some trial durations were very long (e.g., where people switched their attention away from the experiment) and others were very short, the trimmed "interquartile" mean (IQM) was used because this is less sensitive to outliers than the mean. Results showed a significant difference between the three groups ($p < .001$). Participants in the TWEETS ONLY condition had an IQM of 50 seconds (SD = .55 seconds) per trial. In the DIAGRAM WITH TWEETS condition, the IQM was 48 seconds

(SD = .63 seconds) per trial. Participants in the DIAGRAM ONLY condition had an IQM of 30 seconds (SD = .38 seconds) per trial. Pairwise comparisons between DIAGRAM ONLY and TWEETS ONLY conditions; and DIAGRAM ONLY and DIAGRAM WITH TWEETS conditions confirmed that the differences were significant ($p < .001$). In contrast, the DIAGRAM WITH TWEETS and TWEETS ONLY conditions were not significantly different ($p = .31$). Overall, these results suggest that participants were reading the tweets in the relevant conditions (TWEETS ONLY and DIAGRAM WITH TWEETS).

Figure 9

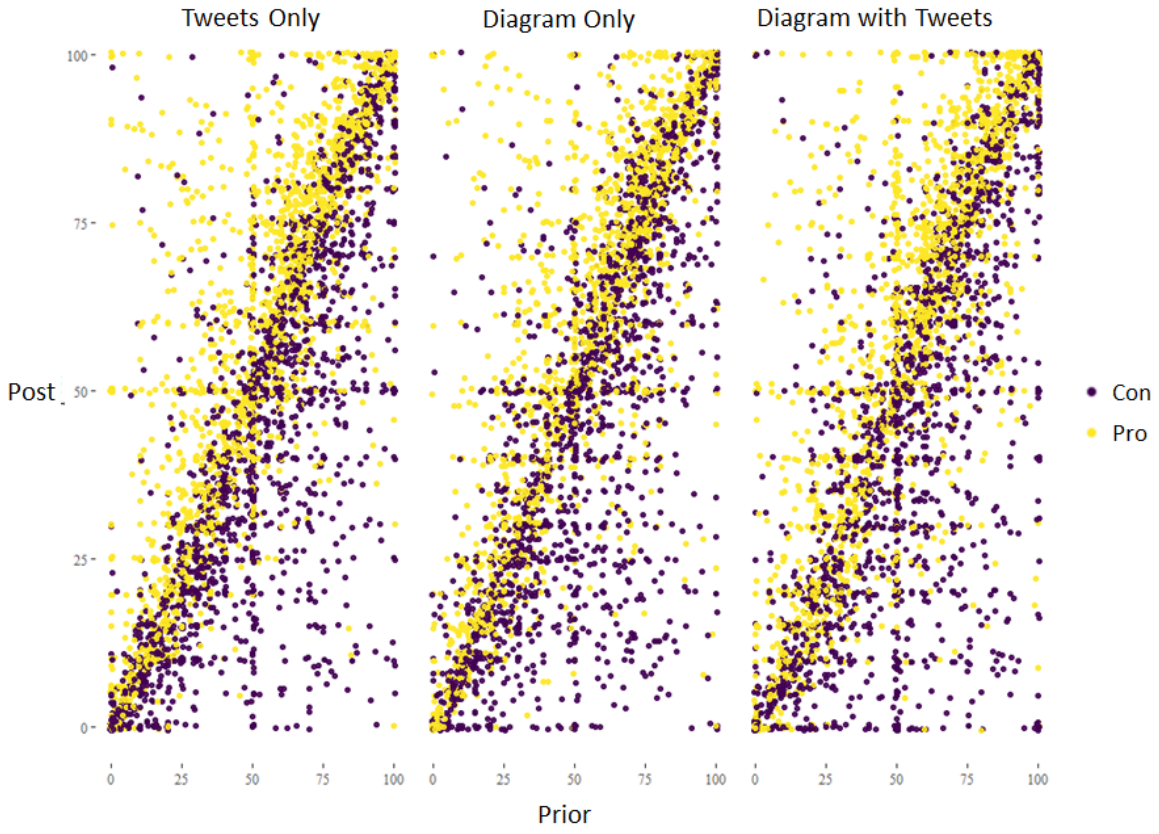
Prior Distribution of Agreement Levels



Note. A linear model was conducted to determine whether there was any significant difference in the prior ratings across INFORMATION LEVEL conditions, with no significant differences ($p > .05$).

Figure 10

Scatterplot of Prior and Post Ratings for Pro and Con Tweets for each INFORMATION LEVEL



To assess the extent to which people updated their beliefs about a claim on the basis of the tweets and/or diagrams presented, the difference (delta) between people's prior and post ratings on a trial-by-trial basis was calculated.[1]

In the TWEETS ONLY condition (see Figure 11), participants had the largest change in agreement rating in the FULL CONSENSUS condition, followed by in the MAJORITY CONSENSUS condition. When the *Target Tweets* were matched numerically by *Opposing Tweets* in the CONTESTED CONSENSUS condition, there was very limited change in the participants' updated

[1] The difference between prior and post ratings (delta) was collapsed across *Pro* and *Con* target tweet trials, to examine the change in ratings after viewing the diagrams/tweets. The sign of delta was adjusted so that the interpretation of a positive value is consistent across *Pro* and *Con* tweets.

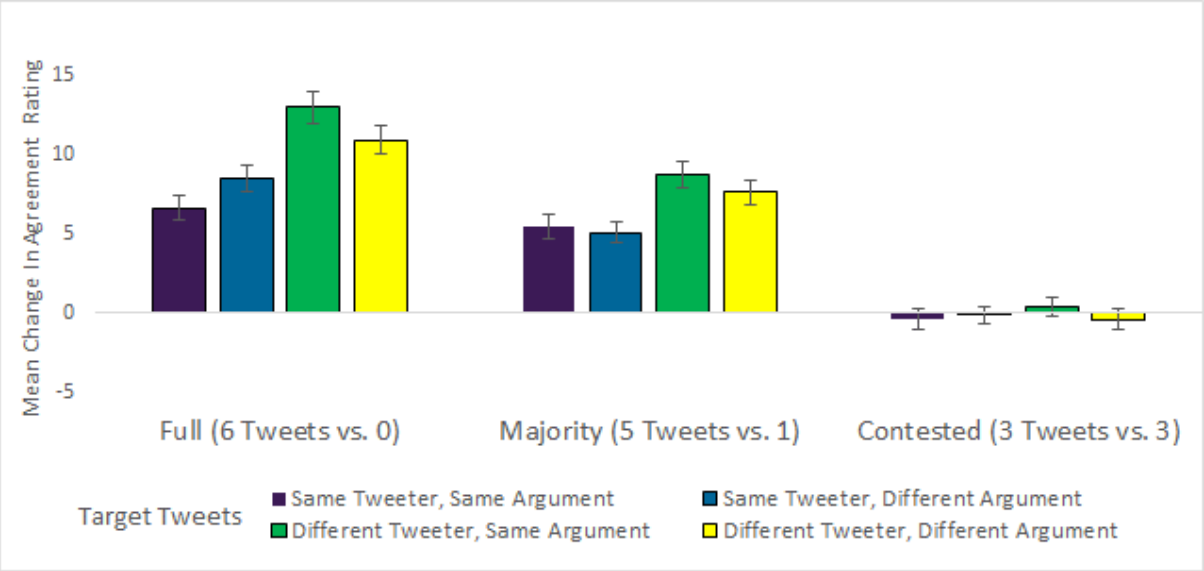
agreement ratings, even though the *Opposing Tweets* were always written by DIFFERENT TWEETERS and used DIVERSE ARGUMENTS, whilst the *Target Tweets* differed in diversity. Overall, people tended to have a higher change in agreement ratings when the tweets were written by DIFFERENT TWEETERS rather than the SAME TWEETER.

In the DIAGRAM ONLY condition (see Figure 12), the largest change in agreement ratings occurred in the FULL and MAJORITY CONSENSUS when the tweets were written by DIFFERENT TWEETERS. In the CONTESTED CONSENSUS condition, people appeared to reverse their agreement rating, indicating they were sensitive to cues of ARGUMENT DIVERSITY and AUTHOR DIVERSITY, even when there was an even quantity of tweets (this is because the *Opposing Tweets* were always diverse). This also suggests that in this DIAGRAM ONLY condition people relied less on the NUMBER OF TWEETS as a cue to revise their agreement.

In the DIAGRAM WITH TWEETS condition (see Figure 13), the changes in agreement ratings were quite similar to that within the DIAGRAM ONLY condition. People had the largest change in agreement ratings in the FULL and MAJORITY CONSENSUS when the tweets were written by DIFFERENT TWEETERS. When there were an equal number of tweets for *Pro* and *Con*, there was sensitivity to AUTHOR DIVERSITY and ARGUMENT DIVERSITY (as was also seen in the DIAGRAM ONLY condition), meaning there was a reduced effect of NUMBER OF TWEETS. This suggests that the diagram is having an effect as when there is the addition of a diagram people seem to be more sensitive to cues of argument quality (i.e., ARGUMENT DIVERSITY and AUTHOR DIVERSITY).

Figure 11

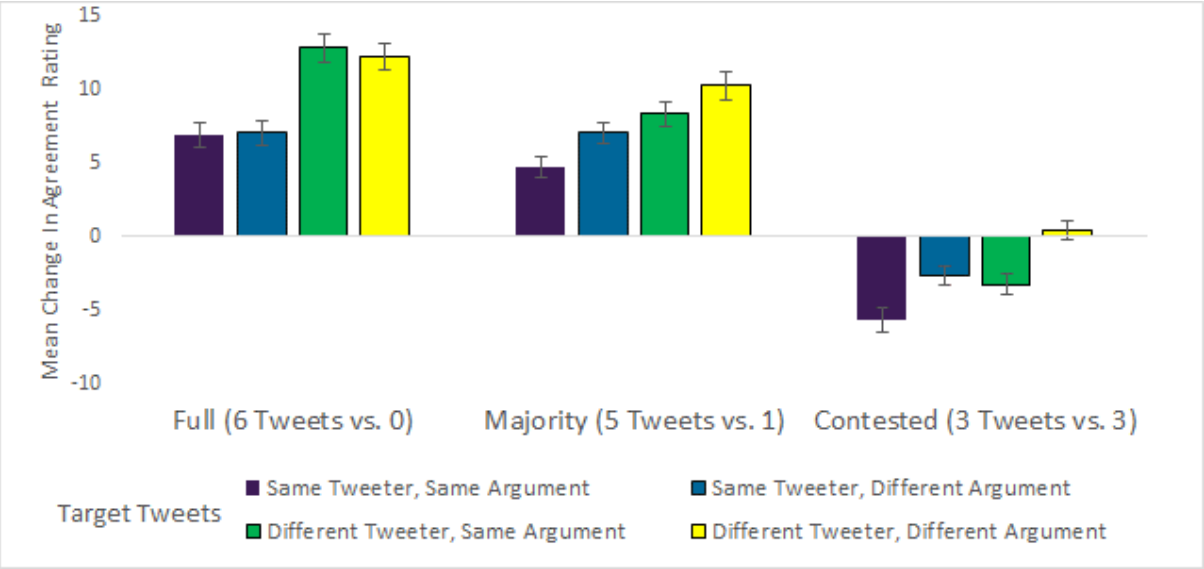
Change in Agreement Ratings for each Variable in TWEETS ONLY Condition



Note. The bars represent mean ± SE.

Figure 12

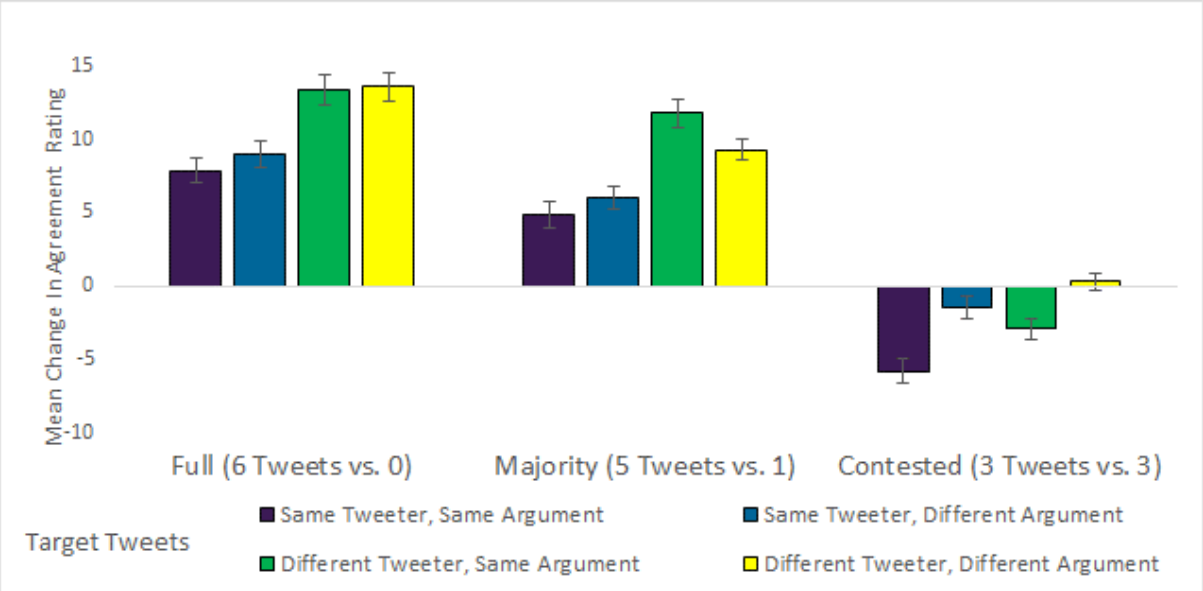
Change in Agreement Ratings for each Variable in the DIAGRAM ONLY Condition



Note. The bars represent mean ± SE.

Figure 13

Change in Agreement Ratings for each Variable in the DIAGRAM WITH TWEETS Condition

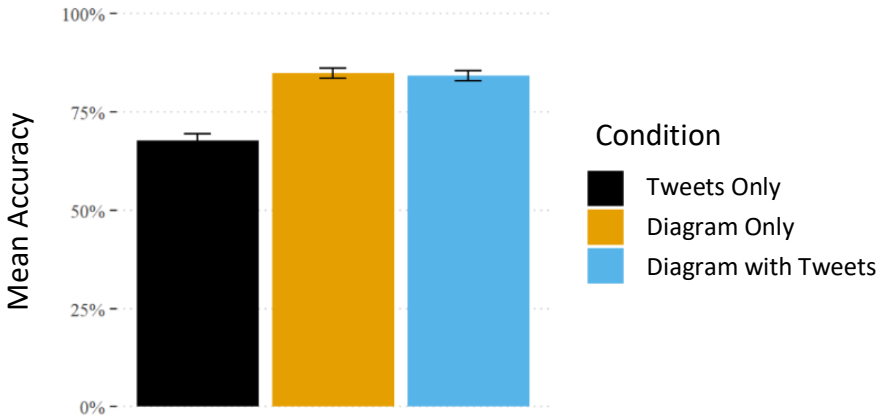


Note. The bars represent mean ± SE.

Taken together, Figures 12-13 suggest that people were more sensitive to cues of consensus quality when the diagrams made these cues more salient. An analysis of response accuracy to the follow-up questions quizzing participants on their recollection of the number of people, tweets and diversity of the final four trials supports this. To quantify the strength of evidence for an effect of condition on accuracy a logistic regression was performed. People were significantly more accurate in the DIAGRAM ONLY and DIAGRAM WITH TWEETS conditions compared with the TWEETS ONLY condition ($p < .001$; see Figure 14).

Figure 14

Mean Accuracy of Attention Check Questions by INFORMATION LEVEL Conditions



3.3 Linear Regression Models

To quantitatively examine the strength of these findings, five nested linear models were compared, with the dependent variable of updated (post) rating. All linear models were fit within R (Version 4.0.3, R Core) using the built-in function lm from the 'lme4' package. Assumptions for linear regression were examined and were met. The first model is a baseline that assumes people's agreement ratings are a function of their prior beliefs and this model accounted for 68.2% of variance, $F(1, 14422) = 30880$, $p < .001$. The second model adds the predictor NUMBER OF TWEETS to see whether people revise their agreement based on the quantity of tweets they view on each side of a claim, and this model accounted for 73.8% of variance, $F(2, 14421) = 20{,}260$, $p < .001$. The third model adds AUTHOR DIVERSITY as a predictor to explore whether people are also sensitive to whether different people post to Twitter, rather than one single person tweeting many times, and this model accounted for 74.2% of the variance, $F(3, 14420) = 13{,}830$, $p < .001$. The fourth model assumes that people will be additionally sensitive to ARGUMENT DIVERSITY and will revise their agreement depending on whether there are a range of arguments used or the same argument is repeated,

and this model also accounted for 74.2% of the variance, $F(4, 14419) = 10{,}380$, $p < .001$. Finally, the fifth model explores whether there is an interaction between the INFORMATION LEVEL and the three predictors, NUMBER OF TWEETS, AUTHOR DIVERSITY and ARGUMENT DIVERSITY, and this model accounted for 74.3% of the variance, $F(10, 14413) = 4164$, $p < .001$. This was the preferred model as this model captures the data better and a summary of the results from all models are shown in Table 4.

As Table 4 shows, all predictors including the interaction terms were significant except for the TWEETS ONLY condition interaction with ARGUMENT DIVERSITY (Argument: Tweets).

The regression analysis revealed an important qualitative reversal in the emphasis that people place on various cues. In the TWEETS ONLY condition, the analysis revealed that the quantity of tweets on each side of the argument had the biggest effect (β = 1.35), with the number of tweeters next (β = .68). This effect reverses in the diagram conditions. In the DIAGRAM ONLY condition, the effect of the quantity of tweets reversed (β = -.47) and sensitivity to the diversity of authors (β = .49) and the diversity of arguments (β = .49) increases. A similar reversal of regression coefficients was also seen in the DIAGRAM WITH TWEETS condition: for the quantity of tweets (β = -.31), the diversity of authors (β = .53) and the diversity of arguments (β = .45).

To explore the size of the effect within Model 5, the "predict" R function was used to hypothetically compare two scenarios in all three information levels: 1) if there were ten more tweets whilst the number of tweeters and arguments were evenly balanced in the regression model; 2) if there were ten more tweeters and ten more arguments whilst the number of tweets, tweeters and arguments were evenly balanced in the regression model.

The model predicts that an additional ten tweets leads to a change in agreement rating of 13.54 in the TWEETS ONLY condition compared to 8.81 in the DIAGRAM ONLY condition and

10.43 in the DIAGRAM WITH TWEETS condition. Conversely, the model predicts that an additional ten tweeters and ten arguments leads to a change in agreement rating of only 5.45 in the TWEETS ONLY condition compared to 15.24 in the DIAGRAM ONLY condition and 15.21 in the DIAGRAM WITH TWEETS condition.

The first example highlights that participants in the TWEETS ONLY condition weighted the NUMBER OF TWEETS as an important cue for their agreement revision, whereas this cue had less importance in both diagram conditions. The second example shows the extra value that additional tweeters and arguments has in the diagrams conditions compared to the control condition (TWEETS ONLY).

Table 4

Nested Regression models

	R^2	*B*	*SE B*	*t*
Model 1	.682			
Constant		6.29	.31	20.56***
Prior		.87	.01	175.72***
Model 2	.738			
Constant		6.55	.28	23.59***
Prior		0.86	.01	192.40***
No. Tweets		1.76	.03	55.42***
Model 3	.742			
Constant		6.58	.28	23.92***
Prior		.86	.01	193.96***
No. Tweets		1.19	.05	24.76***
Author		1.03	.07	15.95***
Model 4	.742			
Constant		6.59	.27	23.95***
Prior		0.86	.01	193.98***
No. Tweets		1.02	.07	18.88***
Author		0.69	.11	15.66***
Argument		0.17	.07	2.64**
Model 5	.743			
Constant		6.60	.28	24.01***
Prior		.86	.10	194.09***
No. Tweets: Tweets		1.35	.10	13.69***
No. Tweets: Diagram		-.47	.14	-3.34***

No. Tweets: Diagram+	-.31	.14	-2.19*
Author:	.68	.11	6.20***
Tweets			
Author: Diagram	.49	.16	3.07**
Author: Diagram+	.53	.16	3.33***
Argument: Tweets	-.14	.11	-1.23
Argument: Diagram	.49	.16	3.07**
Argument: Diagram+	.45	.16	2.85**

Note. R^2 = explained variance. B = regression coefficient. SE B = standard error of regression coefficient. 'prior' = an integer in the range 0-100 indicating the first (prior) agreement rating. 'No. Tweets' = Number of Tweets. 'Author' = Author Diversity. 'Argument' = Argument Diversity. 'Tweets' = Tweets Only Condition. 'Diagram' = Diagram Only Condition. 'Diagram+' = Diagram with Tweets Condition.

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

4. Discussion

The current study aimed to explore whether including a diagram summarising the NUMBER OF TWEETS, AUTHOR DIVERSITY and ARGUMENT DIVERSITY would affect agreement revision relative to the control condition with no diagram. On the basis of the previous work by Ransom et al. (2021), it was apparent that without a reasoning aid, people struggle to utilise information signalling argument quality. This study replicated this finding in the control condition, TWEETS ONLY, where the biggest effect was of quantity: the NUMBER OF TWEETS on either side of a claim. In contrast, as hypothesised, including a reasoning aid reduced the overall effect of the NUMBER OF TWEETS and increased the difference in the persuasiveness of DIVERSE vs NON-DIVERSE arguments. This finding occurred even in the DIAGRAM ONLY condition where the tweet contents were blurred out, supporting the diagram drove the effect. These results indicate that people do prefer DIVERSE TWEETERS and DIVERSE ARGUMENTS over simple quantity cues such as the NUMBER OF TWEETS; they might just be less able to consider cues of quality without the assistance of an aid such as the diagram.

4.1 TWEETS ONLY Condition

The TWEETS ONLY condition was a replication of the study by Ransom et al. (2021), with slight methodological changes including controlling the total number of tweets across the NUMBER OF TWEETS conditions. However, Ransom et al. (2021) did find a perhaps surprising, very slight effect where the same person repeating the same argument was, if anything, slightly more effective than when the same person used different arguments. The current study did not find such an effect, which is likely due to the fact that it was so slight to begin with. This study also helps eliminate the worries Ransom et al. (2021) had regarding a potential weak evidence effect. As their study had little to no effect of argument diversity, it was proposed that this could have been due to a methodological flaw where weaker

arguments were unintentionally included amongst stronger arguments, thereby signalling to participants that there are not many strong reasons to support the claim. However, the current study included additional diverse arguments (up to six rather than four), which would have only amplified the weak evidence effect if it existed – because creating more arguments increases the likelihood that one or more of these arguments are weaker than the others. This indicates that the limited effect of argument diversity was likely because people were not sensitive to ARGUMENT DIVERSITY without the assistance of an aid, which is made more salient in the conditions with diagrams, where effects of ARGUMENT DIVERSITY were found.

The TWEETS ONLY condition findings have added to the conflicting literature regarding people's sensitivity to cues of a quality consensus. The findings lend support to the studies by Ransom et al. (2021) and Yousif et al. (2019) that found people were not sensitive to cues of consensus or argument quality, such as whether people use different arguments to support a claim. The current study also supported Weaver (2007) and the effect of repetition. The largest effect was NUMBER OF TWEETS, but these tweets were sometimes all written by the SAME TWEETER. This means, for instance, that people still considered the same tweeter posting five *Pro* tweets versus one tweeter posting one *Con* tweet a consensus in favour of the *Pro* side, despite it only being an effect of repetition (compared to when different arguments were used).

4.2 Addition of a Diagram

Adding a diagram made people more sensitive to ARGUMENT DIVERSITY and AUTHOR DIVERSITY, with a reduced effect of the NUMBER OF TWEETS on either side of the claim. The fact that people were less sensitive to the NUMBER OF TWEETS when a diagram was present indicates that the diagram successfully promotes awareness of cues of argument quality. Cues about apparent consensus levels (such as the quantity of tweets supporting or refuting a

claim) do not seem to involve deep processing (Martin, 2002). Martin (2002) suggests that this may be because people want to belong to the majority group without actually processing the message, or they may be using a simple persuasion heuristic, such as that the majority is more likely to be correct than the minority. However, the diagram may have encouraged deeper processing of the tweets and cues of argument quality, which also reduced the effect of quantity cues.

The current study also builds upon the work by Yousif et al. (2019), including the troubling and consistent finding that people were unable to distinguish between a "true" and "false" consensus over five different experiments. What was especially unusual about the Yousif et al. (2019) findings, was in Experiment 4 participants were explicitly asked to rate whether they favoured a news article that used unique sources of information (true consensus) or an article that cited the same source (false consensus). However, even the participants that rated that they preferred the news article with unique primary sources still fell prey to the false consensus in the actual experiment. In contrast, the current study suggests that people do prefer a true consensus (when there are multiple people supporting the claim) to a false consensus (when the same person posts multiple times), but are lacking sensitivity to this cue if it is not highlighted for them during the reasoning task. This is likely why there was a major effect of argument quantity in the TWEETS ONLY condition, but strong effects of argument quality with reduced effects of quantity when a diagram was present. A reasoning aid seemed to help draw people's attention to whether there were multiple tweeters and whether they used a range of different arguments, rather than repetition of one tweeter or argument. In short, people are capable of incorporating these cues to consensus quality when they reason, but need support.

4.3 Comparing DIAGRAM ONLY and DIAGRAM WITH TWEETS Conditions

Surprisingly, there were similar effects observed in the DIAGRAM ONLY and DIAGRAM WITH TWEETS conditions. It was hypothesised that the DIAGRAM ONLY condition would have a much smaller effect of ARGUMENT DIVERSITY and AUTHOR DIVERSITY than the DIAGRAM WITH TWEETS condition as it was missing the contents of the tweets. The DIAGRAM ONLY condition displayed a diagram with up to six people and six post icons in it (along with the tweeter names and photos, etc.); in contrast the DIAGRAM WITH TWEETS condition had more information participants could utilise, as they were able to see both the contents of the tweets and the diagram summarising some of this information (stance and diversity of arguments). There are a few potential reasons why these conditions had very similar results. It could be due to the diagram training module shown prior to the experiment for both conditions. The training showed participants the diagrams and the tweets, highlighting specifically how these two features are interconnected. Participants in the DIAGRAM ONLY condition thus had a sense of the ARGUMENT DIVERSITY that was captured by the diagrams. Participants may have felt confident enough in how the diagram operated to not need to view the contents of any tweets (i.e., they knew they had all the information they needed). Another potential explanation could be that in the DIAGRAM WITH TWEETS condition, the participants simply did not utilise or read the tweets, as the diagram was displaying much of the important information found in the tweets anyway. However, further analyses showed that this is likely not the case as participants in the DIAGRAM WITH TWEETS and TWEETS ONLY conditions took significantly longer time to complete each trial than the DIAGRAM ONLY condition, indicating that participants were at least reading the tweets when they were available.

4.4 Diversity of Tweeters and Arguments

The finding that only showing a diagram was enough to alter participant's agreement levels with a claim directly shows that people understand the value of higher argument quality, such as a "true" consensus based on the number of authors and when there are diverse arguments. The category-based induction literature has shown that people seem to follow a diversity principle where diversity amongst the stimuli is more compelling evidence of conclusions (Ranganath et al., 2010). However, the category-based induction literature only seems to highlight this sensitivity indirectly as it does not explore whether people are explicitly aware of a connection between diverse arguments and argument strength (it only shows people two sets of arguments and asks them which set better supports the conclusion). The current study has shown through the DIAGRAM ONLY condition that people are aware of the value of diverse arguments. When the only information accessible to people were icons signalling how many posts there were, how many people were posting and whether the arguments were diverse, people were more compelled by DIVERSE ARGUMENTS and DIFFERENT TWEETERS. This finding demonstrates that people are in fact aware that DIVERSE ARGUMENTS is a stronger cue of argument strength or quality. Comparisons to the category-based induction literature also suggest that people may need more help in utilising this cue for more complex arguments - perhaps it takes extra effort to extract the more high-dimensional diversity when reading the tweets. It is important to note that the category-based induction literature's conception of the diversity principle is relatively different to the argument diversity discussed in the current study. Diversity in category-based induction explores how well the exemplars cover a category (i.e., how adding different premise examples strengthens the conclusion). However, the current study explores how a diversity of *reasons* strengthens agreement with a more complex claim. Nonetheless, it is interesting to report a similar finding that diversity is compelling in the different domains.

4.5 Strengths of the Current Study

The successful replication of Ransom et al. (2021) is one fundamental strength of this study. As previously mentioned, the current study has some methodological improvements, but still replicated the key original findings. This indicates that these results are relatively reliable. The methodological improvements are also a strength of the current study; I ensured participants spent the same amount of reading time in all conditions that displayed tweets, by having an equal total number of tweets in the FULL, MAJORITY and CONTESTED CONSENSUS. Another methodological improvement from Ransom et al. (2021) was the manipulation of NUMBER OF TWEETS within subjects. This also kept the reading time and effort consistent as all participants experienced an equal assortment of FULL, MAJORITY and CONTESTED CONSENSUS trials throughout the study.

The current study built upon the work of Ransom et al. (2021) through novel research on including a reasoning aid. To my knowledge, this kind of aid has not been used in a complex reasoning domain such as the current study. The reasoning aids were based on careful design features such as using distance to convey diversity (Ajjour et al., 2017; Hout et al., 2015). Furthermore, there are a wide variety of topics across the different claims in the study, which strengthens the generalisability of the findings; the results cannot just be due to one particular claim being important to participants, or particular domain knowledge. All claims also had randomised *Pro* and *Con* target arguments, and there was thorough randomisation of many other aspects of the stimuli (e.g., tweeter names and photos) and presentation order.

4.6 Applied Implications

The findings of the current study have significant real-world implications. It is clear that without support, people do not consistently utilise cues of argument or consensus quality

— at least for the kind of claims included in the study. This is especially concerning in an era where social media is a dominant source of information for many people. Social media has made it even more challenging to reason rigorously, with a constant flow of new incoming information, numerous contradicting opinions, visual cues about social support (such as "likes") and an abundance of misleading information. Concerningly, the visual cues about support for information are often distorted and not truly representative on social media, because they are affected by bots or over-active minority opinions (Lee et al., 2021). Algorithms only further perpetuate this issue through prioritising certain content, making it appear more salient and widely accepted than the content may actually be (Zimmer et al, 2019). It is clear that people need assistance in reasoning in a way that is better calibrated with the available evidence on social media.

Unfortunately, the current reasoning tools in place to help people reason more effectively on social media (fact-checkers) can be ineffective and even counterproductive (Allen et al., 2021). The reasoning aid in the current study has shown promising potential in assisting people to utilise cues of consensus quality. As this reasoning aid uses the wisdom of the crowd approach, it avoids many of the issues associated with the currently used fact-checkers. If the diagrams could be successfully automated, the aid could be easily utilised as it uses crowd-sourced opinions and does not need a third-party to individually assess every questionable claim on social media. Crowd-sourced judgements can be extremely accurate and are also viewed as trustworthy as they utilise the general public rather than specific agencies (Collins et al., 2021; Simoiu et al., 2019). This type of reasoning aid could be an important reasoning intervention for people when they are using social media platforms. The diagrams could augment or be used in conjunction with other automated fact checking tools that are being developed by computer scientists (see Shu et al., 2017).

4.7 Limitations of the Current Study

The current study was not without its limitations. All of the tweets were written by the research team and designed to be fairly consistent in plausibility and writing style, but may have had weaknesses. The manipulation of DIVERSE versus NON-DIVERSE arguments may have been vulnerable to human error as some of the claims may have had NON-DIVERSE arguments that were not similar enough, or DIVERSE ARGUMENTS that were not diverse enough. The actual saliency of the repetition or diversity of tweets may have influenced participants' sensitivity for particular claims (i.e., if tweets were not diverse enough for a claim, participants may have been less sensitive to ARGUMENT DIVERSITY as a cue of consensus quality). Another issue was that because of the sheer size of the stimulus set, the NON-DIVERSE repeated argument tweets were not fully randomised. There were not repeated-argument tweets for all six possible diverse arguments; only one of the DIVERSE ARGUMENTS was selected and repeated another five times (eg. see Table 1-2). Thus, the repeated arguments may vary in how compelling they are relative to the other five arguments within a tweet set, which could either reduce or enhance any effect of argument diversity.

All tweets for a topic were presented together; however, in real life, information is not necessarily encountered all together and is generally accrued over time. Post ratings were also made immediately after viewing all of the information, so it is unclear whether the shifts in belief would be enduring for a longer timeframe. The fact that the study required participants to provide an updated rating could have also led to demand effects; participants knew there was an expectation for them to have a change in rating. Therefore, real belief revision may be smaller or require more evidence.

The demographic data also highlights that there was a political bias skewed towards Liberal orientation. This may limit generalisability, especially as research shows that people utilise quantity cues or quality cues of consensus depending on the relevance of the

information to them. When people read claims that are irrelevant to them, the number of arguments, but not the quality of the arguments, is most important (Petty & Cacioppo, 1984). When people read claims relevant to them, they are influenced by the number of arguments, but also the quality of these arguments; simply adding numerous weak arguments does not increase persuasiveness (Petty & Cacioppo, 1984). However, the variety of the topics, along with the fact that the target stances on the topics randomly varied in whether they were in favour of or against the claim, would limit any effect of demographic skew in political identity.

4.8 Future Directions

Firstly, it is essential to address the importance of replication in Psychology, especially with novel research. The current study needs to be repeated as, from my knowledge, there is no other study that uses a diagram representing consensus quality cues to help people reason. Even though the findings of this study were extremely promising, the results still need to be verified through replication.

There are many outstanding research questions for future studies to explore. It would be interesting to see the effect of scaling the reasoning aid. To do this, the people and tweets represented in the diagram could be scaled to, say, 100 and be presented with a sample of tweets that represent the overall consensus quantity and quality information. This would slightly shift the purpose of the diagram into a tool that has a more comprehensive representation of the overall consensus information, not just mirroring the information from the small sample of tweets. The sample of tweets would then reinforce the information found in the diagrams and perhaps allow people to draw stronger conclusions, calibrated with the available evidence. A study exploring the scaling would advance the current study's findings on the usefulness of a reasoning tool where people had even more information accessible to

reason with (i.e., there are now 100 people's agreement levels with the claim to consider, not just six).

There is also potential scope to include other cues of consensus quality in the reasoning aid that were not addressed in the current study. For example, it may be important to explore the role of expert opinion, which elicits more processing of persuasive messages (Clark et al., 2011). It is widely acknowledged that people expect expert sources to provide high-quality, valid information (Clark et al., 2011). It would be interesting to see how differentiating between lay-person and expert sources would impact the perception of the reasoning aid: would it negate the trust in the wisdom of the crowds approach that only uses lay-people's opinions? It could also be useful to highlight distinctions between the similarity of the people, which could be another important cue of consensus quality. For instance, if all of the people represented in the diagram are similar, such as the same gender, political affiliation, or ethnicity, the represented consensus may not be good quality or very reliable. Researchers could manipulate the diversity of the features of the sources to explore whether people are not only sensitive to whether different people are sources of information, but whether these people also have diverse features or qualities. Effects of source-similarity could also be investigated through highlighting certain features of the social network structure. For example, a diagram could depict how people are connected on social media. If people are following each other on Twitter or in the same Facebook "groups", they might not be truly independent sources of information. Furthermore, future studies could expand on the effect of repeated arguments through showing the repeated arguments as shares or 'retweets' of identical posts, rather than as reworded posts.

4.9 Conclusion

The findings from this study suggest that people readily rely on cues of consensus quantity and are not inherently sensitive to cues of consensus quality unless they receive assistance from a reasoning aid. Without an aid, people focused on how many tweets they saw supporting or refuting a claim. This is not necessarily a strong cue of consensus quality, as these tweets may have been written by the same person spamming, or repeat similar evidence; just because there is a large quantity of consistent tweets giving the impression of a consensus does not mean this is a truly compelling consensus with diverse reasoners and diverse original sources. When the reasoning aid was presented with the tweets, people became more sensitive to the cues of consensus quality and the influence of quantity was reduced, suggesting the aid is an effective reasoning tool. Future work should develop the reasoning aid to have a real-world application, where it helps people navigate information on social media.

Reference List

Ajjour, Y., Wachsmuth, H., Kiesel, D., Riehmann, P., Fan, F., Castiglia, G., Adejoh, R., Fröhlich, B., & Stein, B. (2018). Visualization of the topic space of argument search results in args.me. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 60-65.

Allen, J. N. L., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Journal of Science Advances*, *7*(36), 1-10. https://doi.org/10.1126/sciadv.abf4393

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1-70. https://doi.org/10.1037/h0093718

Blutinger, E. J., Shahid, S., Jarou, Z. J., Schneider, S. M., Kang, C. S., & Rosenberg, M. (2021). Translating COVID-19 knowledge to practice: enhancing emergency medicine using the "wisdom of crowds". *Journal of the American College of Emergency Physicians Open*, *2*(1), 1-8. https://doi.org/10.1002/emp2.12356

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), 1-33. https://doi.org/10.1371/journal.pone.0239666

Cacioppo, J. T., & Petty, R. E. (1979). Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of Personality and Social Psychology, 37*(1), 97-109. https://doi.org/10.1037/0022-3514.37.1.97

Calder, B. J., Insko, C. A., & Yandell, B. (1974). The relation of cognitive and memorial processes to persuasion in a simulated jury trial. *Journal of Applied Social Psychology*, *4*(1), 62-93. https://doi.org/10.1111/j.1559-1816.1974.tb02808.x

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current* biology, *20*(13), 1165-1170. https://doi.org/10.1016/j.cub.2010.04.055

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in combating fake news on social media − a survey. *Journal of Information and Telecommunication*, *5*(2), 247-266. https://doi.org/10.1080/24751839.2020.1847379

Coultas, J. C. (2004). When in Rome... an evolutionary perspective on conformity. *Group Processes & Intergroup Relations*, *7*(4), 317-331. https://doi.org/10.1177/1368430204046141

Dowden, B. H. (2017). *Logical Reasoning*. Wadsworth Publishing Company.

Friedrich, J., & Smith, P. (1998). Suppressive influence of weak arguments in mixed-quality messages: an exploration of mechanisms via argument rating, pretesting, and order effects. *Basic and Applied Social Psychology*, *20*(4), 293-304. https://doi.org/10.1207/s15324834basp2004_6

Ganser, C., & Keuschnigg, M. (2018). Social influence strengthens crowd wisdom under voting. *Advances in Complex Systems*, *21*(6-7), 1-24. https://doi.org/10.1142/s0219525918500133

Goldstone, R. L., & Son, J. Y. (2012). Similarity. *Psychology Review*, *100*, 13-36. https://doi.org/10.1093/oxfordhb/9780199734689.013.0010

Gunaratne, C., Baral, N., Rand, W., Garibay, I., Jayalath, C., & Senevirathna, C. (2020). The effect of information overload on online conversation dynamics. *Computational and Mathematical Organization Theory*, *26*(5), 255-276. https://doi.org/10.1007/s10588-020-09314-9

Harkins, S. G., Petty, R. E. (1981). Effects of source magnification of cognitive effort on attitudes: An information-processing view. *Journal of Personality and Social Psychology, 40*(3), 401-413. https://doi.org/10.1037/0022-3514.40.3.401

Harris, A., Sildmäe, O., Speekenbrink, M., & Hahn, U. (2019). The potential power of experience in communications of expert consensus levels. *Journal of Risk Research*, *22*(5), 593-609.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *WIREs Cognitive Science, 9*(3), 13. https://doi.org/10.1002/wcs.1459

Heit, E. (2000). Features of similarity and category-based induction. *Proceedings of the Interdisciplinary Workshop on Categorization and Similarity.* 115-121. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.7764&rep=rep1&type=pdf

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology*. *Learning, Memory, and Cognition*, *20*(2), 411-422. https://doi.org/10.1037/0278-7393.29.2.411

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Journal of Cognitive* Psychology, *111*, 80-102. https://doi.org/10.1016/j.cogpsych.2019.03.001

Hout, M. C., Godwin, H. J., Fitzsimmons, G., Robbins, A., Menneer, T., & Goldinger, S. D. (2016). Using multidimensional scaling to quantify similarity in visual search and beyond. *Attention, Perception, & Psychophysics*, *78*(1), 3–20. https://doi.org/10.3758/s13414-015-1010-6

Innes, M., Dobreva, D., & Innes, H. (2021). Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing. *Contemporary Social Science*, *16*(2), 241-255. https://doi.org/10.1080/21582041.2019.1569714

Kary, A., Newell, B., & Hayes, B. (2018). What makes for compelling science? Evidential diversity in the evaluation of scientific arguments. *Global Environmental Change*, *49*, 186-196. https://doi.org/10.1016/j.gloenvcha.2018.01.004

Kemp, C., Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychology Review*, *116*(1), 20-58. https://doi.org/10.1037/a0014282

Koch, T., Frischlich, L., & Lermer, E. (2021). The effects of warning labels and social endorsement cues on credibility perceptions of and engagement intentions with fake news. *PsyArXiv*, 1-35. https://doi.org/10.31234/osf.io/fw3zq

Larsen, K. S. (1990). The Asch conformity experiment: replication and transhistorical comparisons. *Journal of Social Behavior & Personality*, *5*(4), 163-168.

Lee, S., Liang, F., Hahn, L., Lane, D. S., Weeks, B. E., & Kwak, N. (2021). The Impact of social endorsement cues and manipulability concerns on perceptions of news credibility. *Cyberpsychology, Behavior and Social Networking*, *24*(6), 384-389. https://doi.org/10.1089/cyber.2020.0566

Liu, Y., Chen, F., Kong, W., Yu, H., Zhang, M., Ma, S., & Ru, L. (2012). Identifying web spam with the wisdom of the crowds. *ACM Transactions on the Web*, *6*(1), 1-30. https://doi.org/10.1145/2109205.2109207

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national Academy of Sciences of the United States of America*, *108*(22), 9020-9025. https://doi.org/10.1073/pnas.1008636108

Martin, R., Gardikiotis, A., & Hewstone, M. (2002). Levels of consensus and majority and minority influence. *European Journal of Social Psychology, 32*(5), 645-665. https://doi.org/10.1002/ejsp.113

Medlin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517-532. https://doi.org/10.3758/BF03196515

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS One*, *12*(9), 1-12. https://doi.org/10.1371/journal.pone.0184148

Moore, T., & Clayton, R. (2008). Evaluating the wisdom of crowds in assessing phishing websites. *Financial Cryptography and Data Security*, *5143*, 16-30. https://doi.org/10.1007/978-3-540-85230-8_2

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*(2), 125-135. https://doi.org/10.1037/h0027568

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PloS one*, *8*(11), 1-8. https://doi.org/10.1371/journal.pone.0078433

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2), 185-200. https://doi.org/10.1037/0033-295X.97.2.185

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., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521-2526. https://doi.org/10.1073/pnas.1806781116

Perfors, A., Ransom, K. J., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2759-2764. https://perfors.net/files/2014/PerforsRansomNavarro2014.pdf

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46*(1), 69-81. https://doi.org/10.1037/0022-3514.46.1.69

Ranganath, K., Spellman, B., & Joy-Gaba, J. (2010). Cognitive "category-based induction" research and social "persuasion" research are each about what makes arguments believable: a tale of two literatures. *Perspectives on Psychological Science*, *5*(2), 115-122. https://doi.org/10.1177/1745691610361604

Ransom, K. J., Perfors, A., & Stephens, R. (2021). Social meta-inference and the evidentiary value of consensus. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 833-839.

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, *55*(2), 191-197. https://doi.org/10.1016/j.jmp.2010.10.002

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of* Sciences, *116*(16), 7662-7669. https://doi.org/10.1073/pnas.1805871115

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: Explainable fake news detection. *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 395-405. https://doi.org/10.1145/3292500.3330935

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

Sikder, O., Smith, R. E., Vivo, P., & Livan, G. (2020). A minimalistic model of bias, polarization and misinformation in social networks. *Scientific Reports, 10(1)*, 1-11. https://doi.org/10.1038/s41598-020-62085-w

Simoiu, C., Sumanth, C., Mysore, A., & Goel, S. (2019). Studying the "wisdom of crowds" at scale. *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, 7(1), 171-179.*

Spellman, B. A., López, A., Smith, E. E. (1999). Hypothesis testing: strategy selection for generalizing versus limiting hypotheses. *Thinking & Reasoning, 5(1), 67-91.* https://doi.org/10.1080/135467899394084

Tandoc, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2018). Audiences' acts of authentication in the age of fake news: a conceptual framework. *New Media and Society*, *20*(8), 2745-2763. https://doi.org/10/gc2fmd

Turiel, J., & Aste, T. (2020). *Wisdom of the crowds in forecasting COVID-19 spreading severity*. https://www.researchgate.net/publication/340523862_Wisdom_of_the_crowds_in_forecasting_COVID-19_spreading_severity

Wang, Y., Dai, Y., Li, H., & Song, L. (2021). Social media and attitude change: information booming promote or resist persuasion? *Frontiers in Psychology*, *12*, 1-12. https://doi.org/10.3389/fpsyg.2021.596071

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: a repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, *92*(5), 821-833. https://doi.org/10.1037/0022-3514.92.5.821

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*. *93*(1): 1-13. https://doi.org/10.1016/j.obhdp.2003.08.002

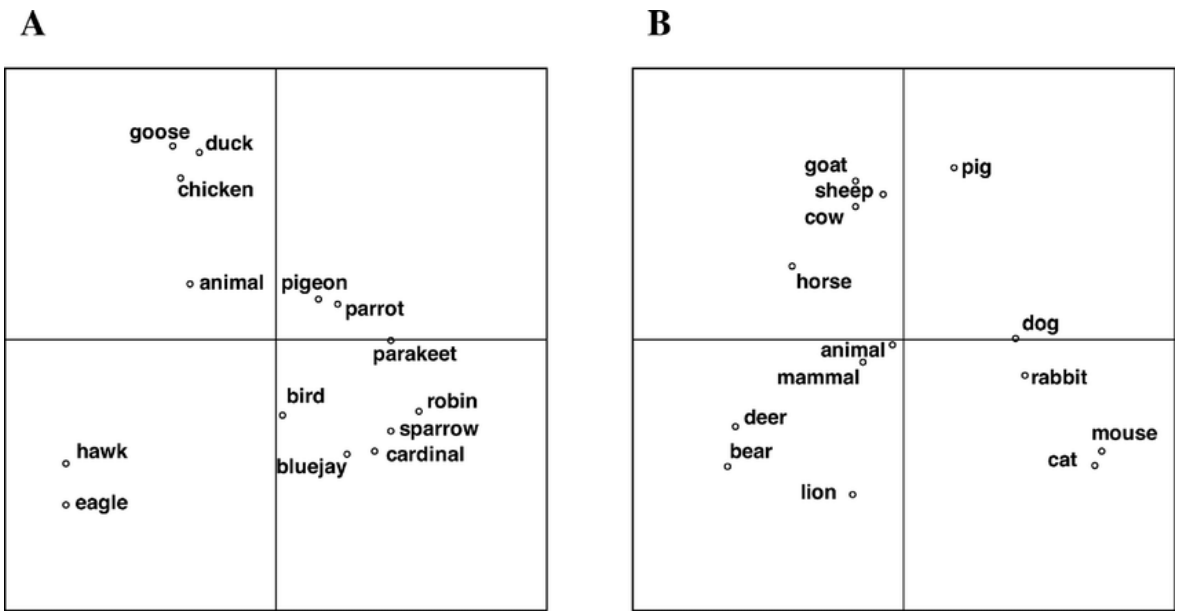
Yousif, S., Aboody, R., & Keil, F. (2019). The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, *30*(8), 1195-1204. https://doi.org/10.1177/0956797619856844

Zimmer, F., Scheibe, K., Stock, M., & Stock, W. (2019). Fake news in social media: bad algorithms or biased users? *Journal of Information Science Theory and Practice*, *7*(2), 40-53. https://doi.org/10.1633/JISTaP.2019.7.2.4

Appendix A

Figure A1

Multidimensional Scaling

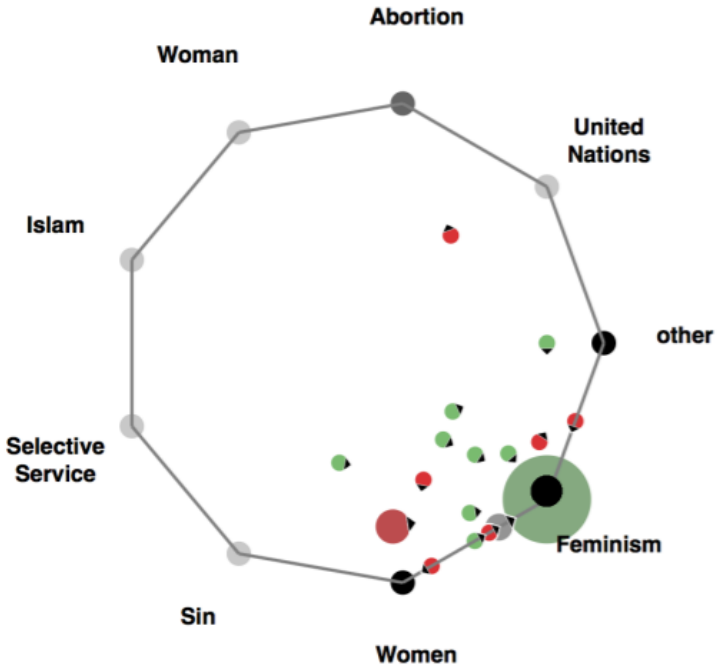


Note. Diagrams are by Goldstone and Son (2021). Two multidimensional scaling solutions for birds (A) and animals (B). Diversity of the birds and animals is illustrated through how far away they are from each other.

Appendix B

Figure B1

Topic Space Visualisation



Note. Diagram of topic space visualisation for the topic of "feminism" by Ajjour et al. (2017). The topic space of "feminism" is represented by the regular polygon and the argument topics surrounding feminism are shown on each vertex. The coloured dots represent specific pro and con arguments about feminism, with their spatial location in relation to the vertices symbolising their diversity.