Cognition 266 (2026) 106257

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

For everyday arguments prior beliefs play a larger role on perceived argument quality than argument quality itself

Calvin Deans-Browne [*], Henrik Singmann

Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

ARTICLE INFO

Keywords:
Argument quality
Argument evaluation
Everyday reasoning
Informal reasoning
Belief-based reasoning

ABSTRACT

Not all arguments are equally convincing, and whilst a given argument may be persuasive to some people, it is often seen as inadequate by others. We are interested in both the individual and argument level differences that make ‘everyday’ arguments such as those on social media persuasive. We investigate this question using a paradigm that consists of two parts. In the first part, we measure participants’ individual beliefs about eight claims each referring to a political topic (e.g., *Abortion should be legal*). In the second part, participants rated the quality of an argument for each of these claims. Arguments were good or bad (Experiments 1 and 2) or good, inconsistent, or authority-based (Experiment 3). Good, inconsistent, and authority-based arguments summarised arguments from an educational bipartisan website, contained internal inconsistencies, or were based on appeals to authority, respectively. We found that participants preferred arguments that were also in line with their beliefs. We also found that participants were able to discriminate the qualities of different arguments – good arguments were rated as better than any other type of argument. In Experiment 3, inconsistent arguments were rated as better than those making appeals to authority. Importantly, the maximum effect of belief was larger than the maximum effect of argument quality. Thus, people do not evaluate arguments independently of the background beliefs held about them, which play at least as large a role in evaluating the quality of the argument as does the actual quality of the argument itself.

1. Introduction

Media literacy is a skill of increasing importance as we are being confronted with information from an ever-growing number of media outlets. There is often no barrier to entry for people to give their opinion online, and the rapid proliferation of poor-quality information, including misinformation and disinformation, has been termed an ‘infodemic’ (Lewandowsky et al., 2022; Zarocostas, 2020). This has had far-reaching real-world consequences including impacts on mental health, misallocation of health resources, and vaccine hesitancy (Borges Do Nascimento et al., 2022). Adjacent in the political sphere, poor-quality information has also disrupted democracy, for example with fake news circulating around both candidates in the 2016 US Presidential election (Allcott & Gentzkow, 2017) encouraging people to vote on false information. From a psychological perspective, it is important to understand how people interpret the information they receive from media sources and integrate it into their belief system, and conversely how our belief system can influence how we reason about the information we receive.

Research on media literacy so far has mostly focused on two aspects, people’s ability to distinguish veridical from fake news and the type of information that people see. For example, Pennycook and Rand (Pennycook & Rand, 2019; see also Pennycook et al., 2023, 2022) found that the propensity to believe fake-news headlines was driven by a lack of analytical thinking, and others (e.g., Bakshy et al., 2015; Cinelli et al., 2021) have found that people are far more likely to encounter news and discussions on social media that are already aligned with their beliefs. In our research we are interested in a different facet of media literacy: how information is interpreted depending on its alignment with one’s beliefs. To study this issue, we focus particularly on information in the context of arguments, where information is commonly exchanged between people with different beliefs.

1.1. What makes an argument convincing?

From a purely rational perspective, the main criterion for the convincingness of an argument should be its quality. Traditionally, research on human reasoning has assumed that what determines the

[*] Corresponding author.
E-mail address: cdeansbrowne@gmail.com (C. Deans-Browne).

https://doi.org/10.1016/j.cognition.2025.106257
Received 26 March 2024; Received in revised form 10 July 2025; Accepted 13 July 2025
Available online 24 September 2025

quality of an argument is solely the form of the argument itself; that is, whether or not it is logically valid. However, this one-dimensional perspective on argument quality does not adequately reflect how people think about the quality of an argument (e.g., Evans, 2002). A more comprehensive perspective instead suggests that there are many factors related to the content of an argument that determine its perceived quality. For example, there is ample evidence to suggest that people perceive arguments based on statistical and causal (as opposed to anecdotal) evidence as being of higher quality (e.g., Hoeken, 2001) and more convincing (Hoeken & Hustinx, 2009; Slusher & Anderson, 1996).

While some factors appear to make arguments more convincing in general, there are also factors that make arguments which are convincing to some people unconvincing to others. For example, Edwards and Smith (1996) and Taber and Lodge (2006) both found that participants were more convinced by arguments that were more aligned with their prior beliefs over those that were not. Stanovich and West (1997) found similar results and additionally compared the magnitude of the effect of prior beliefs with the magnitude of the effect of argument quality (where the quality of each argument was determined by experts on a given topic). Their results showed that, whereas both prior beliefs and argument quality affected the convincingness of an argument, the magnitude of the effect of argument quality was greater. This pattern of results was replicated by Thompson et al. (2012). Thompson et al. also investigated whether there was an interaction between prior beliefs and argument quality (previously emphasised in an influential study by Evans et al., 1983, which used formal arguments as stimuli) and found only weak and inconsistent evidence for its existence.

In this study, we introduce the Everyday Argument Assessment Task which has the goal of disambiguating the effects of prior beliefs and argument quality for evaluations of everyday arguments. Our first research question concerns which of the two components has a larger effect on participants' perceptions of argument quality. To study this question we focus on *disputable* political beliefs – political beliefs that can vary greatly from one person to another and to which there is no one objectively 'correct' belief (e.g., the idiosyncratic beliefs in response to the claim *Abortions should be legal in the US*). By comparing how people with different beliefs respond to the same argument on contentious matters, we can measure the size of the effect of prior beliefs on argument evaluation. Furthermore, by looking at evaluations of arguments of differing quality where the arguments are in response to the same claims, we can also measure the size of the effect of argument quality and compare it with the size of the effect of people's prior beliefs.

1.2. *The everyday argument assessment task*

In three experiments participants saw everyday arguments about disputable political claims (e.g., *Abortions should be legal in the US*) and were tasked with evaluating the quality of the arguments. To determine the believability of the claims, we asked participants to rate the veracity of each claim on a 7-point scale ranging from *extremely false* to *extremely true*[1] (i.e., belief was a continuous independent variable that was measured and not manipulated). The claims presented to participants can be seen in Table 1.

Table 1
Claims and their corresponding topics shown to participants.

Topic (Pretest, Experiment 1, 2 and 3)	Claim (Pretest, Experiment 1, 2 and 3)	Alternative Claim (Experiment 3 Only)
Climate change	Human activity is primarily responsible for climate change.	Human activity is not primarily responsible for climate change.
Abortion	Abortions should be legal in the US.	Abortions should be illegal in the US.
Taking the knee	Kneeling during the national anthem is an appropriate form of protest.	Kneeling during the national anthem is an inappropriate form of protest.
Private prisons	Private prisons are not well run.	Private prisons are well run.
Cancel culture	Cancel culture is bad for society.	Cancel culture is good for society.
Fracking	It is in the United States' best interest to continue fracking.	It is in the United States' best interest to stop fracking.
Habitual offender laws	Habitual offender (or "three strike") laws are an appropriate way to punish reoffenders.	Habitual offender (or "three strike") laws are an inappropriate way to punish reoffenders.
Gun control	Further gun control laws are unnecessary.	Further gun control laws are necessary.
Affirmative Action (Experiment 3 Only)	Affirmative action leads to a less just society.	Affirmative action leads to a more just society.
Secularisation (Experiment 3 Only)	Separating church from state causes more harm than good.	Separating church from state causes more good than harm.

Note. Each claim was either left-aligned or right-aligned so that participants saw roughly the same number of claims they agreed and disagreed with. In Experiments 1 and 2, participants only saw the first eight topics and corresponding claims in the first and second columns of the table. The first four claims in the table were left aligned, and the second four were right aligned, and participants in Experiments 1 and 2 only ever saw these versions of the claims. In Experiment 3, we added new items (revolving around topics Affirmative Action and Secularisation) and changed the experiment so that each participant could see a left or a right aligned version of a claim for each topic that aligned with a left and right aligned version of the argument they would see subsequently.

We manipulated argument quality following an approach by Hopkins et al. (2016). We manipulated the informal evidential support of the arguments by varying how well the information presented in each argument was connected to the argument's central claim (i.e., its conclusion). 'Good' arguments contained evidence that provided strong support for the claim that was either statistical (e.g., *The United States' gun-related homicide rate is 25 times higher than the average of 22 other comparable high-income nations*) or causal (e.g., *When we heat our homes, power our cars, and run our factories, the emissions released cause our planet to warm*) in nature. Evidence for 'bad' arguments was substantially weaker, containing various flaws including circular reasoning (what were essentially restatements of the claim), appeals to authority, appeals to popularity, and appeals to tradition (see Table 2 for examples of good arguments with causal evidence, and bad arguments based on appeals to authority and tradition). In Experiment 3, we more systematically manipulate the 'bad' arguments with the introduction of 'inconsistent' and 'authority-based' arguments that are internally inconsistent and based on appeals to authority respectively (see *Experiment 3*, especially Table 3 below for example arguments).

To ensure participants understood our definition of argument quality, they were specifically asked to evaluate the arguments with regard to how well the information in the argument supported the argument's claim independent of their belief regarding the claim. An example of different good and bad arguments a participant could have seen for one claim is shown in Table 2; the complete set of all arguments used in the study can be found in the supplemental online materials on OSF.

As a manipulation check for our definition of good and bad arguments used in Experiments 1 and 2 we performed a pretest. In each trial

[1] We acknowledge that from a formal (e.g., logic or probability theory) point of view the usage of *extremely true* and *extremely false* appears questionable (as opposed to, for example, *extremely likely to be true* and *extremely unlikely to be true*). However, a graded use of truth as we use it in the current study seems very much in line with linguistic practices (Henderson, 2021). Given that our interest was in people's everyday beliefs and argument quality perception, we therefore decided to use this phrasing. Furthermore, none of the participants in any of the experiments reported here mentioned the anchors of the truth scale in their comments as a source of confusion (all participants were asked for any comments at the end of the experiment).

Table 2
Example pro and con arguments for claim *Abortions should be legal in the US*.

	Good argument	Bad argument
Pro argument	Abortions under Roe v. Wade balanced two fundamental rights; the right of the pregnant woman to bodily autonomy and the right of the unborn child to life. The unborn child only has the potential for life as we know it when they can survive outside the womb, and abortions had to occur before this stage under this ruling. Consequently, abortions can be consistent with both fundamental rights. Abortion should therefore be legal in the US.	It is important that abortion is legal, as it is a woman's right. Roe v. Wade declared abortion as a "fundamental right" and enshrined this in American law in 1973. This means that before Roe v. Wade was recently overturned, accessible abortion was legal for nearly 50 years. It is important that abortion is legal, as historically, abortion has been legal for a substantial amount of time. Abortion should therefore be legal in the US.
Con argument	Conception is the obvious place to pinpoint where life begins, as it starts the process in which a human being starts to form. To perform an abortion is to end this process, so it is therefore equivalent to terminating a human life. Allowing this process is clearly at odds with The Declaration of Independence, as this entitles us all to "Life, Liberty and the Pursuit of Happiness". Abortion should therefore be illegal in the US.	Many influential people hold the view that a fetus is considered as having human rights from the moment of conception, and question the morality of abortion. This includes former President Donald Trump and former Senator Sarah Palin. In fact, not only was the initial anti-abortion movement in the United States led by physicians and feminists alike, but the current Republican Party's platform officially advocates an anti-abortion position. Abortion should therefore be illegal in the US.

Note: An example of two 'good' and two 'bad' arguments participants could have seen in response to the statement *Abortions should be legal in the US*. Each participant only saw one argument (either a good or a bad argument) for each claim. In Experiments 1 and 2, half the arguments participants saw were in defence of the claim (i.e., pro arguments) and the other half of the arguments participants saw challenged the claim (i.e., con arguments).

Table 3
Example Left (Right) leaning arguments for claim *Abortions should be legal (illegal) in the US*.

	Inconsistent	Authority
Left leaning argument	Abortions are safe procedures that protect lives. Women who are denied abortions are also more likely to later have poorer mental and physical health, alongside financial problems. Instead of promoting abortions, increased access to birth control, health insurance, and sexual education would make abortions unnecessary. Abortions promote the idea that human lives are disposable when inconvenient. Abortions protect the bodily autonomy of women – a fundamental human right. Therefore, abortions should be legal in the US.	Many Americans argue that the recent overturning of Roe v Wade, the legislation that granted US citizens the right to abortion, marks a step backwards in the progress of human rights. Vice President and current Presidential Nominee Kamala Harris is a vocal pro-choice advocate, and before the legislation was overturned stated that "If the court overturns Row v Wade it will be a direct assault on freedom". Therefore, abortions should be legal in the US.
Right leaning argument	Instead of promoting abortions, increased access to birth control, health insurance, and sexual education would make abortions unnecessary. Abortions are safe procedures that protect lives. Women who are denied abortions are also more likely to later have poorer mental and physical health, alongside financial problems. Abortions protect the bodily autonomy of women – a fundamental human right. Abortions promote the idea that human lives are disposable when inconvenient. Therefore, abortions should be illegal in the US.	Many politicians are glad that Row v Wade, the legislation granting Americans the right to abortion, has recently been overturned. Mike Pence, former Vice-President of the United States under Donald Trump, is a Pro-Life advocate. During a visit to Florence Baptist Temple, he told the roughly 1500 congregants that "Many more are with us than are with them. Don't ever doubt it. Life is winning in America". Therefore, abortions should be illegal in the US.

Note: An example of two *inconsistent* and two *authority-based* arguments participants could have seen in response to the left-leaning claim *Abortions should be legal in the US* or right-leaning claim *Abortions should be illegal in the US* respectively.

of the pretest, one good and one bad argument for the same claim were presented alongside each other and participants had to choose which argument was better (i.e., a 2-alternative forced choice task). The purpose of the pretest was to show that in principle people can distinguish the good from the bad arguments. To foreshadow the results of the pretest, participants could distinguish good from bad arguments with above chance accuracy when both types of arguments were presented alongside each other.

Our expectations for the results from the Everyday Argument Assessment Task were based on the literature on argument evaluation discussed above. We expected participants to recognise that arguments with better evidence were of better quality, but also expected participants to perceive arguments that were already in line with their beliefs as being of better quality as well. This pattern would result in a main effect of argument quality and belief consistency. More specifically, we would expect good arguments to be rated as being of better quality than bad arguments on average, and for arguments more in line with participants' prior beliefs to be rated as being of better quality on average than arguments that were less in line with their prior beliefs.

We were also interested in a potential interaction pattern. It could be the case that participants always think an argument is good if they are in strong agreement with what the argument is saying, and only evaluate the argument by its evidential quality when they are not so strongly aligned with the argument. In other words, participants might have a blind spot for the evidential quality of arguments they agree with. This pattern would be demonstrated by the interaction between argument quality and belief consistency.

2. Pretest of materials used in Experiments 1 and 2

2.1. Method

2.1.1. Participants

A total of 66 participants pretested the materials used in Experiments 1 and 2. Of those participants, 17 failed the attention checks, leaving 49 participants (27 male, 21 female, 1 did not disclose the gender) from whom we analysed data. Participants were recruited through Prolific and restricted to be residents of the USA. Of the participants whose data we analysed; four were 18–24 years of age, 26 were 25–34 years of age, 13 were 35–44 years of age, four were 45–54 years of age, one was 65 years of age or older and one participant did not disclose their age. Over half of our sample (67 %) were either currently in or had completed university at the time of the experiment. The sample was mostly comprised of Democrats; 26 participants identified as Democrat, 10 were Independent/did not identify with a political party and only 12 identified as Republican (one participant did not disclose their political orientation).

2.1.2. Materials

Our material consisted of the eight claims shown in Table 1 that are relevant to the Pretest and the associated arguments. All materials were created specifically for the purpose of this study.

Each claim revolved around one of eight different political topics, relevant to American political discourse, listed in Table 1. Half of the claims were left-leaning (e.g., *Abortions should be legal in the US*), while the other half were right-leaning (e.g., *Further gun control laws are unnecessary*). The political alignment of each claim was fixed so that the same claims were always either left-leaning (i.e., claims related to Climate change, Abortion, Taking the knee, and Private prisons) or right-leaning (i.e., claims related to Cancel culture, Fracking, Habitual offender laws, and Gun control).

Every argument revolved around one of the political claims. For each claim we created four arguments: two arguments that supported it (*pro* arguments) and two arguments that challenged it (*con* arguments). In addition, one pro argument and one con argument were good arguments, and the other pro argument and con argument were bad arguments.

Good and bad arguments differed in the strength of their evidence. Good arguments were summarised versions of existing prevailing arguments in the discourse (predominantly from the educational bipartisan website www.procon.org). These arguments contained evidence that was either statistical or causal in nature. Bad arguments on the other hand contained argument fallacies (mainly circular reasoning in the form of what were essentially restatements of the claim, but also appeals to authority, appeals to popularity or appeals to tradition). We wrote each argument ourselves and every argument was 75 words in length. An example of a good, bad, pro, and con argument for one topic can be seen in Table 2. A full list of the 32 arguments in our study can be found in the online supplemental materials on OSF.

2.1.3. *Design*

Every participant worked on each of the eight topics listed in Table 1 that are relevant to the Pretest. For each topic, participants saw a good and a bad argument that either defended (i.e., were pro arguments) or challenged (i.e., were con arguments) the relevant claim in Table 1. For example, a participant could see either a good and a bad argument which each conclude that abortions should be legal in the US, or a good and a bad argument which each conclude that abortions should be illegal in the US. In total, each participant saw eight of the 16 pairs of good and bad arguments. The task itself was a two-alternative forced choice task, where for each topic, participants had to decide which of the two arguments presented was the better argument at making its case.

The Pretest had one independent variable, argument support. Argument support refers to whether the pair of arguments being shown were pro arguments or con arguments. Of the argument pairs shown to participants, four pairs of arguments (i.e., eight arguments) were pro arguments, and the other four pairs of arguments were con arguments, making argument support a within-subjects variable with levels *pro* and *con*. For which topics the argument pairs were pro and for which they were con was randomised for each participant.

The dependent variable of the Pretest was the probability of choosing the good argument from each argument pair consisting of one good and one bad argument.

2.1.4. *Procedure*

All experiments presented in this paper were approved by the Psychology Department's Ethics Committee. To ensure that the sensitive and potentially controversial nature of the materials did not create any psychological harm for our participants, participants in all studies presented in this paper were informed about the political nature of the task and the political topics discussed prior to the experiment. Participants in all experiments were also told that they would have to read arguments about the political topics that were discussed and make ratings about the quality of the arguments. Each participant gave their consent before taking part in the study and was debriefed after having taken part. Within the debrief, to ensure our participants did not leave the study misinformed, participants in all experiments were provided with links to websites (the majority of links leading to pages from the website www.procon.org) that lead them to unbiased good quality information regarding the political topics discussed in the experiment.

The procedure for the Pretest is illustrated in Appendix A. Each participant completed eight trials, with each trial concerning arguments related to one of the eight topics listed in Table 1 relevant to the Pretest. The order in which the topics were presented, which topics were defended by pro argument pairs and challenged with con argument pairs, as well as the position of the good argument in relation to the bad argument (i.e., above it or below it) in each argument pair was randomly determined for each participant (within the constraint that each participant saw 4 pro arguments and 4 con arguments).

For each topic, participants were first shown a short paragraph that briefly described the current context (e.g., an introduction to the debate surrounding abortion). After the introductory paragraph, participants were immediately shown a good and a bad argument that either supported or challenged the same claim together as a pair (e.g., either a pro-good and a pro-bad argument or a con-good and a con-bad argument about abortion). Participants were told to assume the arguments were made in good faith and due diligence was exercised to ensure the details were factually correct. Participants' task was to select which of the arguments they thought was the better of the two.

We included two additional topics as attention checks to ensure that participants were paying attention to the arguments presented to them. For these additional topics, participants were shown two arguments in favour of untrue and surprising claims (*All people are cannibals* and *Children are older than their biological parents*) and were told at the end of one of the arguments in each argument pair which argument to select as the answer. The combination of providing the correct response only at the end of one of the arguments in each argument pair and the requirement that both attention check items had to be answered correctly for inclusion in the Pretest may explain the relatively high failure rate of the attention checks (≈ 25 %). In the main experiments reported below, the rate with which participants failed the attention checks was noticeably smaller ($\leq$ 13 %). After all eight trials (plus the two attention check trials) were completed, participants answered basic demographic questions (including their age, level of education, political orientation and self-reported conservatism) and were debriefed.

2.2. *Results and discussion*

Results from the Pretest are shown in Fig. 1. As can be seen, for each set of good and bad arguments participants selected the good argument with a probability of well above 50 %. Furthermore, for all but one argument pair the 95 % binomial confidence interval did not include 0.5 (i.e., the chance level threshold). The only exception was the pair of con arguments for Gun control where the lower bound extended just below 0.5 (95 % CI [0.48, 0.83]). This indicates that across all arguments participants generally agreed with our designation that the good arguments provided stronger evidence for the corresponding claims than the bad arguments. In other words, people can in principle detect the evidential quality of the arguments. The question we will address in Experiment 1 is the role participants' beliefs play when rating the evidential quality of the arguments when shown individually and not adjacently to an argument of differing quality.

3. Experiment 1

3.1. *Method*

3.1.1. *Participants*

A total of 115 participants took part in Experiment 1. Of those participants, 14 failed the attention checks, leaving 101 (43 male, 54 female, 4 preferred not to disclose) participants in Experiment 1 from whom we analysed data.

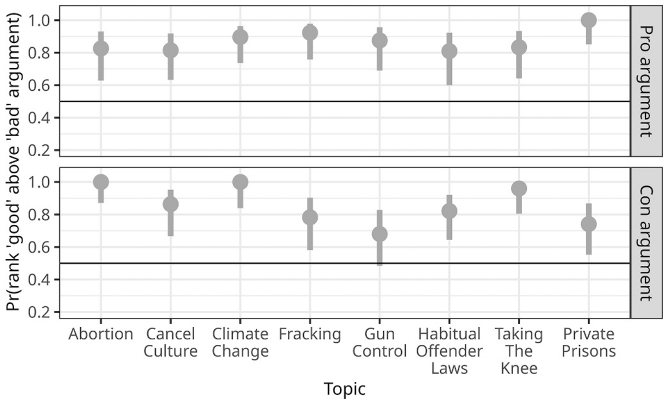


Fig. 1. Probability participants recognise good arguments as better than the bad arguments.
Note. Results from the Pretest. Each data point shows the probability of selecting the good argument out of a pair of good and bad arguments. The error bars show corresponding Wilson binomial confidence intervals. The horizontal line at y = 0.5 is the expected probability if participants selected arguments at chance level. Only one pair of arguments had a confidence interval that crossed the chance level threshold.

Participants were recruited through Prolific and restricted to be residents of the USA. Of the participants whose data we analysed; 10 were 18–24 years of age, 41 were 25–34 years of age, 31 were 35–44 years of age, seven were 45–54 years of age, six were 45–54 years of age, and six were 65 years of age or older. Over half (67 %) were either currently in or had completed university at the time of the experiment. As with the Pretest, the sample was mostly comprised of Democrats. 58 participants identified as Democrat, 24 were Independent/did not identify with a political party and only 19 identified as Republican.

3.1.2. Materials

Materials used in Experiment 1 were the same as those used in the Pretest and described in Section 2.1.2 (Materials).

3.1.3. Design

In Experiment 1, every participant worked on each of the eight topics listed in Table 1 that are relevant to Experiment 1. In the first part of the experiment, for each of the topics, they read the claim as it was in Table 1 and then rated their belief about the claim. In the second part of the experiment, they read one argument related to each claim and then rated the quality of the argument. Argument quality ratings comprised our dependent variable and were made on a 6-point scale ranging from *extremely bad* (represented by 1) to *extremely good* (represented by 6).

Experiment 1 had three independent variables, one of which was continuous. The first continuous independent variable was participants' belief ratings for each claim on a 7-point scale ranging from *extremely false* (represented by −3) to *extremely true* (represented by 3).[2] These claims (e.g., *Abortion should be legal in the US*) formed the basis of the conclusion for the pro arguments (e.g., *Abortion should therefore be legal in the US*); the conclusion for the con arguments were opposite to the conclusions of the pro arguments (e.g., *Abortion should therefore be illegal in the US*).

The other two independent variables, which were factors, were *argument support* and *argument quality*. Argument support related to whether the argument participants saw was in line with the claim they had seen (i.e., a pro argument) or challenged the claim they had seen (i.e., a con argument). Each participant saw eight claims, with arguments for half the claims being pro arguments and arguments for the other half of the claims being con arguments (like in the Pretest). Thus, argument support was a within-subjects factor with levels *pro* and *con*. The second factor, argument quality, related to whether the argument participants saw was good or bad (as operationalised in Section 2.1.2 - *Materials*). For each of the pro and con arguments, half were good, and the other half were bad. Thus, argument quality was also a within-subjects factor with levels *good* and *bad*.

The claims themselves were balanced so that half were left-leaning (e.g., *Abortion should be legal in the US*) while the other half were right-leaning (e.g., *Further gun control laws are unnecessary*). Political leaning was nested in the claims so that the same four claims were always left leaning (i.e., claims related to Climate change, Abortion, Taking the knee, and Private prisons) and the other four claims were always right leaning (i.e., claims related to Cancel culture, Fracking, Habitual offender laws, and Gun control). This was done to ensure that each participant agreed and disagreed with roughly the same number of claims and arguments being made.

The design of the Experiment 1 was such that for both left-leaning and right-leaning claims, participants saw one pro argument that was good (pro-good), one pro argument that was bad (pro-bad), one con argument that was good (con-good) and one con argument that was bad (con-bad). Participants therefore saw eight arguments in total, one argument for each topic. For which topic an argument was pro-good, pro-bad, con-good, or con-bad was randomised for each participant.

3.1.4. Procedure

The procedure for Experiment 1 is illustrated in Appendix B. All participants worked on each of the eight topics listed in Table 1 that are relevant to Experiment 1.

Experiment 1 was split into two different parts. In the first part we introduced participants to the topics and collected their belief ratings about the claims. In the second part, we presented the corresponding arguments and collected the argument quality ratings. In each part, the order in which the topics were presented was randomly determined for each participant.

In the first part of the experiment, participants first saw a short passage introducing them to a topic. Participants then saw a statement related to the topic of the passage (e.g., *Abortion should be legal in the US*). Participants had to rate their belief about the statement on a 7-point scale (*extremely false – extremely true*). After participants made a belief rating for one topic, they proceeded to make a belief rating for the next topic.

After making belief ratings for all eight topics and the two attention check topics (see below), participants proceeded to the second part of the study in which they were asked to make argument quality ratings. Participants were shown one argument at a time and had to rate the quality of this argument on a 6-point scale (*extremely bad – extremely good*). After participants rated the quality of all arguments, they completed a basic demographics questionnaire (including questions about their age, level of education, political orientation, and self-reported conservatism) before they were debriefed.

We included two attention check topics to ensure participants were attending to the stimuli. For these topics, participants had to make belief ratings for the claims *All people are cannibals* and *Children are older than their biological parents*. To pass the attention check participants had to rate the former claim as at least *mostly false* (i.e., equal to or smaller than −1 on the belief scale), the latter claim as at least *mostly true* (i.e., equal to or greater than 1 on the belief scale). When participants then saw the arguments for each of these claims, they were told within the argument exactly how to rate the argument on the scale. Participants had to make correct belief ratings and argument quality ratings for these items in order to be included in the analysis.

[2] Each experiment reported here included a second continuous variable that represented participants' confidence about their belief ratings, *meta-beliefs*, which was collected directly after each belief rating. As this variable did not yield any interesting results, we choose not to report it in the main text. Details of this variable can be found in Appendix C.

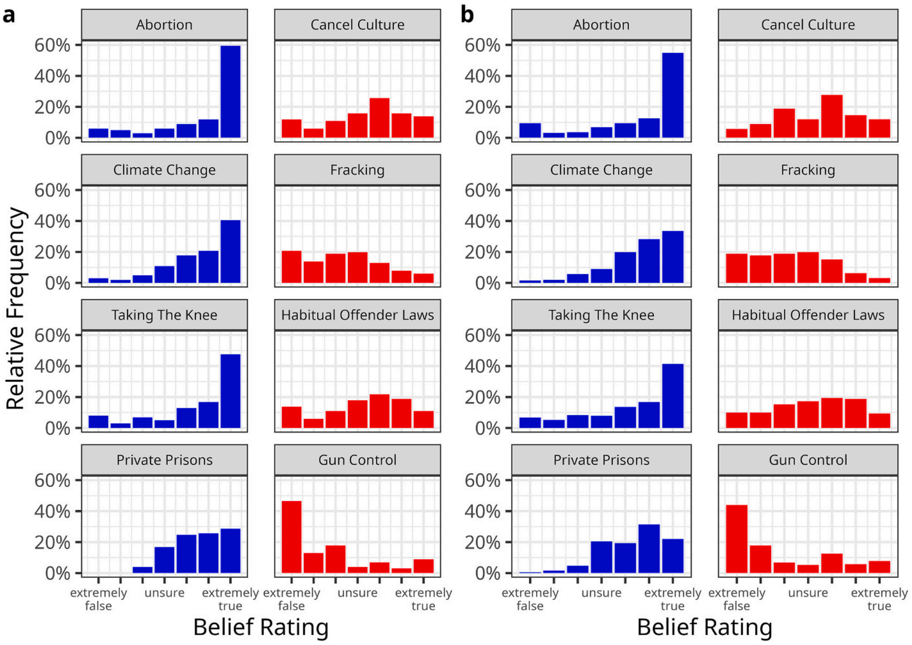


Fig. 2. Belief ratings for each claim for Experiments 1 and 2.
Note. The relative frequency (percentage) of belief ratings that were selected for each topic in Experiment 1 (a) and Experiment 2 (b). For Experiment 2, the panels reflect the number of belief ratings across both levels of argument order. Blue bars represent responses for topics that were left-aligned and red bars represent responses for topics that were right-aligned.

3.2. *Results*

3.2.1. *Distribution of belief ratings across topics*

One of the goals of the Everyday Argument Assessment Task was to investigate how people reason about arguments centred around disputable beliefs. To assess whether the beliefs espoused in the claims are in fact disputed, Fig. 2 shows the belief ratings for each claim for both Experiments 1 and 2. The blue bars show the responses to left-leaning claims and the red bars show the responses to right-leaning claims. Independent of the political leaning we can see that the beliefs are disputed; for almost all claims, at least some people indicated the claim is *extremely true* while others indicate the claim is *extremely false* (i.e., for almost all claims all points on the belief scale were represented in our data). In addition, we can see that participants are overall more left-leaning, which potentially reflects that our sample is mainly comprised of Democrats (as described in section 3.1.1. - *Participants*). The distributions of the blue left-leaning claims are clearly left-skewed whereas the distributions of the red right-leaning claims are more uniform with the distributions of some claims even showing a right-skew.

3.2.2. *Effects of belief on argument quality ratings*

There are two main research questions we aim to address with the Everyday Argument Assessment Task; 1) to what extent people evaluate an everyday argument by the quality of the evidence presented in said argument, and 2) to what extent a person's prior beliefs about the subject of an argument influence how they evaluate said argument. In addition, we are also interested in a potential interaction between the two, where strong agreement with the argument might cause someone to evaluate said argument less on the basis of its quality and more on the basis that they agree with the overall message. In the following analysis, we are therefore interested in three effects respectively; the main effect of argument quality, the effect of belief consistency (the interaction between belief and argument support), and the interaction between argument quality and belief consistency (three-way interaction between belief, argument support, and argument quality).

To address our research questions, we analysed participants' argument quality ratings (from 1 to 6) using a linear mixed model with fixed effects argument quality (good vs bad), argument support (pro vs con), belief (continuous -3 to 3 scale), belief squared (squared values of belief to test for a quadratic interaction between belief and argument quality), and all interactions. We estimated crossed random effects with by-participant and by-topic random terms. We initially attempted to employ the maximal model justified by the design. As the maximal model produced a singular fit, we simplified the model successively until this was not the case. We report the results for the final model here.[3] The pattern of significant and non-significant results remained the same for all random effect structures tested.

The main results for Experiment 1 are shown in Fig. 3. In the figure we can see a clear effect of argument quality; for both pro and con arguments, good arguments were consistently rated as better than bad arguments. This was supported by a significant main effect of argument quality, $F(1, 11.99) = 12.93, p = .004$. Argument quality ratings for good arguments were on average 0.69 points (95 % CI [0.312, 1.07]) higher than argument quality ratings to bad arguments.

Fig. 3 also shows a clear effect of belief consistency; there was a positive relationship between belief ratings and argument quality

[3] The maximal model included by-participant random effects and by-topic random effects. The topic grouping factor had the same intercepts and slopes as the fixed effects structure of the model. The participant grouping factor had the same intercepts and slopes as the fixed effects structure of the model excluding any random slopes involving by-argument quality interactions. A model with more random slopes would be unidentifiable. Our final model was a result of successively simplifying the maximal model until there were no model convergence issues (i.e., until the model was no longer producing a singular fit). The final model employed by-participant random intercepts and random slopes (without correlations) for the main effects of argument quality, belief, belief squared and argument support as well as belief by argument support and belief squared by argument support interactions. It also employed by-topic random intercepts and by-topic random slopes (without correlations) for main effects of argument quality, belief, belief squared and argument support as well as belief by argument support, belief squared by argument support and argument support by argument quality interactions. Full details of the process for reducing the random effect structure can be found in the online supplemental materials.

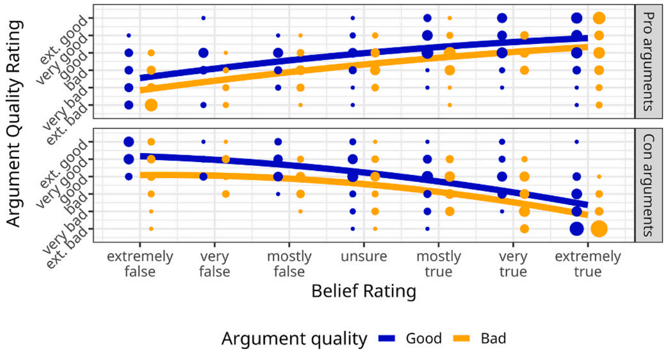


Fig. 3. Argument quality ratings as a function of belief consistency for Experiment 1.

Note. The dots show individual responses and the curved lines show predictions from the linear mixed model. Blue dots represent argument quality ratings to good arguments, orange dots represent argument quality ratings to bad arguments. The size of the dots represents the number of argument quality rating responses for the corresponding belief rating, with larger dots representing a larger number of responses. Data points are dodged so that responses for good and bad arguments do not overlap. Model predictions are based on the fixed effects of the final model including all interactions. Ext. = extremely.

ratings for pro arguments, and a negative relationship between belief ratings and argument quality ratings for con arguments. In line with this visual pattern, we found a significant belief by argument support interaction, $F(1, 11.69) = 175.27, p < .001$. For every additional point of belief, pro arguments were associated with an argument quality rating that was on average 0.40 points (95 % CI [0.32, 0.48]) higher and con arguments were associated with an argument quality rating that was on average 0.43 (95 % CI [0.35, 0.50]) points lower.

The overall effect of belief consistency is given by the average difference in argument quality ratings between participants at either end of the belief scale (i.e., the difference in predicted argument quality ratings between a claim rated as *extremely true* and a claim rated as *extremely false*) which was 2.39 points for pro arguments and 2.60 points for con arguments. As the main effect of argument quality was a 0.69 point difference between ratings for good arguments and ratings for bad arguments, we can interpret the effect or argument consistency as being around three times the magnitude of the main effect of argument quality.

Finally, Fig. 4 shows the results separately for each topic with the predictions derived from the by-topic random effects. The figure clearly shows the main patterns of Experiment 1 – significant effects of argument quality and belief consistency. While the main effect of argument quality is not evident for all topics (e.g., pro-abortion), it does hold for the vast majority of topics. In Experiment 2, which uses a larger sample size and the same materials, we can see an effect of argument quality for almost all topics (see Fig. 6 below; see also Appendix E for the number of responses per argument and experiment). Furthermore, in Experiment 3 (Section 5), we more systematically manipulate what constitutes a bad argument and find essentially the same effect of argument quality across all topics (see Fig. 8 below). Thus, the effect of argument quality does not appear to differ systematically across topics.

The effect of belief consistency, on the other hand, can be seen for every argument shown in Fig. 4. For each argument, participants whose prior beliefs are in line with the overall message of the argument (i.e., participants whose belief is *extremely true* for pro arguments and *extremely false* for con arguments) tend to rate arguments as better than participants whose beliefs are not in line with the arguments (i.e., whose belief is *extremely false* for pro arguments and *extremely true* for con arguments). In other words, the overall effect we see in Fig. 3 is not an artifact resulting from aggregating across topics; what determines participants argument quality ratings for the same argument, to a large degree, is their prior belief about a topic.

In none of our analysis of Experiment 1 did we find a significant interaction between argument quality and belief consistency (see Appendix D for a detailed description). We did see a main effect of belief squared, $\beta = -0.04$, $F(1, 9.26) = 9.07$, $p = .014$, suggesting a slight quadratic effect of the belief. However, this is a small effect and only meaningful when interacting with argument quality (which was not significant, see Appendix D), so we do not discuss this further.

3.3. *Discussion*

Experiment 1 established the main results patterns we find in the Everyday Argument Assessment Task. Participants can distinguish good from bad arguments. However, participants' prior beliefs of the arguments' conclusions play an even larger role on their argument quality ratings. The average difference in argument quality ratings between participants at either end of the belief scale (≈ 2.5) was around thrice the average difference in argument quality ratings (≈ 0.7) between good and bad arguments. Furthermore, we did not find evidence for an interaction between belief consistency and argument quality.

4. Experiment 2

Experiment 1 showed that prior beliefs play a larger role in participants' argument evaluation judgements than argument quality itself. The goal of Experiment 2 was to replicate this finding while controlling for a potential experimental confound; the order in which participants provide belief ratings and argument quality ratings and the resulting demand characteristics.

In Experiment 1, participants were first asked to provide their beliefs about a claim and subsequently asked to evaluate an argument related to that claim. The ordering of the task might cause participants to infer that their beliefs are relevant to their argument evaluation rating (even though we told them to make the argument quality rating independent of their beliefs). To address this possibility, in Experiment 2, we manipulated the order in which participants rate the claims and evaluate the arguments. Half of the participants first rate their belief concerning the claims and then evaluate arguments about these claims as in Experiment 1. The other half of participants first evaluate the arguments and *then* rate their beliefs about the claims that were argued about. If the results in Experiment 1 are due to order effects, we would expect to see the same pattern of results found in Experiment 1 only for the group of participants in Experiment 2 who work on the two tasks in the same order as participants in Experiment 1.

We also considered the possibility that participants might be answering the questions in a way as to emphasise their beliefs about the topics, even when this is not what is being asked of them. Existing literature suggests that socio-political beliefs can form part of a person's sense of self (e.g., Bonomi et al., 2021), and some individuals might be making belief ratings and argument quality ratings for a topic in a single direction only to demonstrate that their view is important. Anecdotally it is clear that many participants felt it important to have their opinions heard, as the textbox at the end of Experiment 1 intended for feedback about the experiment was often filled with justifications about their beliefs concerning the topics discussed in the experiment. In order to address this possibility, in Experiment 2 we highlighted all aspects of the procedure at the very beginning of the experiment, making it very clear when we wanted participants to give their own opinion – for their belief ratings – and when we wanted participants to try to be objective with their judgements – for their argument quality ratings. If participants demonstrating their opinions where it is not relevant was driving the findings of Experiment 1 and the aforementioned change in procedure addressed this issue, then we would expect the belief consistency effect to be reduced for both order manipulations in Experiment 2.

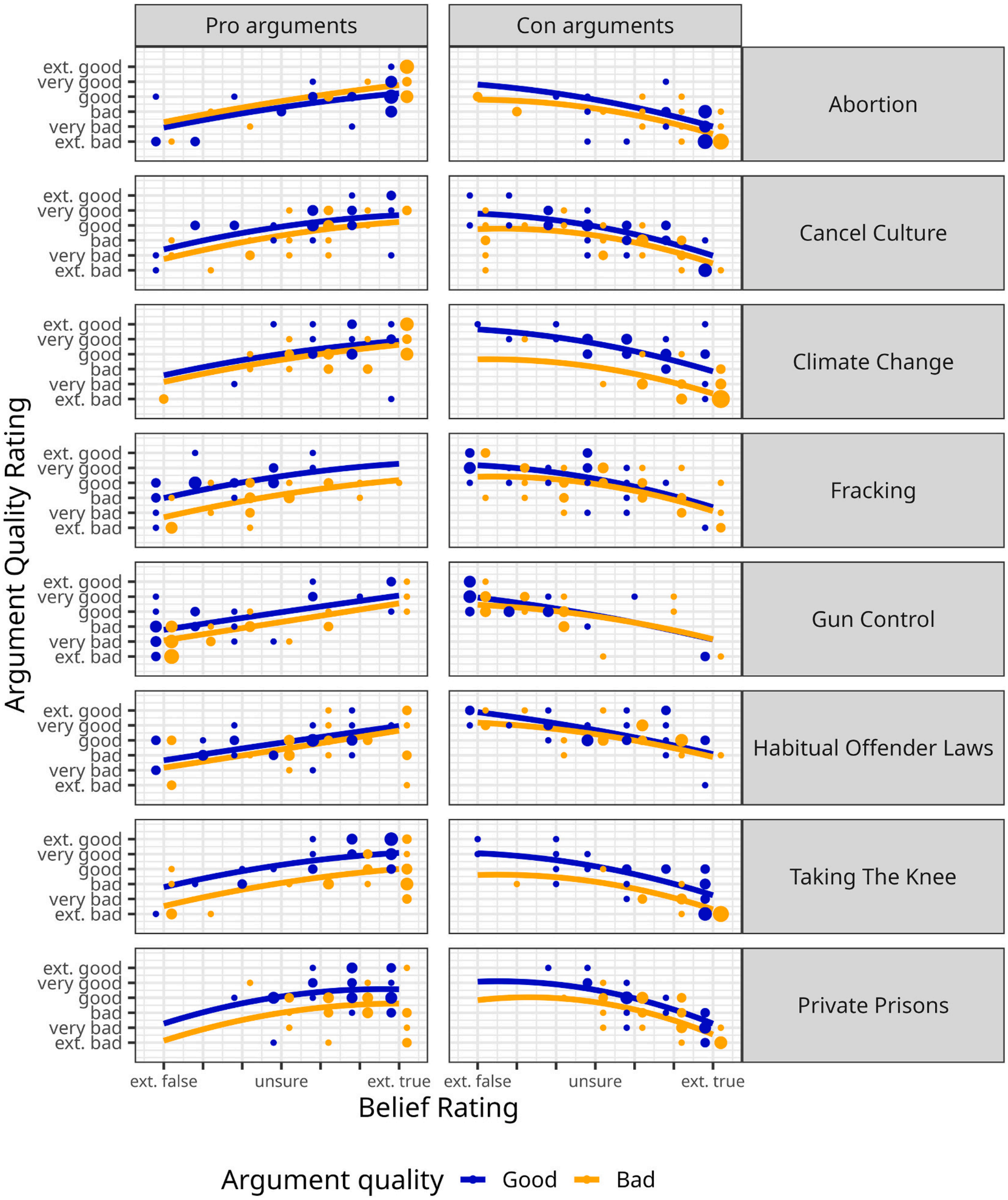


Fig. 4. Argument quality ratings as a function of belief consistency for each argument in Experiment 1.
Note. Results of Experiment 1 conditional on the topic and the level of argument support. Each line and colour in each panel shows responses to exactly one argument (i.e., there is no aggregation across items within a panel). The dots show individual responses and the curved lines show predictions from the linear mixed model. Blue dots represent argument quality ratings to good arguments in the data, orange dots represent argument quality ratings to bad arguments in the data, and the size of the dots represents the number of argument quality rating responses for the corresponding belief rating. Data points are dodged so that responses for good and bad arguments do not overlap. Model predictions are based on the fixed effects of the final model and the random effects of the by-topic grouping factor. Ext. = extremely.

4.1. Methods

4.1.1. Participants

A total of 200 participants took part in Experiment 2. 100 participants were assigned to the *argument-second* condition in which they were presented with and rated the quality of the arguments *after* they were presented with and rated their belief about the claims as was the procedure in Experiment 1. The other 100 participants were assigned to the *argument-first* condition in which they were presented with and rated the quality of the arguments *before* they were presented with and rated their belief about the claims. Of those participants, 9 failed the attention checks, which left 191 participants (101 male, 83 female, 7 prefer not to say) from whom we analysed data. Of these participants, 96 were in the argument-second condition and 95 were in the argument-first condition.

Participants were recruited through Prolific and restricted to residents of the USA. Of the participants whose data we analysed; 36 were

18–24 years of age, 62 were 25–34 years of age, 42 were 35–44 years of age, 25 were 45–54 years of age, 14 were 55–64 years of age and 11 were 65 years of age or older (one participant did not disclose their age). In contrast to our sample in Experiment 1, only around 15 % of our sample in Experiment 2 were either currently in or had completed university at the time of the experiment. As with Experiment 1, the sample was mostly comprised of Democrats; 115 participants identified as Democrat, 30 were Independent/did not identify with a political party and only 45 participants identified as Republican (one participant declined to disclose their political orientation).

4.1.2. *Materials and design*

The materials used in Experiment 2 were mostly the same as were used in Experiment 1. The only materials that were different between the two experiments were the claims and argument conclusions for the abortion topic, which changed from *Abortion should remain legal/be illegal in the US* in Experiment 1 to *Abortion should be legal/be illegal in the US* in Experiment 2 following the change in US abortion laws in the time between the two experiments (i.e., the supreme court ruling overturning Roe v Wade in 2022). The design of Experiment 2 was also mostly identical to that of the main study in Experiment 1, with the additional manipulation of whether participants saw and rated the quality of the arguments before they made belief ratings or after they made belief ratings.

4.1.3. *Procedure*

The procedure was mostly identical to that of Experiment 1 except for two key differences; the instructions given to participants and the additional manipulation of the order in which the arguments were presented relative to the claims. With regard to the instructions, at the beginning of the experiment participants were now told the procedure for the rest of the experiment in detail. Important details included:

- Participants would see a claim and make a belief rating about the claim (with an example of what a claim and a corresponding belief rating question would look like).
- Participants would see an argument and rate how well the argument made its case. Each argument would either defend the claim participants make their belief rating about (e.g., *Abortion should be legal in the US*), or challenge said claim (e.g., *Abortion should be illegal in the US*). Examples of an argument, possible claims participants could see related to this argument and a corresponding argument quality rating question were provided.
- The quality of the argument is determined only by the evidence in the argument, and the quality of the argument should be rated independently of what the participant believes about the claim.
- Participants would make belief ratings, for which we were interested in their own personal opinions. Participants would also make argument quality ratings, which we wanted them to do objectively ‘independent of [their] beliefs’.

Participants were also reminded of these instructions as they became relevant throughout the course of the experiment. The instructions at the beginning of and throughout the experiment were adjusted for each argument order condition (argument-first and argument-second) so that the instructions were consistent with the procedure of each condition.

As with the Pretest and Experiment 1, after participants completed all trials (including attention check trials), they answered basic demographic questions (including their age, level of education, political orientation and self-reported conservatism) and were debriefed.

4.2. *Results*

The aim of Experiment 2 was to replicate the findings of Experiment 1 while controlling for possible confounds. Therefore, we used a similar mixed model as in Experiment 1 to assess the influence of participants' beliefs on their argument quality ratings. The fixed effects were argument quality (good vs bad), argument support (pro vs con), belief (continuous scale from -3 to 3), belief squared (belief values squared to investigate the quadratic interaction between belief and argument quality), argument order (argument-first vs argument-second), and all interactions. We estimated crossed random effects with by-participant and by-topic random terms. As with Experiment 1 we began with the

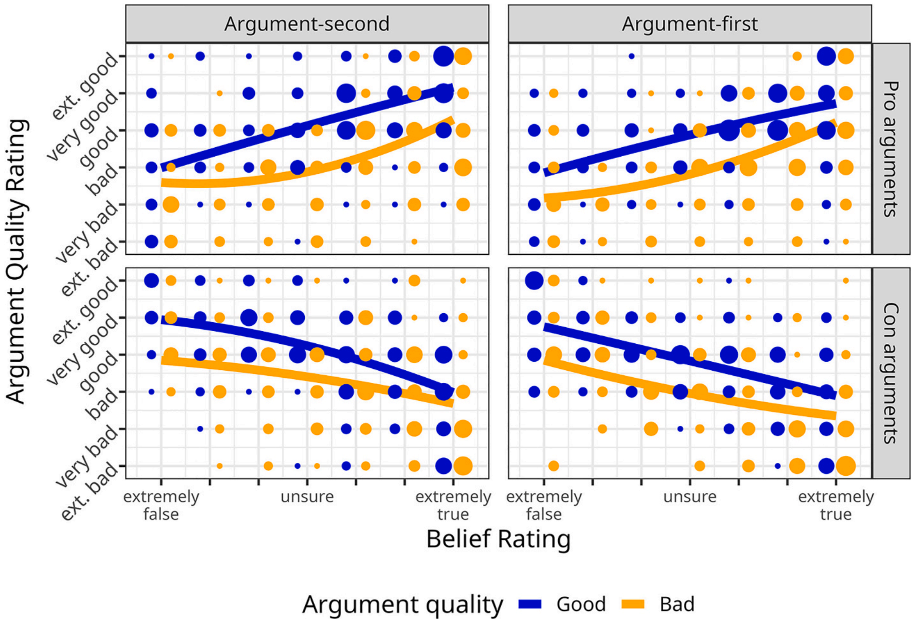


Fig. 5. Argument quality ratings as a function of belief consistency, argument quality, argument support, and argument order for Experiment 2.
Note. The dots show individual responses and the curved lines show predictions from the linear mixed model with the quadratic term. Blue dots represent argument quality ratings to good arguments in the data, orange dots represent argument quality ratings to bad arguments in the data, and the size of the dots represents the number of argument quality rating responses for the corresponding belief rating. Data points are dodged so that responses for good and bad arguments do not overlap. Model predictions are based on the fixed effects of the final model including all (significant and non-significant) interactions. Ext. = extremely.

maximal model justified by the design and reduced this until the model converged without a singular fit.[4] The pattern of significant and non-significant results remained the same for all random effect structures tested except for where explicitly mentioned below.

The main results for Experiment 2 are shown in Fig. 5. The figure and model show a clear main effect of argument quality, $F(1, 10.30) = 58.91$, $p < .001$, as was also evident in Experiment 1. Argument quality ratings for good arguments were on average 0.84 points (95 % CI [0.57, 1.10]) higher than argument quality ratings for bad arguments. There was no evidence that the main effect of the argument quality was moderated by argument order, as the interaction between argument quality and argument order was not significant, $F(1, 24.92) = 0.13$, $p = .720$.

We can also see the effect of belief consistency in Fig. 5, which corresponds to a positive relationship between belief ratings and argument quality ratings for pro arguments (where higher belief ratings correspond to greater agreement with the argument), and a negative relationship between belief and argument quality ratings for con arguments (where higher belief ratings corresponds to less agreement with the argument). In line with this visual pattern, we found a significant belief by argument support interaction, $F(1, 241.02) = 214.60, p < .001$. For every additional point of belief, pro arguments were associated with an argument quality rating that was on average 0.32 points, 95 % CI [0.26, 0.39], higher and con arguments were associated with an argument quality rating that was on average 0.27 points, 95 % CI [0.21, 0.33], lower. We can also see from Fig. 5 that the pattern of the belief consistency effect does not differ greatly between levels of argument order, which is reflected in the non-significant three-way interaction between belief, argument support, and argument order, $F(1, 161.78) = 0.11$, $p = .745$.

The average difference in argument quality ratings between participants at either end of the belief scale (i.e., the difference in predicted argument quality ratings between a claim rated as *extremely true* and a claim rated as *extremely false*) was 1.95 points for pro arguments and 1.62 points for con arguments. As the main effect of argument quality (the average difference in argument quality ratings between good arguments and bad arguments) was only 0.84, we can interpret this effect of belief consistency as being around twice the size as the effect of argument quality. These results replicate Experiment 1; for both experiments the effect of belief consistency was larger than the effect of argument quality.

In contrast to Experiment 1, we found some evidence of a belief consistency by argument quality interaction. However, this pattern was generally quite weak and not consistent across the levels of the argument support factor (see Appendix D for details). Furthermore, this interaction was also only significant in some of the random effects structures tested for the model (see supplemental material on OSF for more details).

There was also a small but significant main effect of argument order itself, $F(1, 39.36) = 7.49$, $p = .009$. Argument quality ratings in the argument-second condition were on average 0.25 points (95 % CI [0.07, 0.43]) higher than ratings in the argument-first condition. None of the interactions with argument order (including those mentioned previously and the less interesting interactions that were not) reached statistical significance (smallest $p = .074$).

Finally, Fig. 6 shows the results separately for each topic with the predictions derived from the by-topic random effects. As with Fig. 4, we can clearly see main effects of argument quality for almost all topics, and a main effect of belief consistency for all of the arguments. Participants generally judge good arguments as better than bad arguments for each topic (i.e., the pattern is more consistent compared to Experiment 1). As with Experiment 1, issues remain in knowing what accounted for some of the variation in the effect of argument quality across topics (e.g., in pro Climate Change vs pro Fracking argument in Fig. 6), which we address in Experiment 3 by manipulating all bad arguments in the same way. Like Experiment 1, the effect of belief consistency was much more consistent across topics; for all topics participants judge arguments aligned line with their beliefs as better than arguments which are not.

4.3. Discussion

The results of Experiment 2 replicated the results of Experiment 1, which did not appear to be moderated by argument order. Firstly, we again observed the belief consistency effect for both pro and con arguments, albeit descriptively slightly weaker than in Experiment 1 (the average effect per point on the belief rating scale was around ± 0.4 in Experiment 1 and around ± 0.3 in Experiment 2). Secondly, we also replicated the main effect of argument quality with a similar magnitude (0.7 in Experiment 1 and 0.8 in Experiment 2). The overall effect of belief consistency was again much larger than the effect of argument quality. The average difference in argument quality ratings at either end of the belief scale (≈ 1.8) was around twice the average difference in argument quality ratings (≈ 0.8) between good and bad arguments.

In contrast to Experiment 1, we found some evidence for the belief consistency by argument quality interaction, but this evidence was very weak. Furthermore, the descriptive patterns we found were not consistent across levels of the argument support factor. The data suggests there might be a quadratic interaction for pro arguments, and a linear interaction for con arguments (see supplemental materials on OSF for details).

5. Experiment 3[5]

In our first two experiments we established two important phenomena in the Everyday Argument Assessment Task; when evaluating arguments, participants are attentive to the quality of the evidence in the argument, but also have the tendency to evaluate arguments in line with their beliefs more favourably than arguments which are not. As shown in Experiment 2, these patterns were independent of demand characteristics and the order in which participants worked on the task.

One shortcoming of the previous experiments was that while what constitutes a good argument was fairly well controlled – the good arguments were based on information from a non-partisan website – this was not the case for bad arguments. Some bad arguments were circular,

[4] For the by-topic random effects term, the maximal model included random intercepts and random slopes for all main effects (argument quality, argument support, belief, belief squared, argument order) and their interactions. The by-participant random term included random intercepts and random slopes for all main effects excluding argument order and interactions excluding those involving argument order and argument quality. By-participant random terms for argument order were not included in the model as argument order did not vary for participants whereas it did within items. By-participant random slopes for interactions involving argument quality were excluded from the model as the model was unidentifiable otherwise. By-topic random intercepts and slopes was identical to the fixed effects structure in the model. As the maximal model showed convergence issues, we simplified it successively until there were no convergence issues (i.e., until the model no longer produced a singular fit). We arrived at a final model which employed by-participant and by-topic random intercepts and slopes without correlations. By-participant random slopes were included for main effects of belief, belief squared, argument support, and argument quality plus the belief by argument support and belief squared by argument support interactions. By-topic random slopes were included for main effects of belief, belief squared, argument support, argument quality, and argument order plus interactions of belief by argument support, belief by argument quality, belief squared by argument support, belief squared by argument quality, argument support by argument quality, argument quality by argument order and argument support by argument order. Full details can be found in the online supplemental materials on OSF.

[5] The results of a similar experiment as reported here have been published as part of Deans-Browne et al. (2024). The results reported here are from a completely new experiment with an improved design and improved materials.

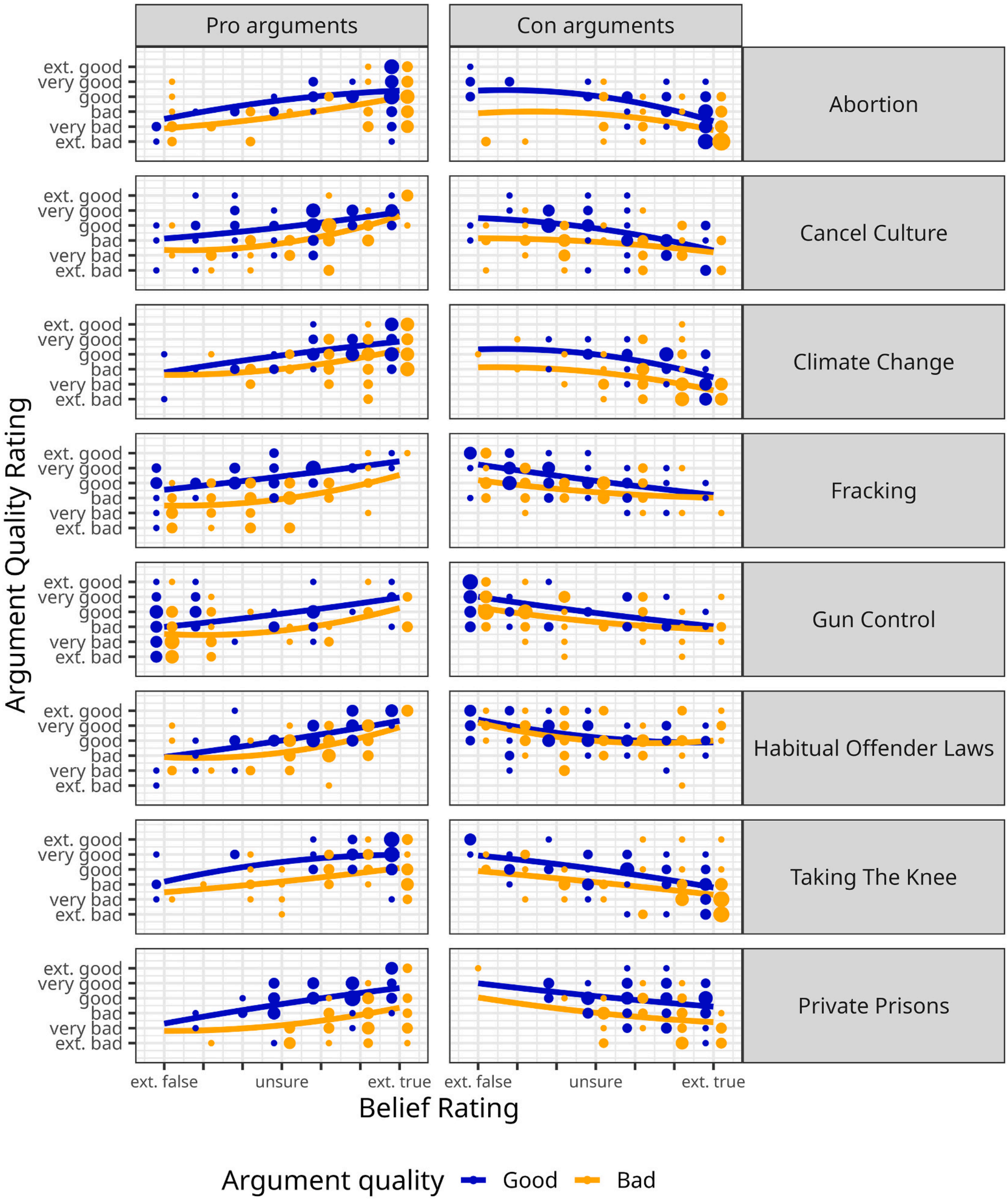


Fig. 6. Argument quality ratings as a function of belief consistency for each argument Experiment 2.
Note. Results of Experiment 2 conditional on the topic and the level of argument support. See Fig. 4 for details.

such as the pro-bad example in Table 2, and others were based on appeals to authority, such as the con-good example in Table 2. The goal of Experiment 3 was to understand which specific features make an everyday argument ’bad’. To this end, we replaced the unsystematically manipulated bad arguments with systematically manipulated ones. In Experiment 3 we manipulated the bad arguments in two different ways: half the bad arguments had inconsistent evidence, and the other half were based on appeals to authority.

The structure of inconsistent arguments was such that while some of the evidence was in support of the claim espoused at the end of the argument, the rest of the evidence instead opposed the claim (see example in Table 3). As a consequence, reading an inconsistent argument attentively in its entirety was difficult as the overall narrative was confusing and the argument as a whole did not make much sense. The only way that an inconsistent argument could make sense to a reader would be for them to ignore the inconsistent parts in the middle of the argument.

The inconsistent arguments were contrasted with arguments based on appeals to authority (Harris et al., 2016). These arguments reasoned that participants should believe the claim being espoused because it is supported by a well-known but non-expert authority figure (i.e., a politician, celebrity, or media personality). Hence, the evidence for the claim provided by these arguments was limited. The arguments based on appeals to authority provide a good contrast to the inconsistent arguments, as unlike the inconsistent arguments they only present evidence going in one direction (either in support of or against the claim) and are easy to understand when scrutinised. As the inconsistent arguments were difficult to parse, we expected participants to rate them as worse than the authority-based arguments on average, which we expected to be rated as worse than the good arguments overall.

5.1. *Methods*

The methodology of Experiment 3 generally followed that of Experiment 1; participants first provided belief ratings for claims and then provided argument quality ratings. The main difference to Experiment 1 was that participants saw three types of arguments: good, inconsistent, and authority-based arguments. In addition, we simplified the design and removed the argument support factor. Instead of having pro and con arguments relative to the claim participants rated initially, the claim now always matched the conclusion of the argument. We also added two more topics (Affirmative action and Secularisation of government) that participants worked on.

5.1.1. *Participants*

Participants were recruited through Prolific and restricted to native English speakers in the USA. A total of 119 participants took part in the study. Of those, 16 failed attention checks and two did not speak English natively (a pre-requisite for participation). This left us with 101 participants (47 male, 52 female, 1 other, 1 did not disclose) from whom we analysed data. Of those participants whose data we analysed; 9 were 18–24 years of age, 31 were 25–34 years of age, 30 were 35–44 years of age, 14 were 45–54 years of age, 11 were 55–64 years of age and 5 were 65 years of age or older (one participant did not disclose their age). 68 % of our sample was either currently in or had completed university at the time of the experiment. In this experiment we balanced the sample for political orientation; 54 participants identified as Democrat and 45 identified as Republican (1 identified as an Independent/did not identify with a political party).

As with the Pretest, Experiment 1, and Experiment 2, after participants completed all trials (including attention check trials) they answered basic demographic questions (including their age, level of education, political orientation, and self-reported conservatism) and were debriefed.

5.1.2. *Materials*

In total the materials consisted of claims and arguments for ten topics. We created new claims and arguments revolving around the topics of Affirmative action (*Affirmative action leads to a more just/unjust society*) and Secularisation of government (*Separating church from state causes more good than harm/harm than good*). We did this to improve the power of our study, which is largely determined by the number of topics worked on across participants due to use of crossed random effects for participants and topics (Westfall et al., 2014). Participants worked on eight topics (as in Experiments 1 and 2) with the eight topics randomly selected from the pool of ten topics anew for each participant. Each participant provided responses for one claim and one argument for each topic; either a good argument, an inconsistent argument, or an authority-based argument. Half the arguments participants saw were for left-leaning claims (e.g., concluding *Abortions should be legal in the US*) and the other half of the arguments participants saw were for right-leaning claims (e.g., concluding *Abortions should be illegal in the US*).

One important change compared to the previous experiments was that the claims participants rated initially depended on which argument they were shown later (i.e., each claim was now always in line with the argument they were subsequently presented with). For example, if a participant were asked to rate an argument that concluded *Abortions should be legal in the US,* then their belief rating would be for this same claim. Likewise, if they saw an argument that concluded *Abortions should be illegal in the US,* then this would be the claim they provided a belief rating for. This meant that all arguments were pro the claim participants saw, making *argument support* a redundant factor. This was done to make the task easier for participants, as they were now only evaluating arguments that supported the claims they had seen. Each participant still saw an equal number of left-leaning and right-leaning claims as they did before, which meant each participant still agreed and disagreed with roughly half the number of items they were presented with.

The good arguments used in Experiment 3 were essentially the same as were used in Experiment 1 (some minor alterations were made to further improve the readability of some of the arguments, see supplemental Materials on OSF). In this experiment, bad arguments were now split into those that were inconsistent and those that were based on appeals to authority.

Inconsistent arguments contained evidence supporting the overall conclusion, but also contained evidence that went against the overall conclusion. Like the good arguments, the evidence was from arguments already established in the current discourse. The evidence in the inconsistent arguments was also good evidence in that it strongly supported or opposed the overall conclusion of the inconsistent argument. However, the argument as a whole was inconsistent as it contained both evidence in favour of and in opposition to the overall conclusion (see Table 3 for an example). All inconsistent arguments followed the same general ‘sandwich’ structure; they started with evidence in line with the conclusion, followed by evidence opposing of the conclusion, finally followed by more evidence in line with the conclusion (i.e., the inconsistent information was ‘sandwiched’ between two pieces of information that opposed it). One special feature of the inconsistent arguments is that for both left-leaning and right-leaning arguments for a given topic, the arguments contain essentially the same sentences (with the exception of the final concluding sentence), just in a different order (which can be seen by a careful look at the inconsistent arguments in Table 3).

Authority-based arguments on the other hand emphasised the endorsement of an authority figure as evidence for its conclusion. These arguments were based on real statements that celebrities, politicians, or organisations had made regarding various political issues. In this way, these arguments did not mislead participants on what authority figures had actually said to the best of our knowledge. The evidence in these arguments only weakly supported the arguments' conclusions in that the main evidence provided was simply that the conclusion was endorsed by an authority figure. However, unlike the inconsistent arguments and more like the good arguments, arguments based on appeals to authority were consistent in that they only contained statements in support of the overall conclusion.

5.1.3. Design

The design of Experiment 3 was similar to that of Experiment 1. However, in Experiment 3 participants saw good, inconsistent, and authority-based arguments. Half of the arguments participants saw (i.e., four arguments) were good, whilst a quarter of the arguments participants saw (i.e., two arguments) were inconsistent, and a quarter were based on appeals to authority.

For each type of argument participants saw (good, inconsistent, or authority-based), half were left-leaning and the other half were right leaning. As discussed in the previous section (5.1.2. – *Materials*), each claim was now in the direction of the argument (i.e., all arguments were pro on the *argument support* factor). As such, argument support was redundant as a factor. Instead, we now included argument leaning (i.e., the effect of an argument being left-leaning vs right-leaning) as a factor.

5.1.4. Procedure

The procedure was the same as that of Experiment 1, described in Section 2.1.4. - *Procedure* and illustrated in Appendix B. The only difference was that participants worked on eight topics now from a pool of ten possible topics, plus the two attention checks items.

5.1.5. Results

5.1.5.1. Distribution of belief ratings across topics. To assess whether the beliefs espoused in the claims were disputed in Experiment 3 as they were in Experiments 1 and 2, Fig. 7 shows the belief ratings for each claim in Experiment 3. The blue bars in column a show the responses to left-leaning claims and the red bars show the responses to right-leaning claims. Recall that in Experiment 3, there was a left-leaning and a right-leaning version of a claim for each topic, and there were two additional topics that participants could see (Affirmative action and Secularisation).

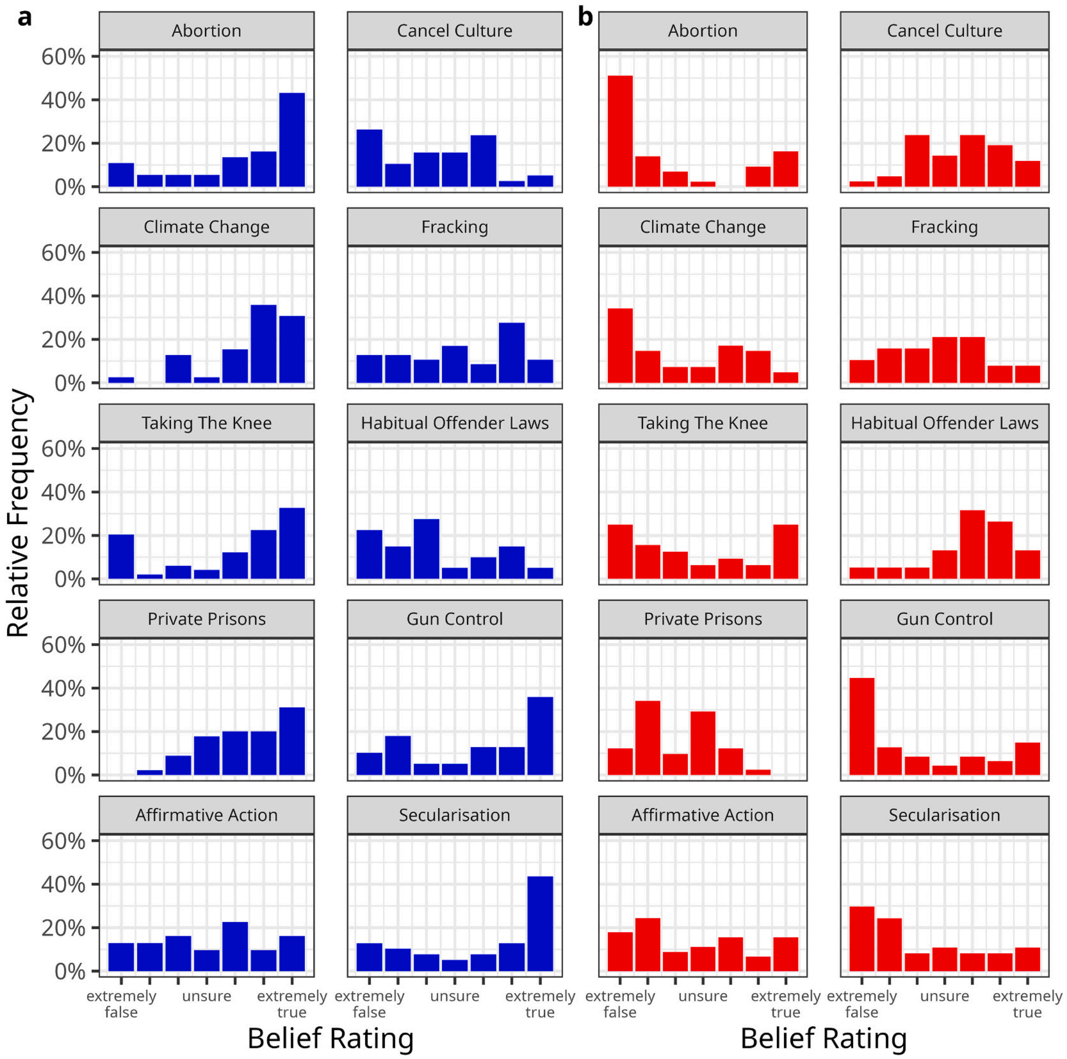


Fig. 7. Relative frequency distributions of belief ratings for each claim in Experiment 3.
Note. The relative frequency (percentage) of belief ratings that were selected for each topic in Experiment 3. Blue bars in column a represent responses to claims that were left-leaning and red bars in column b represent responses for claims that were right-leaning.

Independent of the political leaning we can see that the beliefs are disputed as they were in Experiments 1 and 2. For almost all claims, at least some people indicated the claim is *extremely true* while others indicate the claim is *extremely false* (i.e., for almost all claims all points on the belief scale were represented in our data). And while we still see some evidence for a more left-leaning bias, we also see some items that show a markedly bimodal distribution with peaks on both ends of the scale (e.g., Abortion and Taking the knee). There is also one topic, Habitual offender laws, which shows a pattern against the overall trend (i.e., a right-leaning bias). Finally, two topics (i.e., Fracking and Affirmative action) show an almost uniform pattern.

5.1.5.2. Main analysis. The main purpose of Experiment 3 was to investigate what aspects of an argument make someone more likely to perceive it as worse. The authority-based arguments, whilst only evidenced by the endorsement of a famous individual, were much easier to understand and make sense of than the inconsistent arguments if read carefully (for examples of both see Table 3). We therefore expected inconsistent arguments to be rated as lower on average than authority-based arguments, but for both types of arguments to be rated as worse than the 'good' arguments that were based on the current discourse for the topics discussed in the experiment. We also expected to replicate the belief consistency effect found in Experiments 1 and 2, where participants rate arguments in line with their prior beliefs as better on average than arguments that are not in line with their beliefs.

To address our main research question, we analysed participants' argument quality ratings (from 1 to 6) using a linear mixed model from fixed factors argument type (good, inconsistent, authority-based), argument leaning (left-leaning vs right-leaning), belief (continuous -3 to 3 scale), belief squared (squared values of belief to investigate the quadratic interaction between belief and argument type), and all interactions. We estimated crossed random effects with by-participant and by-topic random terms. We initially attempted to employ the maximal model justified by the design. As the maximal model produced a singular fit, we simplified the model successively until this was not the case.[6] The pattern of significant and non-significant results overall remained the same for all random effect structures tested with some exceptions mentioned explicitly below.

The main results for Experiment 3 are shown in Fig. 8. In the figure we can see a clear effect of argument type in both panels that differed from our expectation. As expected, good arguments received the highest ratings on average. However, the second highest rated arguments were inconsistent arguments followed by authority-based arguments. This visual pattern was supported by a significant main effect of argument type, $F(2, 292.46) = 39.39$, $p < .001$.

Investigation of the marginal means reveal responses for good arguments were on average 0.70 points (95 % CI [0.45, 0.95], $t(417) = 6.70$, $p < .001$) higher than inconsistent arguments, which were themselves on average 0.60 points (95 % CI [0.29, 0.91], $t(110) = 4.72$, $p < .001$) higher than authority-based arguments. The difference of 1.30 points (95 % CI [1.00, 1.59], $t(117) = 10.72$, $p < .001$) between good and authority-based arguments was unsurprisingly also significant (the three reported p-values for pairwise comparisons were adjusted using the Holm method).

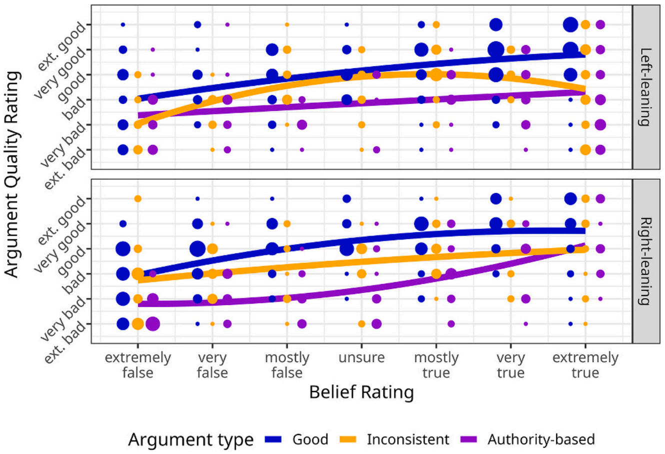


Fig. 8. Argument quality ratings as a function of belief consistency and argument leaning for Experiment 3.
Note. The dots show individual responses and the curved lines show predictions from the linear mixed model. Blue dots represent argument quality ratings to good arguments, orange dots represent argument quality ratings to inconsistent arguments, and purple dots represent argument quality ratings to authority-based arguments. The size of the dots represents the number of argument quality rating responses for the corresponding belief rating, with larger dots representing a larger number of responses. Data points are dodged so that responses for good, inconsistent and authority-based arguments do not overlap. Model predictions are based on the fixed effects of the final model that includes the non-significant three-way interaction terms. Ext. = extremely.

In the final mixed model, there was also a significant interaction between argument type and argument leaning, $F(2, 577.65) = 3.26$, $p < .039$. This interaction was not significant in the maximal model, $F(2, 18.28) = 2.60$, $p < .102$. Inspection of the interaction in the final model revealed the same ordering of conditions based on significance tests for both argument leaning conditions (i.e., good > inconsistent > authority-based, see supplemental materials on OSF for full analysis). Thus, argument leaning did not moderate the effect argument quality in a substantively relevant manner.

Fig. 8 also shows a clear effect of belief consistency. For all argument types (good, inconsistent, and authority-based), for both levels of argument leaning, there was a positive association between belief in the claim and the argument quality rating itself. In line with this visual pattern, we found a significant main effect of belief, $F(1, 476.13) = 131.07$, $p < .001$. Every additional point of belief was associated with an argument quality rating that was on average 0.26 points (95 % CI [0.21, 0.31]) higher. Despite the visual impression potentially suggesting an attenuated effect of belief for the inconsistent arguments, the interaction between belief and argument type was not significant ($F(2, 599.75) = 1.31$, $p = .270$), and all three marginal slopes were positive and significant (largest $p < .001$), suggesting that the linear effect of belief was similar for all argument types. Importantly, there was no significant interaction between belief consistency and argument leaning ($F(1, 498.53) = 2.30$, $p = .130$), suggesting that the effect of belief consistency was similar for both left-leaning and right-leaning arguments.

The total effect of belief consistency (i.e., the difference in predicted argument quality ratings between a claim rated as *extremely true* and a claim rated as *extremely false*) was 1.57. As before, the magnitude of this effect was larger than the largest effect of argument type (i.e., the average difference in argument quality ratings between good arguments and authority-based arguments) which showed a 1.30 point difference.

There was not a consistent nor particularly interesting pattern of interaction between argument type and belief consistency. Details of this

[6] For the by-topic and by-participant random effects terms, the maximal model included random intercepts and random slopes for all main effects and correlations between the intercepts and slopes. The by-topic random effects terms also included random slopes for all interactions. As the maximal model showed convergence issues, we simplified it successively until there were no convergence issues (i.e., until the model no longer produced a singular fit). We arrived at a final model which employed by-participant and by-topic random intercepts and slopes that did not model the correlation between intercepts and slopes. By-participant random slopes were included for main effects of argument leaning and argument type. By-topic random slopes were included for the main effect of argument leaning. Full details can be found in the online supplemental materials on OSF.

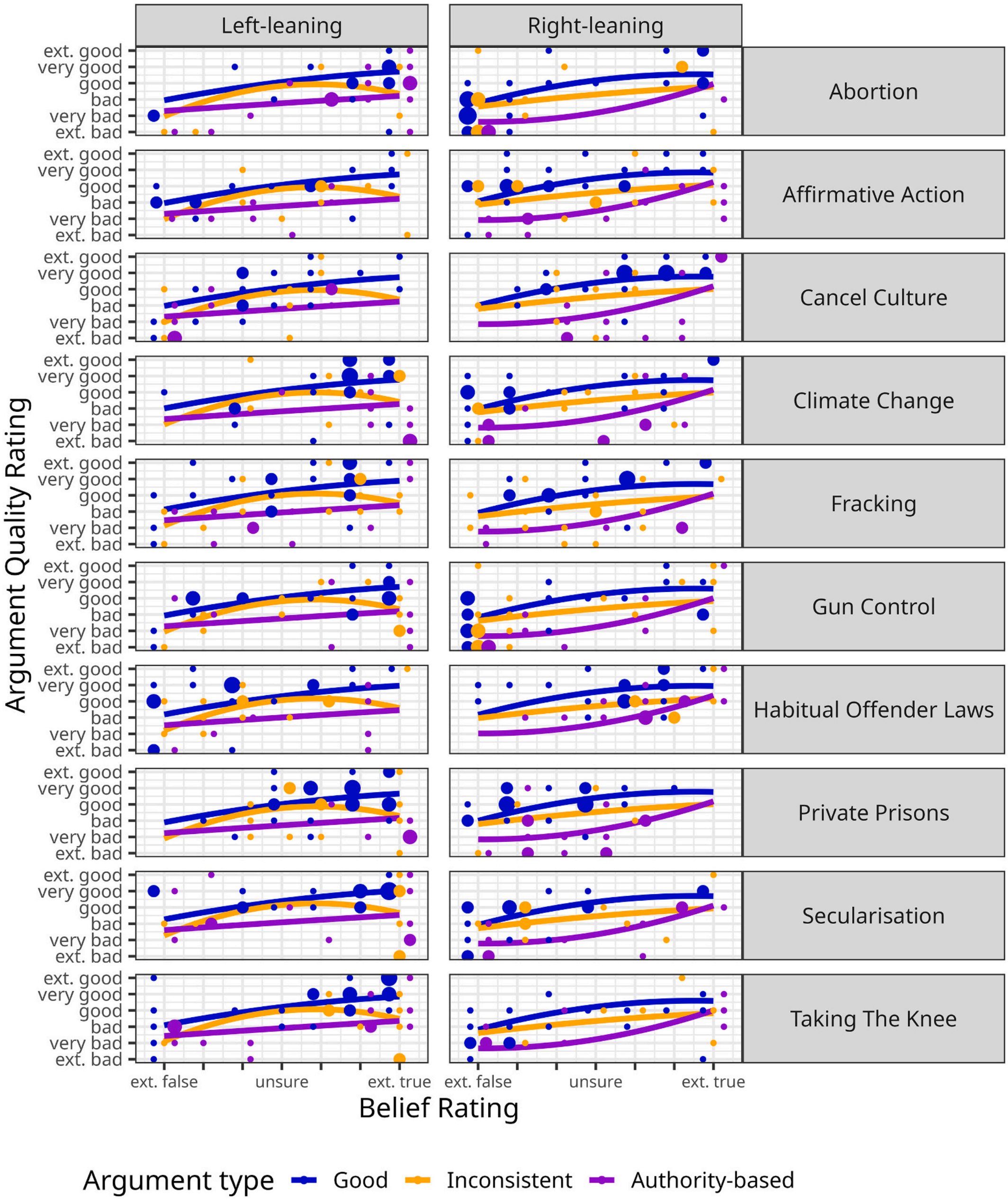


Fig. 9. Argument quality ratings as a function of belief consistency for each argument in Experiment 3.

Note. Results of Experiment 3 conditional on the topic and the level of argument support. Each line and colour in each panel shows responses to exactly one argument (i.e., there is no aggregation across items within a panel). The dots show individual responses and the curved lines show predictions from the linear mixed model. Blue dots represent argument quality ratings to good arguments, orange dots represent argument quality ratings to inconsistent arguments, and purple dots represent argument quality ratings to authority-based arguments. The size of the dots represents the number of argument quality rating responses for the corresponding belief rating, with larger dots representing a larger number of responses. Data points are dodged so that responses for good, inconsistent and authority-based arguments do not overlap. Model predictions are based on the fixed effects of the final model that includes the non-significant three-way interaction terms. Ext. = extremely.

interaction and how it was moderated by the effect of argument leaning are presented in Appendix D.

As with Experiments 1 and 2, we also see that the main patterns of interest are evident on a by-topic level as illustrated in Fig. 9. As with Experiments 1 and 2, the effect of belief consistency is evident for all topics. It is worth mentioning that in Experiment 3 where the manipulation of argument type is more clearly operationalised than the manipulation of argument quality in Experiments 1 and 2, the effect of argument type is more consistent across topics than the effect of argument quality is in the previous experiments. The consistent pattern of results across topics suggests that the main results are not an artifact of aggregating the data across stimuli.

5.1.5.3. Analysis of inconsistent arguments. One shortcoming of the results presented so far is that they do not establish a causal link from participants' beliefs to their argument quality ratings. The problem is that we did not manipulate participants' beliefs but only measured them. Thus, a possible alternative interpretation to the results presented so far is that different participants find different arguments differentially convincing (i.e., instead of the beliefs being responsible for the argument evaluations, the argument evaluations are responsible for the beliefs). For example, because all of the arguments are about well-known issues, most participants might have already seen the central points made in each argument and based their beliefs upon how convincing they found these points.

The inconsistent arguments of Experiment 3 provide us with a way of addressing the aforementioned shortcoming. Recall that the two different inconsistent arguments for each topic share the same content and pretty much exactly the same sentences and only differ in two aspects; the conclusion and the ordering of the sentences (see Table 3 for an example). For the left-leaning versions, each inconsistent argument begins with left-leaning point(s), followed by right-leaning point(s), followed again by left-leaning point(s), followed by the left-leaning conclusion. For the right-leaning inconsistent arguments the same points are made using the exact same phrasing, but in the inverted order: right-leaning point(s), followed by left-leaning point(s), followed by right-leaning point(s), and ending with a right-leaning conclusion. Thus, if what determines participants' argument quality ratings is how convincing they find these points independent of their beliefs regarding the claims, then we should not see a difference in participants' argument quality ratings for left-leaning versus right-leaning inconsistent arguments.

To check whether political leaning of the participants has an effect but not the political leaning of the claim, we reanalysed participants' responses to the inconsistent arguments. In all experiments presented in this paper, we asked participants basic demographic questions including a question about their political orientation (If you had to choose between Democrats and Republicans, how would you identify your political affiliation? 1 = Strongly Democrat, 7 = Strongly Republican) and a question about their self-reported social conservatism (In general, how liberal or conservative are you on social issues? 1 = Strongly Liberal, 7 = Strongly Conservative). In this analysis of Experiment 3, we categorised each participant's political leaning based on a composite *conservatism* score by summing their political orientation score with their self-reported conservatism score and then dividing the sum by two to get a combined average of both scores. After excluding participants who did not provide us with information on their political beliefs (1 participant), we were left with data from 53 left-leaning participants and 43 right-leaning participants.

We ran a linear mixed model with fixed effects of argument leaning (left-leaning vs right-leaning), conservatism (−3 to 3: most liberal - most conservative), conservatism squared (squared values of conservatism to test for quadratic effect), and all interactions with participants' argument quality rating of the inconsistent arguments as the dependent variable. We initially attempted to employ the maximal model justified by the design, which contained crossed random effects with by-participant and by-topic random terms. As the maximal model produced a singular fit, we simplified the model successively until this was not the case,[7] resulting in a final model that contained only by-participant random intercepts. The pattern of significant and non-significant results overall remained the same for all random effect structures tested.

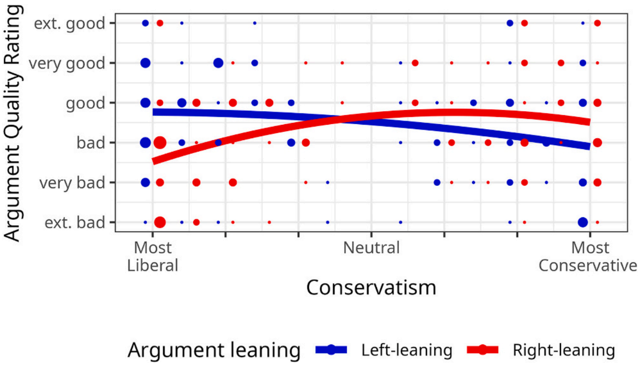


Fig. 10. Argument quality ratings of inconsistent arguments as a function of argument leaning and participant conservatism.
Note. Results of responses to inconsistent arguments in Experiment 3. The dots show individual responses and the curved lines show predictions from the linear mixed model. Blue dots represent argument quality ratings to left-leaning arguments and red dots represent argument quality ratings to right-leaning arguments. The size of the dots represents the number of argument quality rating responses for the corresponding belief rating, with larger dots representing a larger number of responses. Data points are dodged so that responses for left-leaning and right-leaning arguments do not overlap. Model predictions are based on the fixed effects of the final model. Ext. = extremely.

Fig. 10 shows the results of the mixed model on responses to the inconsistent arguments. The figure does not clearly show a main effect of conservatism or argument leaning, as points to the left of the figure are not clearly higher or lower than the points to the right of the figure, nor is one regression line in the figure clearly higher than the other. In line with this visual impression, the main effect of participant political orientation was not significant ($F(1, 93) = 0.05$, $p = .831$) nor was the main effect of argument leaning ($F(1, 93) = 0.15$, $p = .701$).

However, Fig. 10 does illustrate a conservatism by argument leaning interaction, as the regression lines in the figure cross over each other. In line with this visual impression, the participant political orientation by argument leaning interaction was significant, $F(1, 93) = 15.52, p < .001$. Each point in increased conservatism of a participant was associated with argument quality ratings for right-leaning arguments that were on average 0.17 (95 % CI [0.04, 0.29]) points higher, and with argument quality ratings for left-leaning arguments that were 0.15 (95 % CI [0.03, 0.26]) points lower. The main effect of conservatism squared was not significant (F(1, 93) = 1.90, $p = .171$), nor was its interaction with argument leaning ($F(1, 93) = 0.78$, $p = .379$). Together, this suggests that even for the inconsistent arguments that make the same points and use the same sentences, what matters is participants' beliefs regarding the claim and not the content of the argument itself.

[7] The maximal model as justified by the design contained all fixed effects, a by-topic random effect structure that replicated the fixed effects structure and by-participant intercepts. This model however had a singular/boundary fit estimation, so we successively simplified the model until it did not have a singular/boundary fit estimation (details can be found in the supplemental materials on OSF). The final model contained the aforementioned fixed effects and by-participant intercepts only.

5.1.6. Discussion

In Experiment 3 we were mainly interested in two questions; whether people preferred arguments with internal inconsistencies over those that were internally consistent but based on appeals to authority, and whether the perception of these inconsistent and authority-based arguments were equally influenced by participants' prior beliefs. The results from Experiment 3 suggest that participants have a preference for the inconsistent arguments over the authority-based arguments. We also replicated the pattern that, independent of argument type, argument evaluations were correlated with participants' prior beliefs.

We were initially surprised that participants on average preferred inconsistent arguments to consistent arguments based on appeals to authority, as the inconsistent arguments made little sense and were difficult to understand when examined closely (see examples in Table 3). We speculate that this preference might be because participants are willing to overlook inconsistencies in arguments, and prefer the causal/statistical evidence in the inconsistent arguments that is consistent with the arguments' conclusion over the evidence based on appeals to authority in the authority-based arguments. This is supported by existing literature suggesting that people value causal/statistical evidence (e.g., Hoeken, 2001; Hoeken & Hustinx, 2009; Slusher & Anderson, 1996; Tobin & Weary, 2008) and are also good at spontaneously explaining away inconsistencies (Khemlani & Johnson-Laird, 2012). An alternative explanation for this result could be that participants recognised the personalities in the authority-based arguments were not knowledgeable about the claims being made and downweighed the evidence from them accordingly (Harris et al., 2016).

The rest of the results from Experiment 3 were in line with what we expected given the results from Experiments 1 and 2. Good arguments were rated as better than both inconsistent and authority-based arguments, and participants on average rated arguments in line with their beliefs as being of better quality than arguments that were not. Finally, we again found that the maximum effect of belief (i.e., the difference in predicted argument quality ratings between a claim rated as *extremely true* and a claim rated as *extremely false*) was greater than the largest effect of argument quality (i.e., the average difference in argument quality ratings between good arguments and authority-based arguments), even though the magnitude of this difference was attenuated compared to Experiments 1 and 2.

Analysis of the inconsistent arguments in isolation also suggested that the results did not arise from inter-individual differences in the convincingness of arguments across participants. If participants had seen the arguments we presented them beforehand, and these arguments had in turn informed their beliefs, then we would expect participants with similar beliefs to rate essentially identical arguments (in which only the order of sentences is changed) similarly irrespective of how they were framed (i.e., as left-leaning vs right-leaning). Instead, we see that participants with similar beliefs rate essentially identical arguments differently when the political framing of these arguments also differs.[8]

6. General discussion

Our research question was concerned with the degree to which our prior beliefs influence the way we reason about everyday political arguments. In three studies, we asked participants to rate their belief about political claims (e.g., *Abortion should be legal in the US*) and their perception of the quality of good and bad arguments related to these claims. Results showed that participants could distinguish between good and bad arguments; their average argument quality ratings were higher for good compared to bad arguments. We also found that their argument quality ratings were highly correlated with their belief ratings. Importantly, the effect of belief was larger than the effect of argument quality – the difference in argument quality ratings on opposite ends of the belief scale was larger than the effect of the quality of the argument itself. In Experiment 3, we also found that participants are more sensitive to some flaws in arguments than they are to others. More specifically, participants thought arguments that were internally inconsistent were to an extent better than arguments that were based on appeals to authority, even though the inconsistent arguments made very little sense when looked at closely. Finally, Experiment 3 ruled out that the observed effects are primarily driven by participants' prior exposure to the information given in the presented arguments. For inconsistent arguments both left-leaning and right-leaning arguments made the same points using the same sentences, just in a different order and with a different conclusion. Still, we found what matters were participants' prior beliefs on the issues and not the arguments themselves.

The stimuli used in our Everyday Argument Assessment Task were both ecologically valid and well controlled. Participants rated claims about topics appealing to both sides of the political spectrum; half the topics they were asked about had left-leaning claims and the rest had right-leaning claims. This ensured that each participant saw a roughly even mixture of arguments that agreed with and that were at odds with what they believed. The manipulation of informal argument quality in Experiments 1 and 2 (inspired by the stimuli used in Hopkins et al., 2016), was also validated in a Pretest, demonstrating that the good and bad arguments used in our study differed in their informal quality in a way that participants could detect. Experiment 3 furthermore systematically manipulated the 'bad' arguments. As clear from the by-topic analysis, the results discussed are not stimulus specific, but can for the most part be seen for each topic.

We believe one of the main contributions of our study is showing that for people's perception of the quality of 'everyday' informal arguments, their prior beliefs play at least as large a role as what is said in the argument itself. This finding held across the three experiments reported in the present paper. However, it is clear this effect cannot hold universally. For example, if we added typographical or grammatical errors into the bad arguments, at some point their perceived quality would drop so far that the effect of quality would exceed the effect of belief. The problem with such a manipulation would be that it would remove the ecological validity of the bad arguments. Furthermore, even if we found that the effect of argument quality exceeded the effect of prior beliefs for some ecologically valid arguments, in our opinion this would not change our main message: If we ask people to objectively judge the quality of an argument, they cannot do so without their beliefs playing a major role.

We see our study as a contribution to the ongoing discussion on media literacy in a digital world. What our results essentially show is that the same piece of information – such as a newspaper article or social media post – can be interpreted very differently depending on someone's prior beliefs. While some accounts suggest this is the result of faulty reasoning (e.g., Aspernäs et al., 2023; Čavojová et al., 2018; Evans et al., 1983; Gampa et al., 2019; Lord et al., 1979), Hahn and Oaksford (2007) propose an account that explains this effect as a consequence of rational belief updating. In a Bayesian framework, the perceived quality of an argument is given by posterior beliefs that are a function of the prior belief a person holds and the quality of the evidence presented in the argument. People who start with a lower prior belief are therefore also expected to give a lower argument quality rating assuming they update their beliefs in a rational Bayesian manner. Similar updating models have even been shown to predict rational belief polarisation (e.g., Cook & Lewandowsky, 2016; Jern et al., 2014).

Further questions as to the mechanism of argument evaluation are also raised from Experiment 3, which gives some insight into the criteria by which people judge the quality of arguments. Despite the inconsistent arguments making little sense when examined properly, we found

[8] One may object to this interpretation on the basis that participants may have neglected the central (i.e., inconsistent) parts of the inconsistent arguments. Whereas this might be the case to some degree, participants still rated the inconsistent arguments as worse than the good arguments. Thus, to the degree that participants were able to distinguish good from inconsistent arguments, they must have read the full inconsistent arguments.

arguments with internal inconsistencies were rated as better on average than arguments without internal inconsistencies, but which were based only on appeals to authority. We speculate this could be because participants overlook the inconsistencies in the inconsistent arguments, but this raises further questions as to what exactly people tend overlook in arguments, what things people usually attend to and what is being retained from the arguments they look at.

One of the take-aways from our study is that the effect of belief-aligned argument evaluation needs to be taken into account when considering interventions targeted at reducing the negative effect of misinformation, a topic which has arguably received the largest attention in the study of media literacy (e.g., Lewandowsky et al., 2017; Pennycook & Rand, 2019; Van der Linden, 2022). For example, in their consensus statement on fighting health misinformation, the APA recommends people “Leverage trusted sources to counter misinformation and provide accurate [health] information" (Van Der Linden et al., 2023). In line with the Bayesian account of belief updating (Hahn & Oaksford, 2007) our results show that such interventions, which rely on people updating their beliefs about misinformation from accurate information, may have limited effects on those people whose beliefs are furthest away from the truth. However, effective interventions should ideally have the strongest effect on those most affected by misinformation. Another way to understand our findings is that the negative effect of misinformation is in principle smallest for those whose prior beliefs are already most strongly aligned with the accurate information. Thus, maybe little intervention is necessary in such cases.

Taking a broader perspective, we believe that addressing a complex real-world issue such as the ‘infodemic’ requires a comprehensive and multifaceted approach. The facet we are attempting to address here is trying to uncover the main factors, such as prior beliefs, affecting the degree with which people integrate new information into their belief systems. A better understanding of this fundamental question will enable better methods for combating the negative effects of misinformation in the future.

Finally, much of the research into misinformation seems to be squarely aligned with what Chater and Loewenstein (Chater & Loewenstein, 2023; see also Hagmann et al., 2023) call the i-frame, the idea that policy interventions, such as combating the effect of misinformation, should focus on the individual (e.g., Van Der Linden et al., 2023). Chater and Loewenstein contrast this with the s-frame, the idea that policy interventions need to change the underlying system. Given the overall rather modest effects of existing i-frame interventions targeting this issue, we concur with Chater and Loewenstein that combating misinformation requires s-frame interventions, such as tighter regulations of social media companies.

A starting point for s-frame interventions from our research could be the finding that even people who have relatively weak beliefs aligned with misinformation will be biased in their evaluation of information that is accurate. This suggests that the negative effect of misinformation might be particularly problematic in rapidly evolving situations when there is no existing prior belief, because of the potential that the initial misinformation biases how all subsequent information will be perceived (see also Pilgrim et al., 2024). Unfortunately, algorithms that are designed to increase engagement of users on social media platforms have the tendency to also proliferate misinformation (e.g., Menczer, 2021). As a case in point, after the Hamas attack on Israel on 7. October 2023, all major social media companies struggled to curb the spread of misinformation through their platforms (e.g., Gogarty et al., 2023; Milmo & O’Carroll, 2023). The spread of pro-Hamas information even resulted in a letter from Osama bin Laden justifying the 9/11 terror attacks to go viral, first on TikTok and then on X/Twitter (Montgomery, 2023). We believe that without sufficient pressure from policy makers, social media companies have no incentive to adapt their algorithms such that the spread of misinformation is curtailed for situations in which news is rapidly evolving.

7. Conclusions

In the Everyday Argument Assessment Task, participants did not evaluate the quality of real-world arguments independently of what they believed. This happened despite participants explicitly being told to evaluate the quality of each argument objectively. Participants were able to discriminate between good and bad arguments, and interestingly on average preferred inconsistent arguments compared to consistent arguments based on appeals to authority amongst the bad arguments they saw. However, the strongest effect was the tendency for participants to rate arguments as being of better quality when those arguments were also more in-line with their beliefs. Our findings suggest that people can interpret belief-consistent information very differently from belief-inconsistent information, which we highlight should be taken into account when coming up with potential interventions to reduce the spread of misinformation that is becoming ever-present in the digital age.

CRediT authorship contribution statement

Calvin Deans-Browne: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Henrik Singmann:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Funding

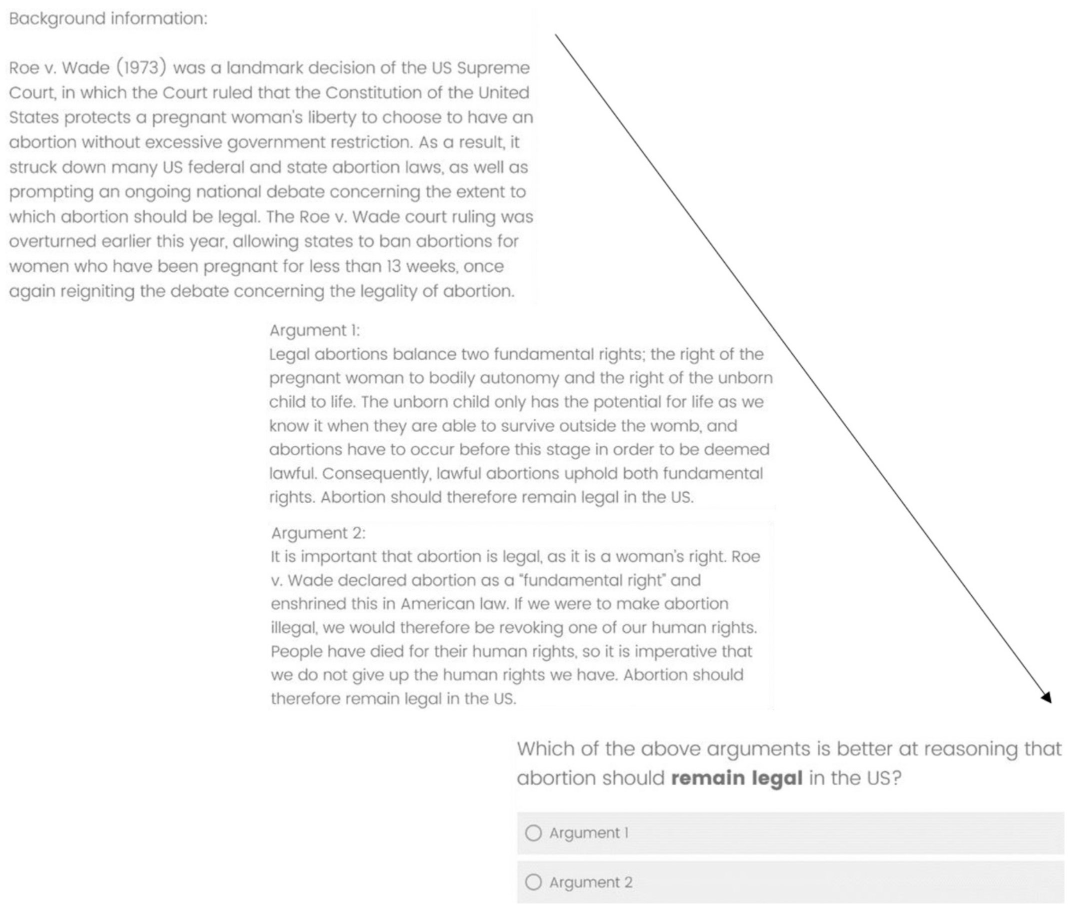
This research was supported by Grant 100014_179121 from the Swiss National Science Foundation (SNSF) to HS.

Declaration of competing interest

We have no conflicts of interests to disclose.

Appendix A. Pretest procedure

Pretest Procedure



Background information:

Roe v. Wade (1973) was a landmark decision of the US Supreme Court, in which the Court ruled that the Constitution of the United States protects a pregnant woman's liberty to choose to have an abortion without excessive government restriction. As a result, it struck down many US federal and state abortion laws, as well as prompting an ongoing national debate concerning the extent to which abortion should be legal. The Roe v. Wade court ruling was overturned earlier this year, allowing states to ban abortions for women who have been pregnant for less than 13 weeks, once again reigniting the debate concerning the legality of abortion.

Argument 1:
Legal abortions balance two fundamental rights; the right of the pregnant woman to bodily autonomy and the right of the unborn child to life. The unborn child only has the potential for life as we know it when they are able to survive outside the womb, and abortions have to occur before this stage in order to be deemed lawful. Consequently, lawful abortions uphold both fundamental rights. Abortion should therefore remain legal in the US.

Argument 2:
It is important that abortion is legal, as it is a woman's right. Roe v. Wade declared abortion as a "fundamental right" and enshrined this in American law. If we were to make abortion illegal, we would therefore be revoking one of our human rights. People have died for their human rights, so it is imperative that we do not give up the human rights we have. Abortion should therefore remain legal in the US.

Which of the above arguments is better at reasoning that abortion should **remain legal** in the US?

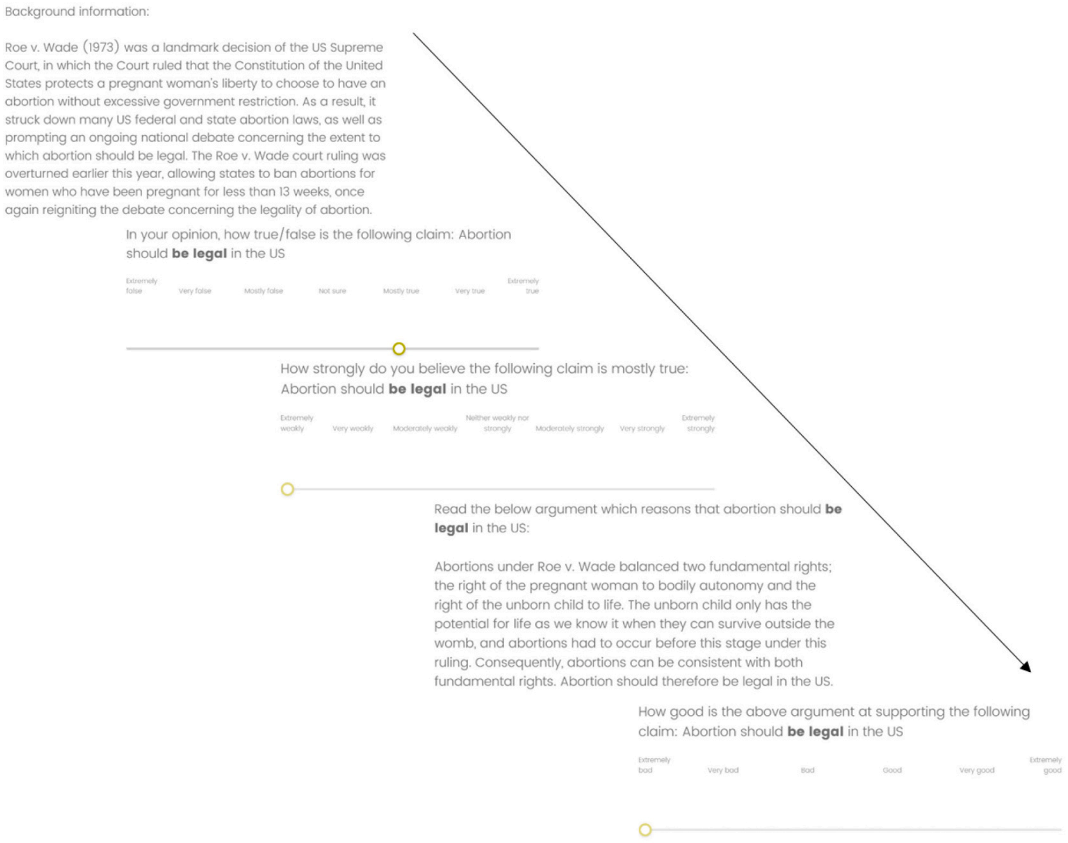
○ Argument 1

○ Argument 2

Note. Screenshots of the two-alternative forced-choice task in the Pretest. Participants were first shown a short passage that briefly described the context of the political issue of relevance. After the introductory paragraph, participants were immediately shown a good and a bad argument that either supported or challenged the same political claim (here a pro-good and a pro-bad argument) together as a pair. Participants' task was to select which of the arguments they thought was the better of the two.

Appendix B. Everyday argument assessment task procedure

Everyday Argument Assessment Task Procedure



Background information:

Roe v. Wade (1973) was a landmark decision of the US Supreme Court, in which the Court ruled that the Constitution of the United States protects a pregnant woman's liberty to choose to have an abortion without excessive government restriction. As a result, it struck down many US federal and state abortion laws, as well as prompting an ongoing national debate concerning the extent to which abortion should be legal. The Roe v. Wade court ruling was overturned earlier this year, allowing states to ban abortions for women who have been pregnant for less than 13 weeks, once again reigniting the debate concerning the legality of abortion.

In your opinion, how true/false is the following claim: Abortion should **be legal** in the US

Extremely false — Very false — Mostly false — Not sure — Mostly true — Very true — Extremely true

How strongly do you believe the following claim is mostly true: Abortion should **be legal** in the US

Extremely weakly — Very weakly — Moderately weakly — Neither weakly nor strongly — Moderately strongly — Very strongly — Extremely strongly

Read the below argument which reasons that abortion should **be legal** in the US:

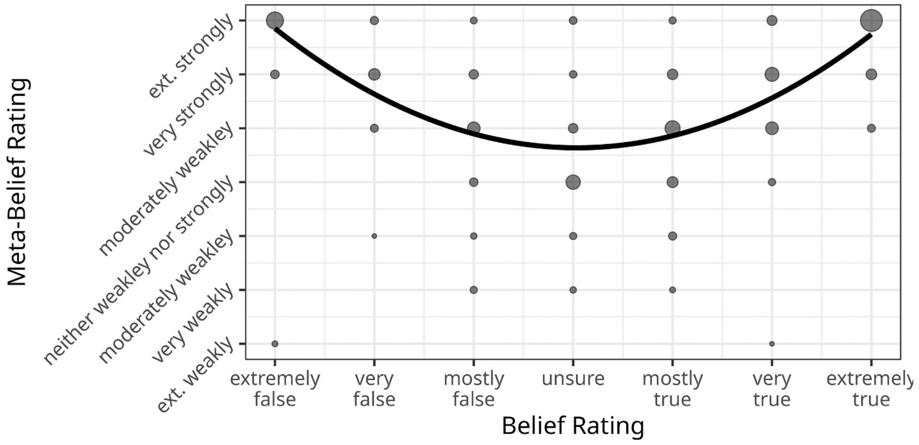
Abortions under Roe v. Wade balanced two fundamental rights; the right of the pregnant woman to bodily autonomy and the right of the unborn child to life. The unborn child only has the potential for life as we know it when they can survive outside the womb, and abortions had to occur before this stage under this ruling. Consequently, abortions can be consistent with both fundamental rights. Abortion should therefore be legal in the US.

How good is the above argument at supporting the following claim: Abortion should **be legal** in the US

Extremely bad — Very bad — Bad — Good — Very good — Extremely good

Note. Screenshots of the Everyday Argument Evaluation Task as used in Experiment 1. The Everyday Argument Evaluation Task consists of two parts. In the first part, participants first saw a short passage introducing them to a political topic and then had to provide their belief rating about a related statement (e.g., Abortion should be legal in the US) as well as their meta-belief rating (both on a 7-point scale; analysis of meta-belief ratings did not yield anything interesting so we exclude much reference to it throughout the manuscript). After making both ratings participants proceeded to the next topic. After making their belief ratings for all eight topics participants proceeded to the second part of the study in which they were asked to make argument quality ratings. Participants were shown one argument at a time (either a good or a bad argument for each claim in Experiments 1 and 2, or either a good, inconsistent or authority-based argument for each claim in Experiment 3) and had to rate the quality of this argument on a 6-point scale (*extremely bad – extremely good*).

Appendix C. Relationship between belief and meta-belief



Note. The grey dots show individual responses and the curved black line show predictions from the linear mixed model. The size of the dots represent the number of meta-belief rating responses for the corresponding belief judgement, with larger dots representing a larger number of responses. Model predictions are based on the fixed effects of the linear-mixed effects model reported in the main text. The u-shape of the prediction line indicates that there is a quadratic relationship between belief and meta-belief. Ext. = extremely.

In addition to the belief ratings, we collected a second rating to gauge participants' beliefs – the meta-belief ratings. The belief rating indicated how much participants believed in the truth of the claim, while the meta-belief rating indicated how strongly they held the corresponding beliefs. The reason for collecting both ratings was to see whether the strength of belief (as measured by the meta-belief rating) differed in a systematic and potentially meaningful manner across belief ratings.

The u-shaped pattern in the graph suggests that participants who had more extreme beliefs at either end of the belief scale tended to also feel more strongly about said beliefs. We considered the extent to which this relationship was quadratic by running a linear mixed effects model (e.g., Singmann & Kellen, 2019); meta-beliefs were predicted from fixed variables belief (continuous −3 to 3 scale) and belief squared (i.e., the quadratic effect of belief) using crossed random effects for participants and topics. The full model can be seen in the online supplemental materials.

As suggested by the figure, we found a significant quadratic effect of belief, $F(1, 13.00) = 697.69$, $p < .001$, but no linear effect of belief, $F(1, 5.73) = 0.47$, $p = .522$. This indicates that meta-belief ratings were smallest when belief ratings were at the midpoint of the scale and larger when belief ratings are further from the midpoint of the scale in either direction. Given this strong (quadratic) correspondence between meta-belief ratings and belief ratings, in the following we only focus on the belief ratings as our measure of participants' beliefs.

We also performed an exploratory analysis replacing the belief and belief squared ratings with meta-belief ratings for the fixed effects structure reported in Experiment 1. These analyses did not produce any noteworthy results. Full details of this model can also be found in the online supplemental materials.

Appendix D. Argument quality by belief consistency interaction analysis

Details of all following analysis can be found in the OSF that can be accessed using the following link: https://osf.io/f9h6a/

Interaction analysis of Experiment 1

Results regarding the belief consistency by argument quality interaction are less straightforward to infer from Fig. 3. There are at least two different possible data patterns that could result in an interaction. One possibility is that the ability to discriminate between good and bad arguments is associated with the (linear) strength of belief consistency. This pattern would be illustrated in Fig. 3 if the two prediction lines for good and bad arguments in both the pro and con grids converged at just one end of the belief scale. We do not see this in Fig. 3, as despite some suggestion of the prediction lines converging at the high end of the scale for con arguments, the prediction lines for both pro and con arguments are mostly parallel. In line with this visual assessment, the three-way interaction between belief, argument support, and argument quality was not significant, $F(1, 435.24) = 0.56$, $p = .456$.

A second possibility for the interaction is that participants' ability to discriminate between good and bad arguments depends on the extremity of their beliefs. In other words, we would expect a quadratic effect of belief on discriminability between good and bad arguments such that discriminability is worse when belief consistency is either extremely high or extremely low. This pattern would be illustrated in Fig. 3 if the difference in argument quality ratings between good and bad arguments for both pro and con arguments was largest at the centre of the belief scale and smallest at the edges of the scale. This also does not appear to be the case, given that the prediction lines for both pro and con arguments in Fig. 3 seem mostly parallel. In line with this, neither the squared belief by argument quality interaction, $F(1, 544.39) = 0.13$, $p = .722$, nor the squared belief by argument support by argument quality interaction, $F(1, 603.90) = 0.39$, $p = .532$, were significant.

Interaction analysis of Experiment 2

Inspection of Fig. 5 suggests that there might be a belief consistency by argument quality interaction. For pro arguments, discriminability (i.e., the difference in argument quality ratings between good and bad arguments) seems smaller at both ends of the belief scale compared to the middle. In contrast, for the con arguments, discriminability seems smaller at the right end of the belief scale. In line with these visual patterns, we see a significant three-way interaction between belief, argument support, and argument quality, $F(1, 406.05) = 4.66$, $p = .031$, but no significant three-way interaction between belief squared by argument support by argument quality, $F(1, 902.77) = 2.20$, $p = .138$. Furthermore, the p-value of the belief squared by argument quality interaction is above but near the .05 threshold, $F(1, 7.44) = 4.85$, $p = .061$. While this pattern appears to provide some evidence for a belief consistency by argument quality interaction, none of these three interactions is significant in the maximal model justified by the design (all $ps >$ .07). This indicates that the evidence for such interactions was very weak. Importantly, none of the interactions involving argument order reached significance ($ps > .07$).

Given the significant linear belief consistency by argument quality interaction and nearly significant quadratic belief consistency by argument quality interaction, we performed exploratory analyses investigating the interaction between good and bad arguments for both pro and con arguments. Our initial prediction was that for both pro and con arguments, participants should be worse at discriminating between good and bad arguments when those arguments are less aligned with what the participants believe (in other words, people do not attend to the quality of the argument when they are already in agreement with what the argument has to say). However, when we compare the slopes between good and bad arguments individually for both pro and con arguments, we see that the effect of belief on good and bad arguments are not significantly different (*slope difference* $= 0.026$, $t(9.91) = 0.48$, $p = .641$ for pro arguments; *slope difference* $= -0.097$, $t(9.47) = -1.84$, $p = .098$ for con arguments). We also theorized that belief might have a quadratic effect on argument quality, where participants with strong beliefs at either end of the scale are worse at differentiating good and bad arguments compared to participants who hold more neutral beliefs towards the middle of the scale. We found this pattern of results (significant belief squared by argument quality interaction) was consistent for pro arguments ($t(22.0) = -2.64$, $p = .015$) but not for con arguments ($t(21.2) = -0.71$, $p = .485$).

In Fig. 6, we can see that the difference in argument quality ratings between good and bad pro arguments is greater for participants in the middle of the scale compared to participants at either end of the scale, in line with our exploratory analysis of the quadratic interaction. For some of the con items, we see a pattern indicative of a linear interaction (where participants rate good and bad arguments more similarly at one edge of the scale compared to the other), though we did not see any support for this in our exploratory analysis.

Interaction analysis of Experiment 3

We tested for an interaction between argument type and belief consistency, similar to the interaction pattern between argument quality and belief consistency we tested for in Experiments 1 and 2. The interaction between argument type and belief consistency was not significant, $F(2, 599.75) =$ 1.31 $p = .270$), though the three way interaction between argument leaning, argument type, and belief consistency was, $F(2, 440.11) = 3.32$ $p = .037$). Importantly, all six slopes (i.e., for the three argument types in the two argument-leaning conditions) were significantly positive ($ps < .01$). To check for the source of the interaction we compared the three slopes within each argument leaning condition to each other (without controlling for multiple testing). For left-leaning arguments, only the slopes for good and authority-based arguments differ significantly from each other (*difference* $= 0.14$, $t(659) = 2.08$, $p = .038$), remaining $ps > .341$. For right-leaning arguments, only the slopes for inconsistent and authority-based arguments significantly differed from each other (*difference* $= 0.19$, $t(485) = -2.31$, $p = .021$), remaining $ps > .187$.

We also found some evidence for a quadratic – that is, non-linear – effect of belief. The argument type by belief squared interaction was significant ($F(2, 615.13) = 5.36$, $p = .005$). It is difficult to judge exactly what is driving this effect from Fig. 7, as the quadratic pattern of results is not consistent across levels of argument leaning (e.g., the quadratic effect for inconsistent and authority-based arguments looks especially different between the left-leaning and right leaning arguments). The marginal quadratic effects suggests the quadratic effect is significantly negative for both good arguments ($\beta = -0.04$, $t(636) = -2.42$, $p = .016$) and inconsistent arguments ($\beta = -0.07$, $t(516) = -3{,}13$, $p = .002$), but not significantly different from 0 for authority-based arguments ($\beta = 0.03$, $t(661) = 1.32$, $p = .186$). Furthermore, the difference in quadratic estimates is significant for the comparison between good and authority-based arguments (*coefficient difference* $= -0.08$, $t(705) = -2.48$, $p = .026$), and between inconsistent and authority-based arguments (*coefficient difference* $= -0.11$, $t(546) = -3.15$, $p = .005$) but not between good and inconsistent arguments (*coefficient difference* $= 0.03$, $t(662) = 1.08$, $p = .282$; Holm adjusted for three comparisons).

The argument leaning by argument type by belief squared interaction was significant in the final model, $F(2, 653.17) = 4.02$, $p = .018$, but not the maximal model, $F(2, 10.19) = 2.33$, $p = .146$). Looking at the marginal effects, we can see that the quadratic effect of belief in the final model for left-leaning arguments is only (negatively) significant for inconsistent arguments ($\beta = -0.13$, $t(585) = -4.10$, $p < .001$) and for right-leaning arguments is only negatively significant for authority-based arguments ($\beta = -0.15$, $t(646) = -2.32$, $p = .020$) and only positively significant for authority-based arguments ($\beta = 0.07$, $t(737) = 1.97$, $p = .049$). Full analysis can be found in the supplemental materials on OSF.

Summary

We did not find evidence that the effect of argument quality was moderated by participants' belief ratings in Experiment 1. We did find weak evidence of this interaction in Experiment 2. Furthermore, the pattern of the interaction observed in Experiment 2 differed between pro and con arguments. For pro arguments the interaction appeared in quadratic form – discriminability was worse at the edges compared to the midpoint of the belief scale. For con arguments it looked linear for some topics – discriminability was better when arguments were more consistent with participants' beliefs. Overall, there was little evidence for any type of consistent interaction pattern, which appears to be in line with the results from several other studies investigating similar interactions between belief and argument quality for informal arguments (e.g., McCrudden et al., 2017; Thompson et al., 2012; Wolfe & Kurby, 2017). For Experiment 3, we found some evidence of an interaction, though nothing that was consistent for both left-leaning and right-leaning arguments.

Appendix E. Number of responses for each argument

Topic	Number of Responses					
	Experiment 1		Experiment 2		Experiment 3	
	Pro	Con	Pro	Con	Left-leaning	Right-leaning
Abortion	50	51	98	93	37	43
Cancel culture	45	56	100	91	38	42
Climate change	45	56	105	86	39	41
Fracking	46	55	96	95	47	38
Gun control laws	58	43	95	96	39	47
Habitual Offender laws	53	48	91	100	40	38
Kneeling during the national anthem	55	46	81	110	49	32
Private prisons	52	49	98	93	45	41
Affirmative action	–	–	–	–	31	45
Secularisation	–	–	–	–	39	37

Data availability

The data and supplementary materials are available via the following OSF link: https://osf.io/f9h6a/.

References

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

Aspernäs, J., Erlandsson, A., & Nilsson, A. (2023). Motivated formal reasoning: Ideological belief bias in syllogistic reasoning across diverse political issues. *Thinking & Reasoning, 29*(1), 43–69. https://doi.org/10.1080/13546783.2022.2038268

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Bonomi, G., Gennaioli, N., & Tabellini, G. (2021). Identity, beliefs, and political conflict. *The Quarterly Journal of Economics, 136*(4), 2371–2411. https://doi.org/10.1093/qje/qjab034

Borges Do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization, 100*(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Čavojová, V., Šrol, J., & Adamus, M. (2018). My point is valid, yours is not: Myside bias in reasoning about abortion. *Journal of Cognitive Psychology, 30*(7), 656–669. https://doi.org/10.1080/20445911.2018.1518961

Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences, 46*, Article e147. https://doi.org/10.1017/S0140525X22002023

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences, 118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science, 8*(1), 160–179. https://doi.org/10.1111/tops.12186

Deans-Browne, C. C. J. L., Baitanu, A., Dubinska, Y., & Singmann, H. (2024). Inconsistent Arguments are Perceived as Better Than Appeals to Authority: An Extension of the Everyday Belief Bias. In *, 46. Proceedings of the Annual Meeting of the Cognitive Science Society*.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*(1), 5.

Evans, J. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*(3), 295–306. https://doi.org/10.3758/BF03196976

Gampa, A., Wojcik, S. P., Motyl, M., Nosek, B. A., & Ditto, P. H. (2019). (Ideo)logical reasoning: Ideology impairs sound reasoning. *Social Psychological and Personality Science, 10*(8), 1075–1083. https://doi.org/10.1177/1948550619829059

Gogarty, K., Geonzon, J., Winstanley, J., & Carter, C. (2023, October 9). *Musk's X allows misinformation about Hamas' war on Israel to proliferate*. Media Matters for America. https://www.mediamatters.org/twitter/musks-x-allows-misinformation-about-hamas-war-israel-proliferate.

Hagmann, D., Liao, Y., Chater, N., & Loewenstein, G. (2023). *Costly distractions: Focusing on individual behavior undermines support for systemic reforms (SSRN scholarly paper 4426034)*. https://doi.org/10.2139/ssrn.4426034

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*(3), 704–732. https://doi.org/10.1037/0033-295X.114.3.704

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science, 40*(6), 1496–1533. https://doi.org/10.1111/cogs.12276

Henderson, J. (2021). Truth and Gradability. *Journal of Philosophical Logic, 50*(4), 755–779. https://doi.org/10.1007/s10992-020-09584-3

Hoeken, H. (2001). Anecdotal, statistical, and causal evidence: Their perceived and actual persuasiveness. *Argumentation, 15*(4), 425–437. https://doi.org/10.1023/A:1012075630523

Hoeken, H., & Hustinx, L. (2009). When is statistical evidence superior to anecdotal evidence in supporting probability claims? The role of argument type. *Human Communication Research, 35*(4), 491–510. https://doi.org/10.1111/j.1468-2958.2009.01360.x

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition, 155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review, 121*(2), 206–224. https://doi.org/10.1037/a0035941

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139*(3), 486–491. https://doi.org/10.1016/j.actpsy.2012.01.010

Lewandowsky, S., Armaos, K., Bruns, H., Schmid, P., Holford, D. L., Hahn, U., … Cook, J. (2022). When science becomes embroiled in conflict: Recognizing the public's need for debate while combating conspiracies and misinformation. *The Annals of the American Academy of Political and Social Science, 700*(1), 26–40. https://doi.org/10.1177/00027162221084663

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

McCrudden, M. T., Barnes, A., McTigue, E. M., Welch, C., & MacDonald, E. (2017). The effect of perspective-taking on reasoning about strong and weak belief-relevant arguments. *Thinking & Reasoning, 23*(2), 115–133. https://doi.org/10.1080/13546783.2016.1234411

Menczer, F. (2021, September 20). *Facebook's algorithms fueled massive foreign propaganda campaigns during the 2020 election – Here's how algorithms can manipulate you*. The Conversation. http://theconversation.com/facebooks-algorithms-fueled-massive-foreign-propaganda-campaigns-during-the-2020-election-heres-how-algorithms-can-manipulate-you-168229.

Milmo, D., & O'Carroll, L. (2023). *Social media urged to act on violent content after Hamas attack*. The Guardian. https://www.theguardian.com/media/2023/oct/11/social-media-urged-to-act-on-violent-content-after-hamas-attack.

Montgomery, B. (2023). *TikTok 'aggressively' taking down videos promoting Bin Laden 'letter to America.'*. The Guardian. https://www.theguardian.com/technology/2023/nov/16/tiktok-bin-laden-letter-to-america-videos-removal.

Pennycook, G., Bago, B., & McPhetres, J. (2023). Science beliefs, political ideology, and cognitive sophistication. *Journal of Experimental Psychology: General, 152*(1), 80.

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2022). Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning. *Personality and Social Psychology Bulletin, 48*(5), 750–765. https://doi.org/10.1177/01461672211023652

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

Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. (2024). Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition, 245*, Article 105693. https://doi.org/10.1016/j.cognition.2023.105693

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. Spieler, & E. Schumacher (Eds.), *New methods in cognitive psychology* (1st ed., pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405-2.

Slusher, M. P., & Anderson, C. A. (1996). Using causal persuasive arguments to change beliefs and teach new information: The mediating role of explanation availability and evaluation bias in the acceptance of knowledge. *Journal of Educational Psychology, 88*(1), 110.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Thompson, V., Evans, J. S., & B. T.. (2012). Belief bias in informal reasoning. *Thinking & Reasoning, 18*(3), 278–310. https://doi.org/10.1080/13546783.2012.670752

Tobin, S. J., & Weary, G. (2008). The effects of causal uncertainty, causal importance, and initial attitude on attention to causal persuasive arguments. *Social Cognition, 26* (1), 44–65. https://doi.org/10.1521/soco.2008.26.1.44

Van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine, 28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

Van Der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2023). *Using psychological science to understand and fight health misinformation: An APA consensus statement: (506432023-001)*. https://doi.org/10.1037/e506432023-001

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wolfe, M. B., & Kurby, C. A. (2017). Belief in the claim of an argument increases perceived argument soundness. *Discourse Processes, 54*(8), 599–617. https://doi.org/10.1080/0163853X.2015.1137446

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet, 395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X