MCI MANAGEMENT CENTER INNSBRUCK
The Entrepreneurial School® | Die Unternehmerische Hochschule®
6020 Innsbruck, Austria, Universitaetsstrasse 15
Tel. +43 512 2070, office@mci.edu, www.mci.edu

Innsbruck, November 2, 2021

Dear reviewers, dear experts of the Jubiläumfond,

Below, you can find the copy of our manuscript *Correction of misinformation: Developing and comparing techniques to mitigate the belief perseverance bias* submitted for consideration to the European Journal of Operational Research. The manuscript is currently in the second round of review. It is a hidden copy made available only for the purpose of evaluating our research proposal *Mitigating the negative impact of misinformation on individual investors* submitted to the Jubiläumsfond of the OeNB.

Please use the manuscript only for the purposes of evaluating our research proposal and do not distribute the manuscript.

Thank you very much.

With best regards,

Prof. PD Dr. habil. Johannes Siebert
Phone: +43 512 2070 -3138, Fax: -3199
johannes.siebert@mci.edu

Correction of misinformation: Developing and comparing techniques to mitigate the belief perseverance bias

Jana Siebert

Palacky University Olomouc, Faculty of Arts, Department of Applied Economics, 771 80 Olomouc, Czech Republic, jana.siebert@upol.cz

and

Johannes Ulrich Siebert

Management Center Innsbruck, Department Business and Management, 6020 Innsbruck, Austria, Johannes.Siebert@mci.edu (corresponding author)

Abstract

The spread and influence of misinformation have become a matter of concern in society. Research has shown that simple retraction of misinformation is not sufficient to eliminate its influence on individuals. A reason for the failure of simple retractions is the belief perseverance bias. If the opinion or preferences of decision makers are biased by misinformation, decision-support methods cannot be effective in identifying optimal decisions. Thus, the belief perseverance bias and therewith misinformation negatively impact the decision quality of individuals as well as of organizations. However, the research on mitigating the belief perseverance bias after the retraction of misinformation has been limited. Only a few techniques for mitigating the bias have been proposed, and research on comparing various techniques in terms of their effectiveness has been scarce. Moreover, the practical applicability of these techniques is limited. This paper contributes to the research on mitigating the belief perseverance bias after the retraction of misinformation. We propose two debiasing techniques, counter-speech and awareness training, with a higher potential for practical applicability than the existing debiasing techniques. In an experiment, we compare the techniques with the previously proposed counter-explanation debiasing technique and show that all three debiasing techniques mitigate the belief perseverance bias. Moreover, the counter-speech technique performs considerably better in terms of effectiveness than the awareness-training and counter-explanation techniques. By debiasing decision makers' opinion, the proposed techniques help the decision makers become aware of their true preferences, thus increasing the effectiveness of decision-support methods and thereby the decision quality.

Keywords: Behavioural OR, misinformation, belief perseverance bias, debiasing, awareness training, counter-speech, counter-explanation

1 Introduction

Decision making shapes important outcomes for individuals, organizations, and society (Keeney, 1996; Milkman, Chugh, & Bazerman, 2009; Siebert & Keeney, 2015). The discipline of OR focuses on developing and applying highly sophisticated methods to facilitate complex decision making, improve decision quality, and therewith reach better outcomes for individuals (see, e.g., Siebert, Kunz, & Rolf, 2020) as well as for organizations (see, e.g., Badunenko, Kumbhakar, & Lozano-Vivas, 2021; Hübner et al., 2021; Nikolopoulos, Punia, Schäfers, Tsinopoulos, & Vasilakis, 2021). However, decision-support methods can reveal their full potential only when decision makers are aware of their true preferences, i.e. when they are not influenced by biases (Kahneman, 2011). The emerging field of behavioural OR has studied behavioural aspects related to problem-solving and decision support, and a particular focus has been on mitigating cognitive biases (Montibeller & Winterfeldt, 2015). Recently, Lahtinen, Hämäläinen, and Jenytin (2020) designed an approach to mitigate the overall effect of biases in a preference elicitation process. At the same time, they noted that "Yet, preference elicitation is only one phase in the overall decision analysis process. In practice, it is important to pay attention and manage behavioral phenomena in the entire process." (Lahtinen et al., 2020, p. 208). This paper contributes to bias mitigation in the phases of gathering relevant information and forming preferences, particularly in the presence of misinformation.

Biases make people vulnerable to misinformation (see, e.g., Kai Shu, Suhang Wang, Dongwon Lee, & Huan Liu, 2020), and misinformation, in turn, influences peoples' opinion, preferences, and consequentially their decisions (see, e.g., Lewandowsky, Ecker, & Cook, 2017). Misinformation has always been a part of our society. However, the internet and the rise of social media platforms have facilitated its spread. According to the Eurobarometer on fake news and online disinformation (European Commission, 2018b), 37% of the respondents come across fake news every or almost every day, and 83% of the respondents believe that fake news represents a danger to democracy. Misinformation thus can have severe consequences for individuals, organizations, and society. Salient examples are the decisions related to the COVID-19 pandemic (Pennycook, McPhetres, Zhang, Lu, & Rand, 2020; Roozenbeek et al., 2020; Tasnim, Hossain, & Mazumder, 2020; van der Linden, Roozenbeek, & Compton, 2020), climate change (Farrell, 2019; Lawrence & Estow, 2017; Treen, Williams, & O'Neill, 2020), Brexit (Watson, 2018), or the 2016 US presidential election (Bovet & Makse, 2019; Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019).

The term misinformation used in the context of this paper broadly refers to information that is initially presented as true but later appears to be false, regardless of intent to mislead. Thus, misinformation in this context covers everything from timely news coverage of unfolding events requiring occasional corrections of earlier statements (with no intention to mislead the news consumers) over fake news (intentionally designed to mislead the news consumers) to retracted research papers (for reasons ranging from concerns over the quality of the data to fabrication).

There is a clear consensus about the need to tackle misinformation. The European Commission has developed an action plan to proactively address misinformation and protect European Union's democratic system (European Commission, 2018a). The action plan involves detecting misinformation, raising awareness and improving societal resilience, and mobilizing the private sector to tackle misinformation. Reisach (2021) proposed a responsibility-based approach for social media platforms to counter misinformation. Numerous fact-checking organizations aiming at promoting veracity and correctness of reporting have emerged in recent years (Graves & Cherubini, 2016), and the Retraction Watch has been founded to report on retractions of scientific papers (Marcus & Oransky, 2014).

The research shows that simple retraction or correction of misinformation is insufficient to eliminate its influence; misinformation may continue to influence our judgment and reasoning even after it has been retracted or discredited. Lewandowsky, Ecker, Seifert, Schwarz, and Cook (2012) provide a review of cognitive factors that make the retraction or correction of misinformation at the individual level difficult. The reasons for the failure of simple retractions are, among others, the belief perseverance bias and the continued influence effect (Johnson & Seifert, 1994). The belief perseverance bias is the tendency to persevere in beliefs or opinions even after the initial information on which the beliefs or opinions were based has been discredited (Anderson, 2007), while the continued influence effect consists in persistent reliance on information even after it has been discredited (Johnson & Seifert, 1994).

The research has mainly focused on the continued influence effect of misinformation and its mitigation (Connor Desai, Pilditch, & Madsen, 2020; Ecker, Lewandowsky, & Tang, 2010; Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017; Johnson & Seifert, 1994; Lewandowsky et al., 2012; Seifert, 2002), while the research on the belief perseverance bias has been quite limited. The research on the belief perseverance bias has mainly focused on demonstrating the bias in an experimental setting and studying underlying mechanisms (see, e.g., Anderson, 1983, 1989; Anderson, Lepper, & Ross, 1980; Anglin, 2019; Green & Donahue, 2011; Maegherman, Ask, Horselenberg, & van Koppen, 2021). Although several techniques to mitigate the belief perseverance bias have been introduced and their effectiveness tested in experiments (see, e.g., Anderson, 1982; Anderson & Sechler, 1986; Lord, Lepper, & Preston, 1984), the research on comparing various debiasing techniques in terms of their effectiveness in mitigating the belief perseverance bias is scarce. The only and limited comparison of two debiasing techniques has been found in Anderson (1982). Furthermore, the practical applicability of the existing debiasing techniques is limited, particularly in the context of misinformation in the general public.

This paper contributes to the research on techniques to mitigate the belief perseverance bias after retraction of misinformation. In Sec. 1.1, we briefly review three main categories of techniques to tackle misinformation in general. Afterwards, in Sec. 1.2, we focus on the belief perseverance bias. Namely, we briefly discuss its relation to other biases and review relevant techniques designed to mitigate the belief perseverance bias and techniques designed to mitigate other biases, but with the potential to be adaptable to the belief perseverance bias. In Sec. 1.3, we turn our focus to the experimental design. Namely, we briefly review the experimental design commonly used in research on the belief

perseverance bias, identify its disadvantages, and describe the experimental design used in our studies. In Sec. 2 and Sec. 3, we describe our two studies. In Study 1, which serves as a preparatory study, we develop and validate measures of participants' opinion on a particular topic and two manipulation treatments for biasing participants' opinion on the topic and inducing belief perseverance. In Study 2, we develop two debiasing techniques and compare them with an existing debiasing technique in terms of their effectiveness in mitigating the belief perseverance bias. The studies were approved by the ethics committee of the Management Center Innsbruck. In Sec. 4, we discuss the results and provide directions for further research. In Sec. 5, conclusions are made.

1.1 Techniques to tackle misinformation

Numerous techniques to improve the effectiveness of retractions of misinformation have been explored. Lewandowsky et al. (2012) distinguished three main categories of successful techniques: (a) warnings at the time of the initial exposure to misinformation, (b) repetition of the retraction, and (c) corrections telling an alternative story that fills the coherence gap otherwise left by the retraction.

Up-front warnings and inoculation

Ecker et al. (2010) showed that explicit up-front warnings specifically explaining the continued influence effect are more effective in reducing the continued influence of information after retraction than general warnings (such as that the information is sometimes not double-checked before the release). A particular subcategory of techniques belonging to up-front warnings that appear to be effective in reducing the effect of misinformation is inoculation techniques. Inoculation consists in warning people that the information to be presented might be misleading and exposing them to particular examples of how they may be misled. A review of promising inoculation techniques to prevent misinformation is provided by Lewandowsky and van der Linden (2021).

Up-front warning or inoculation can be beneficial in specific situations, such as in a court setting, where jurors are often asked to disregard a piece of information they have heard (Lewandowsky et al., 2012). However, the practical applicability of up-front warnings and inoculation in the context of misinformation in the general public seems to be limited. Firstly, providing standardized up-front warnings with every single piece of information to be published (such as news or research articles) would eventually lose its desired effect as the individuals would get immune to these warnings after being exposed to them repeatedly. Similarly, providing inoculation with every single piece of information to be published would not be efficient as creating inoculation text for every single news or research article to be published is simply not feasible. Secondly, providing an up-front warning or inoculation only with "suspicious" content is difficult as well, as the information about the potential incorrectness or falsity is usually not available at the time of publication. A solution to these problems might be to inoculate the public against the manipulation techniques used to misinform in general instead of designing inoculation for a specific content of the information to be presented (Lewandowsky & van der Linden, 2021). An example of such a real-world application is the online fake news inoculation game *Bad News* introduced by Roozenbeek and van der Linden (2019), in which the players learn

through play about techniques commonly used to produce misinformation. Another solution, useful particularly in the context of fake news, might be to identify up-front the topics susceptible to misinformation and use the up-front warnings and inoculation with these topics. For example, van der Linden et al. (2020) applied inoculation to one such topic - the COVID-19 pandemic. There have already been efforts to identify topics susceptible to misinformation automatically. For example, Del Vicario, Quattrociocchi, Scala, and Zollo (2019) proposed a methodology to identify future fake news topics and validated this methodology on a Facebook dataset by identifying such topics with 77% accuracy. Zhang, Gupta, Kauten, Deokar, and Qin (2019) proposed a text analytics-driven methodology to detect fake news and validated it by achieving 92% classification accuracy with a novel detection system.

Facebook incorporated warnings against misinformation by flagging fake news in 2016 but stopped only one year later after discovering that the "fake news" flag not only did not have the intended effect but was sometimes even backfiring (Meixler, 2017). Indeed, Moravec, Minas, and Dennis (2018) found out that flagging headlines as fake does not affect peoples' judgments about truth. Further, they found out that people are more likely to believe news headlines that are in agreement with their opinions, thus confirming that processing of (fake) news relies on confirmation bias.

Repetition of the retraction

Lewandowsky et al. (2012) recommend using repeated retractions to mitigate the influence of misinformation. Such retractions seem to be helpful, particularly when the misinformation was repeatedly encoded (Ecker, Lewandowsky, Swire, & Chang, 2011). At the same time, Lewandowsky et al. (2012) warn that repeated retractions may paradoxically have an opposite effect. Indeed, repeating the correction may reduce people's confidence in its veracity (Bush, Johnson, & Seifert, 1994). Repeating the original misinformation in retractions could even cause a backfire effect (Schwarz, Sanna, Skurnik, & Yoon, 2007).

Corrections telling an alternative story

Several studies have shown that providing an alternative explanation for why the original information was incorrect reduces the continued influence of misinformation (see, e.g., Johnson and Seifert (1994)). At the same time, simple alternative explanations are generally preferred over complex ones (Lombrozo, 2007). Indeed, providing too many counter-arguments or asking people to think of too many possible counter-arguments may even backfire (Sanna, Schwarz, & Stocker, 2002).

1.2 Belief perseverance bias and approaches to its mitigation

Belief perseverance bias belongs to the group of motivational biases, i.e., biases "in which judgments are influenced by the desirability or undesirability of events, consequences, outcomes, or choices" (Montibeller & Winterfeldt, 2015, p. 1231). Belief perseverance bias is in Encyclopedia of Social Psychology defined as "the tendency to cling to one's initial belief even after receiving new information that contradicts or disconfirms the basis of that belief" (Anderson, 2007, p. 109). There is a close connection of the belief perseverance bias to the confirmation bias (Maegherman et al., 2021; Nickerson, 1998) and the myside bias (Perkins David, 1989). Confirmation bias occurs when there is a desire to

confirm one's belief, leading to unconscious selectivity in the acquisition and use of evidence (Nickerson, 1998). Myside bias consists in generating reasons or arguments consistent with one's belief (Perkins David, 2019). Nisbett and Ross (1980) argue that after creating a hypothesis based on received feedback, people may be prompted to search (selectively) for additional evidence confirming the hypothesis. When the original feedback on which the hypothesis was created is discredited, people may still persevere in their belief resting on the evidence (selectively) found in support of it.

Only a few techniques for mitigating the belief perseverance bias have been proposed. Nevertheless, also techniques originally developed to mitigate other biases can be applied (with appropriate adaptions) to the belief perseverance bias. In the following, we briefly review the most relevant debiasing techniques and discuss their practical applicability.

Counter-explanation

Anderson (1982) argued that belief perseverance might be mitigated by making eminent the plausibility of an opposite or alternative hypothesis or theory. Therefore, he introduced the so-called *counter-explanation* debiasing technique. Counter-explanation (CE), applied after the retraction of misinformation, consists in inducing the subjects to imagine there is evidence supporting the validity of the opposite (or an alternative) hypothesis and try to explain why this opposite (alternative) hypothesis might be true. Considering the categorization of techniques to improve the effectiveness of retractions by Lewandowsky et al. (2012), the CE technique belongs to the category of corrections telling an alternative story. Anderson (1982) showed in an experiment that CE mitigates the belief perseverance bias. The effectiveness of the CE technique in mitigating the belief perseverance bias has also been demonstrated in experiments by Lord et al. (1984) and Anderson and Sechler (1986).

The applicability of the CE debiasing technique in the context of misinformation in the general public is, however, limited; asking individuals (e.g., news consumers or potential customers reading product reviews) to write down reasons why the opposite (or an alternative) hypothesis might be true after discovering that it was misinformation does not seem feasible. Nevertheless, the CE debiasing technique inspired us to develop a new *counter-speech* (CS) debiasing technique with a higher potential for applicability in practice. In the CS debiasing technique, the individuals are asked to read instead of writing down why the initial information might not be true. The CS technique is introduced in Sec. 3.1.2.2. In Study 2, we test both the CS technique and the CE technique and compare them in terms of their effectiveness in mitigating the belief perseverance bias. Although the practical applicability of the CE technique is only limited, we include this technique in our study mainly for comparison purposes.

Inoculation

To mitigate the belief perseverance bias by making salient the plausibility of the opposite or alternative hypothesis or theory, Anderson (1982) proposed, besides the CE technique, also the so-called *inoculation*. The inoculation debiasing technique consists in creating plausible explanations for both (or all) possible hypotheses before reading a particular piece of information with the aim to reduce unwarranted hypothesis perseverance by showing how easily any of the possible hypotheses might be

explained. This should lead to the mitigation of the belief perseverance bias when the initial information is later retracted.

Although the inoculation technique is not applied after the retraction of misinformation but before even reading the initial information that may later be retracted, the technique may still be classified as a correction telling an alternative story in the categorization of techniques to improve the effectiveness of retractions by Lewandowsky et al. (2012). Indeed, the explanations for an alternative (or opposite) hypothesis created beforehand fill (at least partially) the coherence gap otherwise left by the retraction. Although Anderson (1982) showed that inoculation mitigates the belief perseverance bias, its practical applicability is limited; explicitly asking people to create plausible explanations to all possible (or alternative) hypotheses before reading the initial information is infeasible.

The inoculation debiasing technique for mitigating the belief perseverance bias proposed by Anderson (1982) should not be confused with the group of techniques of the same name reviewed in Sec.1.1, as they rely on different mechanisms. Indeed, the former belongs to corrections telling an alternative story, while the latter is categorized as an up-front warning.

Awareness training

Hammond, Keeney, and Raiffa (1998) argued that "[…] the best protection against all psychological traps […] is awareness. […] even if you can't eradicate the distortions ingrained into the way your mind works, you can build tests and disciplines into your decision-making process that can uncover errors in thinking before they become errors in judgments. And taking action to understand and avoid psychological traps can have the added benefit of increasing your confidence in the choices you make" (Hammond et al., 1998, p. 55). However, according to Gaeth and Shanteau (1984), being aware of a potential bias is not sufficient for mitigating the bias, and training explicitly designed for debiasing is necessary. The reason for this is, according to Mowen and Gaeth (1992), that "decision makers may not recognize their own fallibility until they are personally confronted with it" (Mowen & Gaeth, 1992, p. 185).

In relation to the confirmation bias, Nickerson (1998) suggests that "[…] simply being aware of the confirmation bias […] might help one both to be a little cautious about making up one's mind quickly on important issues and to be somewhat more open to opinions that differ from one's own" (Nickerson, 1998, p. 211). Thus, he argues that the impact of awareness training on reducing confirmation bias should be examined more closely. Anderson and Lindsay (1998) recommend education and training to improve society's general reasoning ability to reduce naive theory biases.

Nevertheless, the effectiveness of awareness training in reducing decision biases has not been studied in much detail yet. Aczel, Bago, Szollosi, Foldes, and Lukacs (2015) studied awareness training and analogical training in an experiment with the aim to initiate the exploration of debiasing techniques applicable in a real-life setting and achieving lasting improvement in decision making. Their experiment focused on ten biases (covariation detection, anchoring bias, overconfidence bias, outcome bias, etc.). The belief perseverance bias was, however, not included. Awareness training consisted of a general

introduction of heuristics and biases and the presentation of each bias. In the introduction, information about the duration and the aim of the training was provided, flaws in intuitive decision making were demonstrated by several examples, concluding that our intuition can often misguide us in real life, and the participants received a presentation on how people make mistakes in problems similar to those the participants encountered in the experiment. The presentation of each bias then consisted of a real-life example, an explanation of the bias, and some techniques to avoid the bias.

Also the specific warning about the continued influence effect proposed and studied by Ecker et al. (2010) falls into the category of awareness-training debiasing techniques. The specific warning, applied before reading the particular information, consists in explaining the continued influence effect and demonstrating its operation on two concrete examples. Ecker et al. (2010) showed that the specific warning reduces but does not eliminate the continued influence of misinformation.

In this paper, we apply awareness training to the belief perseverance bias. In Study 2, we compare the awareness-training (AT) debiasing technique with the CS and CE debiasing techniques in terms of their effectiveness.

Analogical training

Analogical training studied by Aczel et al. (2015) bases on analogical encoding – comparison of two situations aiming at discovering common principles and transferring them to new structurally similar situations. Analogical training applied by Aczel et al. (2015) was based on group work and lasted approximately 2 hours. Although their experiment suggested that analogical training could be more successful in mitigating biases than awareness training, its practical applicability in mitigating the belief perseverance bias in the context of misinformation in the general public is limited.

1.3 Experimental design

Experiments on mitigating the belief perseverance bias usually consist of three main steps: 1) manipulation of participants' opinion on a specific topic, 2) retraction of misinformation, and 3) application of a debiasing technique to mitigate the belief perseverance bias. The pioneering experiments on inducing and mitigating the belief perseverance bias by Anderson and colleagues used manipulation of participants' *opinion on the relationship between firefighters' attitude to risk and successfulness in their job*. Indeed, numerous studies confirmed that participants' opinion on this relationship could be manipulated and the belief perseverance bias induced in an experimental setting (see, e.g., Anderson et al., 1980; Anderson, 1982, 1983; Anderson, New, & Speer, 1985; Anderson & Sechler, 1986). We, therefore, adopted manipulation of participants' opinion on this topic for our study.

In studies on the belief perseverance bias, the *posttest-only control group design* has usually been used (see, e.g., Anderson et al., 1980; Anderson, 1982, 1983). This design has two significant drawbacks. It does not allow for determining whether there is a difference between the experimental and control groups before the study, and, more importantly, it does not allow for determining the amount of change between pretest and posttest. The latter drawback is for the studies on mitigating the belief perseverance bias particularly severe. First, it does not allow for determining the amount of change in belief

perseverance caused by applying a debiasing technique. Thus, we can only determine whether there are significant differences between the treatment and the control group. However, we cannot say anything about how effective a debiasing technique is (Was the belief perseverance of the participants considerably reduced or even eliminated? Or was the debiasing even "too strong"?). Second, it is not possible to identify participants who show belief perseverance after the retraction of misinformation. This means that the effectiveness of debiasing techniques is tested on samples containing participants without belief perseverance (this is as reasonable as, e.g., testing a headache treatment on "patients" who do not suffer from headaches). The conclusions about the effectiveness of the debiasing techniques thus could be distorted. To overcome these drawbacks, we use the *pretest-posttest control group design* in our studies. In order to be able to determine changes in participants' opinion during the experiment and thus identify participants with(out) belief perseverance and compare the effectiveness of various debiasing techniques, we measure participants' opinion several times during the experiment.

The same measurement items are commonly used for pretest and posttest in pretest-posttest designs (with or without a control group). For example, Maegherman et al. (2021) used the same sets of items three times within a study on belief perseverance, and Roozenbeek and van der Linden (2019) used the same sets of items for pretest and posttest in their study to test the effectiveness of the inoculation game *Bad News* in increasing resistance to online misinformation. However, using the same sets of measurement items repeatedly within an experiment may impact the results. For example, presenting the same set of items twice in experiments in which participants' performance is tested (such as the ability to spot fake news in the experiment by Roozenbeek and van der Linden (2019)) may cause a *practice effect* (participants improve in the posttest after having "practiced" on the same items in the pretest). Contrarily, presenting the same set of items twice in experiments in which participants' opinion or belief is measured (such as in experiments on the belief perseverance bias) may lead to no significant results as the participants are likely to try to *maintain consistency* (at least to some degree) in their answers. Indeed, Maegherman et al. (2021) failed to observe belief perseverance in their experiment, which might be because they used the same sets of items three times within the experiment. To overcome these problems, we use different sets of measurement items at each measurement time in our experiment.

Nevertheless, using different sets of items for pretest and posttest is related to another problem – the *item order effect*. This means that the order of the sets of items might influence the results of an experiment. Roozenbeek, Maertens, McClanahan, and van der Linden (2021) examined the item order effect in the experiment on the effectiveness of the *Bad News* game conducted by Roozenbeek and van der Linden (2019). They found a significant effect for one order and no effect for the other order of two sets of items. To reduce the item order effect in our study, we use *counterbalancing* - administering the sets of measurement items to different participants in different orders. More precisely, we use *random counterbalancing*, in which the order of the measurement items is randomly determined for each participant. More details on how random counterbalancing is applied in our experiment follow in Sec. 3.1.2.3.

The repeated measurement of participants' opinion in our study requires a relatively large number of items suitable for indicating participants' opinion on the topic. These items need to be first developed and validated. Further, since we intend to use a new treatment to manipulate participants' opinion in our study, the suitability of such treatment for biasing participants' opinion and inducing belief perseverance should be tested first. Therefore, we conduct a preparatory study (Study 1), in which we develop and validate two biasing treatments and numerous items for indicating participants' opinion on the topic. Afterward, we use one validated biasing treatment and a set of validated measurement items in Study 2 to study three debiasing techniques and compare them in terms of their effectiveness in mitigating the belief perseverance bias.

2 Study 1: Testing biasing treatments and measures of opinion

The aim of Study 1 was twofold: 1) to develop a biasing treatment and confirm its suitability for biasing participants' opinion and inducing belief perseverance in an experimental setting, and 2) to develop and validate measures of participants' opinion on the relationship between firefighters' attitude to risk and successfulness in their job (shorty a *risk-attitude & success relationship*).

2.1 Method

2.1.1 Participants

The participants were recruited by Qualtrics©. The data were collected anonymously. The sample N = 92 consisted of 41 females and 51 males, 51 residing in the UK, 27 in the Netherlands, and 14 in Germany. Further, 32 participants were of age between 18 and 23, 30 participants were of age between 24 and 29, and 30 participants were of age between 30 and 35. The median of the time the participants spent on the study was 24.4 minutes (IQR = 11.0).

2.1.2 Materials

2.1.2.1 Biasing

One purpose of Study 1 was to develop a biasing treatment and confirm its suitability for biasing participants' opinion and inducing belief perseverance in an experimental setting. We designed two biasing treatments, one treatment suggesting a positive risk-attitude & success relationship (i.e., suggesting that risk-taking firefighters are more successful in their job than risk-avoiding firefighters), the other treatment suggesting a negative risk-attitude & success relationship (i.e., suggesting that risk-avoiding firefighters are more successful in their job than risk-taking firefighters). Each treatment consisted in presenting 1) an invented summary of an alleged research study suggesting either a positive or negative risk-attitude & success relationship and 2) invented case studies of two firefighters allegedly participating in the study (see Appendix A).

The experiment participants were randomly assigned to one of two treatment groups (shortly TG). One TG (N = 48) received the biasing treatment suggesting a positive risk-attitude & success relationship

(shortly a *positive treatment* and a *positive TG*), while the other TG (N = 44) received the biasing treatment suggesting a negative risk-attitude & success relationship (shortly a *negative treatment* and a *negative TG*).

2.1.2.2 Measures of opinion

Another purpose of Study 1 was to develop and validate measures of participants' opinion on the risk-attitude & success relationship. For this purpose, we adopted (with slight modifications) one measure proposed by (Anderson, 1982) and developed three additional types of measures based on direct comparisons, Likert items, and phi coefficients. Each type of measure is described in more detail below.

Slider

We used, with slight modifications, the measure originally used by (Anderson, 1982). In particular, we asked the participants to indicate their opinion on the risk-attitude & success relationship on a slider scale ranging from -100 to 100 (-100 – absolutely negative relationship, 0 – no relationship, 100 – absolutely positive relationship). Note that (Anderson, 1982) used a scale ranging from -50 (highly negative relationship) to 50 (highly positive relationship). From now on, we will shortly refer to this measure as the *slider*.

Direct comparison

The most obvious way to get participants' opinion on the risk-attitude & success relationship is to ask them directly. Thus, we have created two oppositely worded incomplete direct-comparison statements about the successfulness of firefighters ("In my opinion, risk-taking firefighters tend to be ___ risk-avoiding firefighters." and "In my opinion, risk-avoiding firefighters tend to be ___ risk-taking firefighters.") that were to be completed by choosing from the list of 9 items (1 – extremely less successful than, 5 – as successful as, 9 – extremely more successful than). Each participant was randomly assigned one of these two statements.

The direct-comparison measure (either of the two formulations) is a valid measure of participants' opinion on the risk-attitude & success relationship. If participants' opinion was measured only once within an experiment, the direct-comparison measure would be sufficient. However, since we measure participants' opinion several times within an experiment and intend to use different sets of measures at each measurement time (see Sec. 1.3), we need more measures. We, therefore, use the direct-comparison measure in this study as a reference measure for validating other measures of participants' opinion.

Likert items

We created a list of oppositely worded Likert items about firefighters to be assessed on a 7-point scale (1 – completely disagree, 4 – neither agree nor disagree, 7 – completely agree) with an additional "I do not know" answer option. By the opposite wording of the Likert items, we mean here that one Likert item compares risk-taking firefighters with risk-avoiding firefighters (we will shortly call such Likert item *a positively formulated (Likert) item*), while the other Likert item compares risk-avoiding firefighters with risk-taking firefighters (shortly *a negatively formulated (Likert) item*).

Phi coefficients

Anderson, Lepper, and Ross (1980) and Anderson (1982) used the measures "new items" and "criterion validity" in their experiment. The "new items" measure consisted in computing the intensity of the risk-attitude & success relationship as a simple difference (X% − Y%) of participant's estimations of percentages of successful (denoted as X %) and unsuccessful (denoted as Y %) firefighters advising the risky option in a hypothetical item of the Risky-Conservative Choice test. Similarly, also the measure "criterion validity" consisted in computing the intensity of the risk-attitude & success relationship as the difference (X% − Y%) of participant's estimations of the percentage of risky responses of successful firefighters (X%) and the percentage of risky responses of unsuccessful firefighters (Y%) in the Risky-Conservative Choice test. However, it is not clear how this simple difference should represent the intensity of the risk-attitude & success relationship.

The intensity of the risk-attitude & success relationship could be better described using the *phi coefficient* (sometimes called the mean square contingency coefficient), which is frequently used in statistics to measure the intensity of the relationship between two binary variables. The phi coefficient reaches values between -1 and 1, with 0 representing no relationship/association between the variables, and -1 and 1 representing perfect negative and perfect positive relationship/association between the variables, respectively.

Using the values X and Y above, the phi coefficient for the intensity of the risk-attitude & success relationship is given as

$$\phi = \frac{X - Y}{\sqrt{(X + Y)(200 - X - Y)}} \, .$$

Positive values of φ represent a positive risk-attitude & success relationship, while negative values of φ represent a negative risk-attitude & success relationship. The bigger the absolute value of φ is, the stronger the intensity of the relationship is.

Process of validation of the measures

The process of developing and validating the measures of opinion consisted of four steps. In the first step, we proposed four types of measures for measuring participants' opinion. These were the three types described above (i.e., direct comparisons, Likert items, and phi coefficients) and one additional type (based on pairwise comparison matrices).

In the second step, we assessed the suitability of all four types of measures for indicating participants' opinion in collaboration with 18 experts. The experts were active participants of the 2019 Workshop of the Working Group "Decision Theory and Practice" of the German Society for Operations Research. We presented the research project, described the four types of measures (i.e., direct comparisons, Likert items, phi-coefficients, and pairwise comparison matrices), and distributed questionnaires to the experts. The questionnaires contained a brief description of each type of measure, a particular example of the measure as it would appear in the experiment, and two questions regarding the understandability and the validity of the given type of measure. Namely, the experts were asked to assess a) whether the (type

of) measure (the task behind it to be completed by the experiment participants) is for the participants easy or difficult to understand and b) whether it measures what it is supposed to measure. Afterward, we discussed the pros and cons of all four types of measures. Most experts agreed on the suitability of the measures based on direct comparisons, Likert items, and phi coefficients for measuring participants' opinion in our study. Contrarily, most experts held the opinion that the measures based on pairwise comparison matrices are too complicated for participants and not reliable. Therefore, we abandoned the measures based on pairwise comparison matrices and considered only the measures based on direct comparisons, Likert items, and phi coefficients.

In the third step, we created a list of oppositely worded Likert items and a list of phi-coefficient measures and administered them to three experts for content validation. A final set of nine pairs of Likert items and four phi-coefficient measures (see Appendix B) was chosen based on their feedback.

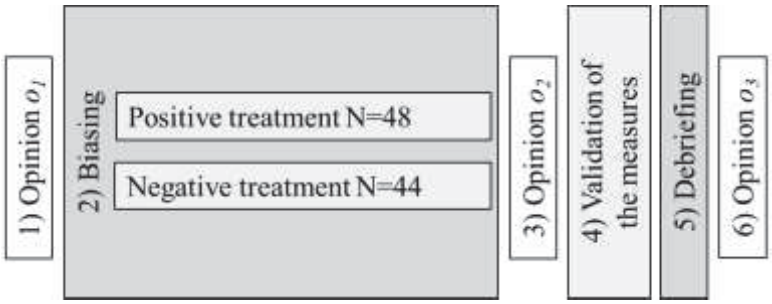
In the fourth step, the measures were empirically validated. The set of nine pairs of Likert items and four phi-coefficient measures was administered to the participants together with one randomly chosen direction-comparison measure and the slider. The order of the measures and questions within each phi-coefficient measure was randomized for each participant to reduce the question order bias. Correlation analysis was performed to assess the concurrent validity of the Likert items and phi-coefficient measures.

2.1.3 Procedure

The design of the experiment is illustrated in Figure 1, together with the sample sizes for the TGs. The experiment consists of 5 steps:

1. Measurement of initial opinion o_1 (at the measurement time t_1): At the beginning of the experiment, each participant completed one randomly chosen direct-comparison measure and the slider measure. The measures were administered to each participant in random order.
2. Manipulation - biasing treatment: The participants were randomly assigned to one of the two biasing TGs and received either a positive or negative treatment (i.e., a biasing treatment suggesting either a positive or negative risk-attitude & success relationship).
3. Measurement of opinion o_2 after biasing (at the measurement time t_2): Same as step 1.
4. Validation of measures: The participants in both TGs completed nine pairs of oppositely worded Likert items and four phi-coefficient measures (see Appendix B). The order of the measures and questions within each phi-coefficient measure was randomized for each participant to reduce the question order bias.
5. Debriefing: The participants were fully debriefed about the real purpose of the study. That is, they were told that the research report and the case studies had been invented and the alleged research study had never taken place.
6. Measurement of opinion o_3 after debriefing (at the measurement time t_3): Same as step 1.

Figure 1: Design of the experiment with sample sizes for the treatment groups.



2.2 Results and discussion

2.2.1 Validation of the biasing treatments

The mean initial opinion of the participants was that risk-taking firefighters are slightly more successful in their job than risk-avoiding firefighters (direct-comparison measure on the 9-point ordinal scale at t_1: N = 92, M_1 = 5.83, SD = 1.53). The formulation of the direct-comparison statements had no significant effect on the answer (positive formulation (N = 50): M_1 = 5.80, SD = 1.62; negative formulation (N = 42): M_1 = 5.86, SD = 1.44), $t(90) = 0.18$, $p = 0.86$, $CI_{95\%} = [-0.70, 0.58]$. The participants in the positive TG (N = 48) changed their opinion in the positive direction at t_2 (M_2 = 7.15, SD = 2.14), $t(47) = -4.00$, p = 1.1E-4, Cohen's effect size $d = 0.58$. The participants in the negative TG (N = 44) changed their opinion in the negative direction at t_2 (M_2 = 2.34, SD = 1.52), $t(43) = 12.66$, p = 2.1E-16, $d = 1.91$. Thus, we can conclude that both biasing treatments biased participants' opinion in the desired direction.

Table 1: Sample sizes, opinion means, and standard deviations for the treatment groups at the measurement times t_1, t_2, and t_3.

Treatment group	N	Opinion means			Standard deviations		
		M_1	M_2	M_3	SD_1	SD_2	SD_3
Positive TG	48	5.69	7.15	6.35	1.67	2.14	1.59
Negative TG	44	5.98	2.34	4.07	1.37	1.52	1.70
all	92	5.83	-	-	1.53	-	-

Participants' opinion moved back toward their original opinion after the debriefing at t_3 in the positive TG (M_3 = 6.35, SD = 1.59) as well as in the negative TG (M_3 = 4.07, SD = 1.70). The change in opinion was significant for both the positive TG ($t_{2,3}(47) = .10$, $p = 0.02$, $d = 0.30$) and the negative TG ($t_{2,3}(43) = -4.95$, p = 6E-6, $d = 0.75$). Nevertheless, the participants demonstrated belief perseverance. Namely, their opinion at t_3 still varied significantly from their initial opinion at t_1 in the positive TG ($t_{1,3}(47) = -2.36$, $p = 0.011$, $d = 0.34$) as well as in the negative TG ($t_{1,3}(43) = 6.79$, p = 1.3E-8, $d = 1.02$). The experiment thus confirmed the suitability of both biasing treatments for biasing participants' opinion and inducing the belief perseverance bias in an experimental setting. Sample sizes, opinion means, and standard deviations for the treatment groups at the measurement times t_1, t_2, and t_3 are shows in Table 1.

2.2.2 Validation of the measures

To assess the concurrent validity of the slider, Likert items, and phi-coefficient measures, correlations of the measures with the direct-comparison measure at the measurement time t_2 were analyzed. Since the scales for the direct-comparison measure and the Likert items are ordinal, Spearman's coefficient ρ was applied.

The correlation analysis showed strong correlations of the direct-comparison measure with the phi-coefficient measures and most Likert items. In particular, except for one pair of oppositely worded Likert items LIK_{9P} and LIK_{9N} ($0.43 < \rho < 0.49$, $p < 2E\text{-}5$) and the slider measure ($\rho = 0.54$, $p = 3.2E\text{-}8$), the correlations of all other measures with the direct-comparison measure were strong, ranging from 0.67 to 0.80 ($M = 0.72$, $SD = 0.04$, $p < 5.5E\text{-}13$). It is also worth mentioning that the correlations of the Likert items LIK_{9P} and LIK_{9N} and the slider measure with all other measures were at most moderate ($0.37 < \rho < 0.69$, $p < 4E\text{-}4$, $M = 0.57$, $SD = 0.07$) and the correlations of the slider measure with the direct-comparison measure at t_1 ($\rho = 0.31$, $p = 0.003$) and t_3 ($\rho = 0.42$, $p = 3.2E\text{-}5$) were only weak. By removing the Likert items LIK_{9P} and LIK_{9N} and the slider measure from the set of measures, the correlations among the remaining measures at t_2 were strong, ranging from 0.65 to 0.98 ($M = 0.80$, $SD = 0.06$).

The correlation analysis showed concurrent validity of eight pairs of Likert items and all four phi-coefficient measures. These were, therefore, adopted as valid measures of participants' opinion on the risk-attitude & success relationship to be used in Study 2.

3 Study 2 - Comparing the effectiveness of debiasing techniques

The aim of Study 2 was twofold: 1) to study the effectiveness of debiasing techniques to mitigate the belief perseverance bias after the retraction of misinformation, and 2) to compare the debiasing techniques in terms of their effectiveness.

3.1 Method

3.1.1 Participants

Overall, data from 366 participants have been collected. Most participants (337) were recruited by Qualtrics© in the U.K. Additionally, we conducted the experiment with 29 first-year business students at an Austrian university of applied science. The data were collected anonymously. The sample $N = 366$ consisted of 196 females and 170 males. Further, 118 participants were of age between 18 and 23, 129 participants were of age between 24 and 29, and 119 participants were of age between 30 and 35. The median of the time the participants spent on the study was 23.3 minutes ($IQR = 11.4$).

3.1.2 Materials

3.1.2.1 Biasing

For Study 2, we used the same topic as in Study 1, i.e., the risk-attitude & success relationship. To not reveal the real purpose of the experiment, it was presented to the participants as a *Study on analytical*

thinking and comprehension of scientific text. To make this more credible for the participants, we included a critical thinking scale (Sosu, 2013), a credulity scale (Kassebaum, 2004), and several tasks allegedly examining participants' comprehension of scientific text.

For biasing participants' opinion and inducing the belief perseverance bias, we used the positive biasing treatment validated in Study 1. Although two biasing treatments (positive and negative) were validated in Study 1, we decided to use only one of them to keep the total number of TGs in this study reasonably low.

Retraction of misinformation was done in the spirit of the alleged purpose of the study. That is, the participants were told that the summary of the research study had been invented with the aim to analyze people's comprehension of scientific text and analytical thinking, and the described research study had never taken place. This retraction is not to be confused with the debriefing about the real purpose of the experiment, which is done at the end of the experiment.

3.1.2.2 Debiasing techniques

The main purpose of Study 2 was to study the effectiveness of debiasing techniques in mitigating the belief perseverance bias. We considered three debiasing techniques, namely the counter-explanation inspired by the debiasing technique of the same name proposed by Anderson (1982) and the counter-speech and awareness-training techniques proposed in this paper.

Counter-explanation

The counter-explanation (CE) debiasing technique applied in our study consists in 1) repeating that the research study presented to the participants was invented, 2) pointing out that the opposite hypothesis might be true, 3) asking the participants to think of and write down at least three arguments supporting the opposite hypothesis (i.e., counter-arguments), and 4) providing an example of such a counter-argument. Thus, the CE technique employs the repetition of the retraction of misinformation and corrections telling an alternative story. Because asking people to think of and write down too many counter-arguments could cause a backfire effect (Sanna et al., 2002), only three counter-arguments are required from the participants. The exact form of the CE treatment applied in Study 2 is shown in Appendix C.

Counter-speech

The counter-speech (CS) debiasing technique builds on the CE debiasing technique. However, in contrast to the CE technique, the CS technique does not require the subjects to actively think of and write down arguments supporting the opposite (or alternative) hypothesis. Instead, it consists in providing some kind of counter-explanation to the subjects, in practice to people who have encountered misinformation. In other words, the subjects are supposed to read rather than write down arguments supporting an alternative hypothesis.

The CS technique applied in our study consists in 1) repeating that the research study presented to the participants was invented, 2) pointing out that the opposite hypothesis might be true, 3) noting that there

are several arguments supporting the opposite hypothesis, 4) providing three arguments supporting the opposite hypothesis (i.e., counter-arguments), and 5) asking the participants to spend some time thinking about the provided arguments and think of other possible arguments. Thus, the CS technique employs the repetition of the retraction of misinformation and corrections telling an alternative story. Because providing too many counter-arguments could cause a backfire effect (Sanna et al., 2002), only three counter-arguments are provided to the participants. The exact form of the CS treatment applied in Study 2 is shown in Appendix C.

Awareness training

The awareness-training (AT) debiasing technique applied in our study consists in 1) repeating that the research study presented to the participants was invented, 2) pointing out that the invented study should therefore have no influence on participants' opinion, 3) introducing belief perseverance as a phenomenon responsible for irrational behaviour, 4) illustrating the effect of belief perseverance on a hypothetical real-life situation, and 5) warning the participants about the traps of belief perseverance. Thus, the AT technique employs the repetition of the retraction of misinformation and a warning explaining the effect of the belief perseverance bias. The exact form of the AT treatment applied in Study 2 is shown in Appendix C.

Control group

A debiasing control group (CG) has been included in the experiment as a benchmark for measuring the effectiveness of the CE, CS, and AT debiasing treatments in mitigating the belief perseverance bias. The participants in the CG were administered the 21-item Proactive Decision-Making Scale (Siebert et al., 2020; Siebert & Kunz, 2016).

Hypotheses

We hypothesize that the debiasing techniques CE, CS, and AT mitigate the belief perseverance bias, while the CG treatment has no effect on the belief perseverance bias.

H_A: The counter-explanation debiasing technique (CE) mitigates the belief perseverance bias.

H_B: The counter-speech debiasing technique (CS) mitigates the belief perseverance bias.

H_C: The awareness-training debiasing technique (AT) mitigates the belief perseverance bias.

H_D: The debiasing control treatment (CG) has no effect on the belief perseverance bias.

3.1.2.3 Measures of opinion

As already discussed in Sec. 1.3, repeated measurement of participants' opinion is used in Study 2 to 1) determine the changes in participants' opinion and 2) identify the participants who show belief perseverance after the retraction of misinformation. As we intended to use different sets of measures at each measurement time in Study 2, a sufficient number of such measures was necessary. Several measures of opinion on the risk-attitude & success relationship were validated in Study 1, namely eight pairs of oppositely worded Likert items and four phi-coefficient measures.

We use one phi-coefficient measure and four Likert items to measure participants' opinion at each measurement time in the experiment. To reduce the impact of the *acquiescence bias* (the tendency to agree with statements regardless of their content; see, i.e., Lavrakas (2008)), we use the same number of positively and negatively formulated Likert items (i.e., two positively and two negatively formulated items) at each measurement time. Further, we apply random counterbalancing to reduce the item order effect. That is, the phi-coefficient measure and the Likert items are chosen randomly from the set of four phi-coefficient measures and the set of eight positively and eight negatively formulated Likert items, respectively, for each participant at each measurement time. Moreover, also the order of the phi-coefficient measure and Likert items, and the order of the questions within the phi-coefficient measure are randomized for each participant at each measurement time.

A composite score defined on the interval scale [1,7] (1 – absolutely negative risk-attitude & success relationship, 4 – no risk-attitude & success relationship, 7 – absolutely positive risk-attitude & success relationship) is computed at each measurement time as an average of the phi-coefficient measure (first transformed to the interval [1,7]) and the average value of the four Likert items. Based on the properties of the composite score, we established expertly the threshold value for opinion change as $\Delta = 0.2$ That is, when the difference of the composite scores at two measurement times is at least 0.2 then we say that there is a change in opinion. Alternatively, if the difference is less than 0.2, then there is no change in opinion.

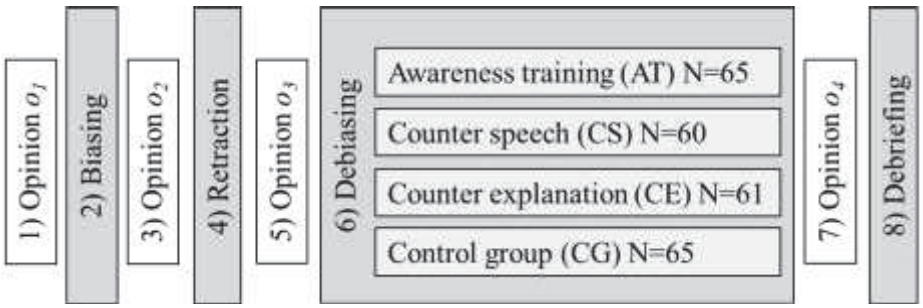
3.1.3 Procedure

The design of the experiment is illustrated in Figure 2, together with the sample sizes for the treatment groups. The experiment consists of 10 steps:

1. Measurement of initial opinion o_1 (at the measurement time t_1): Each participant completed one phi-coefficient measure and evaluated two positively and two negatively formulated Likert items randomly selected from the set of available measures and administered in random order, see Sec. 3.1.2.3.
2. Manipulation – biasing treatment: Each participant received the positive treatment (i.e., the biasing treatment suggesting a positive risk-attitude & success relationship), see Sec. 3.1.2.1.
3. Measurement of opinion o_2 after biasing (at the measurement time t_2): Same as step 1.
4. Retraction of the misinformation: Retraction of misinformation was done in the spirit of the alleged purpose of the study, i.e., the participants were told that the research summary presented to them had been invented with the aim to analyze their comprehension of scientific text and analytical thinking and the described research study had never taken place, see Sec. 3.1.2.1.
5. Measurement of opinion o_3 after retraction (at the measurement time t_3): Same as step 1.
6. Manipulation - debiasing treatment: The participants were randomly assigned to one of the three debiasing TGs (CE, CS, or AT) or a control group (CG), see Sec. 3.1.2.2.
7. Measurement of opinion o_4 after debiasing (at the measurement time t_4): Same as step 1.

8. Debriefing: The participants were debriefed about the real purpose of the experiment.

Figure 2: Design of Study 2 with sample sizes for the treatment groups.

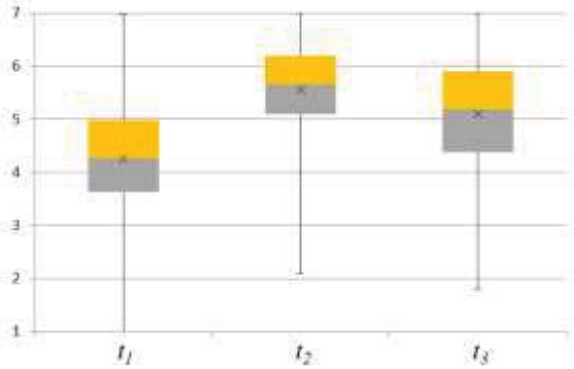


3.2 Results and discussion

3.2.1 Biasing

The mean initial opinion of the participants at t_1 was that risk-taking firefighters are slightly more successful in their job than risk-avoiding firefighters (composite score on the interval scale [1,7]: $M_1 = 4.26$, $SD = 1.04$), which agrees with the findings of Study 1 (direct-comparison measure on the 9-point ordinal scale: $M_1 = 5.83$, $SD = 1.53$). The participants changed their opinion in the positive direction after the biasing treatment at t_2 ($M_2 = 5.55$, $SD = 0.90$), $t_{1,2}(365) = -22.99$, $p = 3E\text{-}73$, Cohen's effect size $d = 1.20$. Thus, the biasing treatment had the desired effect on biasing participants' opinion, which confirms the results obtained in Study 1. Afterward, participants' opinion moved back towards their original opinion after the retraction of misinformation at t_3 ($M_3 = 5.10$, $SD = 1.08$), $t_{2,3}(365) = 8.79$, $p = 2.9E\text{-}17$, $d = 0.46$. Nonetheless, their opinion at t_3 was still significantly different from their initial opinion at t_1, $t_{1,3}(365) = -15.26$, $p = 2.7E\text{-}41$, $d = 0.80$. This result confirmed the presence of the belief perseverance bias by the participants. The boxplots of participants' opinion at times t_1, t_2, and t_3 are shown in Figure 3.

Figure 3: Boxplots of participants' opinion at the measurement times t_1, t_2, and t_3.



3.2.2 Belief perseverance

For analyzing the effectiveness of the debiasing techniques, only the participants who demonstrated belief perseverance have been considered. To identify and eliminate the participants without belief perseverance from the sample, we operationalized belief perseverance as follows. First, when the opinion of a participant moves from the initial opinion at t_1 in the direction corresponding to the biasing treatment at t_2, i.e., when $o_2 \geq o_1 + \Delta$ for the positive treatment, where Δ is a given threshold value for opinion change, we say that the participant has been manipulated by the biasing treatment. We will shortly call this opinion change a *biased opinion*. When a participant with a biased opinion (i.e., $o_2 \geq o_1 + \Delta$) persists on his or her biased opinion even after the retraction of misinformation at t_3, i.e., when $o_3 \geq o_1 + \Delta$, we say that the participant shows belief perseverance. Based on the properties of the composite score, we established the threshold value expertly as $\Delta = 0.2$ (see Sec. 3.1.2.3). To summarize, participants with belief perseverance in our study are such with $o_2 \geq o_1 + 0.2$ and $o_1 \geq o_1 + 0.2$.

Out of 366 participants, 311 participants (85%) showed biased opinion after the biasing treatment (i.e., $o_2 \geq o_1 + 0.2$), and 251 of them (68.5%) showed belief perseverance after the retraction of misinformation ($o_3 \geq o_1 + 0.2$). The sample of $N = 251$ participants with belief perseverance consisted of 138 females and 113 males. Further, 85 participants were of age between 18 and 23, 83 participants were of age between 24 and 29, and 83 participants were of age between 30 and 35.

3.2.3 Effectiveness of the debiasing techniques

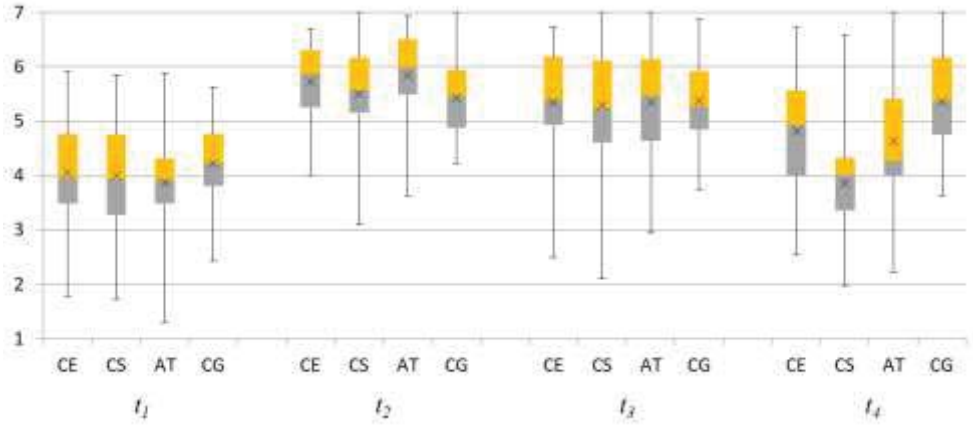
Debiasing techniques aim at returning a biased opinion persevering after the retraction of misinformation to the initial opinion before encountering misinformation (this is too ambitious and in real-world difficult to achieve) or at least reducing the belief perseverance bias (i.e., decreasing the distance of the biased opinion from the initial opinion before encountering misinformation). To assess the effectiveness of the debiasing techniques CE, CS, and AT, the difference between the initial opinion at t_1 and the opinion after the debiasing treatment at t_4, i.e., $o_1 - o_4$, is thus of relevance. The effectiveness of the debiasing techniques will be assessed by comparing the corresponding TGs with the CG. Table 2 shows the relevant statistics for the TGs at each measurement time and for the differences $o_1 - o_4$. Figure 4 shows the boxplots of participants' opinion for the debiasing TGs at each measurement time.

The CG serves as a benchmark for analyzing the effectiveness of the debiasing techniques. There was no change in opinion between the measurement times t_3 ($M_3 = 5.27$, SD = 0.97) and t_4 ($M_4 = 5.25$, $SD = 1.02$) in the CG, $t(64) = 0.41$, $p = 0.68$. Thus, hypothesis H_D was confirmed. TOST equivalence test showed that the opinion at t_3 is equivalent to the opinion at t_4 as $CI_{90\%} = [-0.07, 0.12]$ lies well within the equivalence interval [-0.2,0.2].

Table 2: Opinion means and standard deviations at each measurement time and t-test for the differences $o_1 - o_4$ for the debiasing treatment groups.

Group	N	Means				Standard deviations				$o_1 - o_4$				
		M_1	M_2	M_3	M_4	SD_1	SD_2	SD_3	SD_4	$M_{1\text{-}4}$	$CI_{95\%}$	*t*-stat	p-value	Cohens' *d*
CE	61	4.01	5.70	5.33	4.78	0.94	0.78	1.04	1.09	-0.78	[-1.01,-0.53]	-6.50	8.7E-09	0.83
CS	60	3.99	5.60	5.40	3.87	0.99	0.79	1.01	1.15	0.12	[-0.15,0.39]	0.87	0.39	0.11
AT	65	3.97	5.83	5.32	4.73	0.99	0.79	0.93	1.13	-0.76	[-1.01,-0.52]	-6.15	2.8E-08	0.76
CG	65	4.06	5.55	5.27	5.25	0.86	0.82	0.97	1.02	-1.19	[-1.40,-0.97]	-11.10	7.4E-17	1.38

Figure 4: Boxplots of participants' opinion for the debiasing treatment groups at each measurement time.



One-factor ANOVA on the differences o_1-o_4 revealed a significant effect of the debiasing techniques, $F(3,247) = 20.08$, $p = 1.1E\text{-}11$, $\eta^2 = 0.20$. Planned contrasts with Bonferroni correction further showed a significant reduction of the belief perseverance bias for the CE treatment ($t(247) = 2.41$, $p = 0.017$), the CS treatment ($t(247) = 7.57$, $p = 7.5E\text{-}13$), and the AT treatment ($t(247) = 2.50$, $p = 0.013$) compared to the CG. Thus, the hypotheses H1a, H2b, and H3c were supported, i.e., all three debiasing techniques mitigate the belief perseverance bias. The effect size was medium for the CE and AT techniques ($d = 0.43$ and $d = 0.44$, respectively) and very large for the CS technique ($d = 1.36$). The planned contrasts for the debiasing TGs on the differences o_1-o_4 with the corresponding statistics are shown in Table 3.

Table 3: Planned contrasts for the debiasing treatment groups on the differences o_1-o_4 with the corresponding statistics. Significance level with Bonferroni correction: $\alpha = 0.017$.

Contrasts	CE	CS	AT	CG	M	$CI_{95\%}$	*t*-stat	*p*-value	Cohen's *d*
Contrast 1	1			-1	0.41	[0.08, 0.75]	2.41	0.017	0.43
Contrast 2		1		-1	1.31	[0.97, 1.64]	7.57	7.5E-13	1.36
Contrast 3			1	-1	0.42	[0.09, 0.76]	2.50	0.013	0.44

Paired t-test for the CS treatment showed no significant difference between participants' opinion at t_1 ($M_1 = 3.99$, $SD = 0.99$) and t_4 ($M = 3.87$, $SD = 1.15$), $t_{1,4}(59) = 0.87$, $p = 0.39$, $M_{1\text{-}4} = 0.12$, $CI_{95\%} = [-0.15, 0.39]$, $d = 0.11$. Contrarily, there is a significant difference between participants' opinion at t_1 and t_4 for the CE treatment, $t_{1,4}(60) = -6.50$, $p = 8.7E\text{-}09$, $M_{1\text{-}4} = -0.77$, $CI_{95\%} = [-1.01, -0.53]$, $d = 0.83$, as well as for the AT treatment, $t_{1,4}(64) = -6.15$, $p = 2.8E\text{-}08$, $M_{1\text{-}4} = -0.76$, $CI_{95\%} = [-1.01, -0.52]$, $d = 0.76$. Thus, the CS technique is the most effective in mitigating the belief perseverance bias among the three techniques. Moreover, the non-significant t-test ($t_{1,4}(59 = 0.87$, $p = 0.38$) and the confidence interval $CI_{95\%} = [-0.15,$

0.39] containing 0 suggest that the CS technique could even eliminate the belief perseverance bias. Nevertheless, the equivalence of participants' opinion at t_1 and t_4 was not confirmed by the TOST analysis of equivalence as $CI_{90\%} = [-0.11, 0.35]$ does not lie within the equivalence interval [-0.2, 0.2]. The positive mean difference of participants' opinion at t_1 and t_4 ($M_{1\text{-}4} = 0.12$) also suggests that the CS technique could even work too strongly and push participants' opinion in the opposite direction, although the effect size is very low, $d = 0.11$.

Closer analysis of the differences o_1 - o_4 for the CE and AT debiasing techniques showed that there is no significant difference in their effectiveness, $t(124) = 0.05$, $p = 0.96$, M = -0.01, $CI_{95\%} = [-0.35, 0.33]$, $d = 0.01$. The TOST equivalence test did not confirm the equivalence of the CE and AT debiasing techniques in terms of their effectiveness as $CI_{90\%} = [-0.28, 0.29]$ does not lie within the equivalence interval [-0.2, 0.2]. Nevertheless, they are close to being equivalent as the $CI_{90\%}$ is relatively close to the equivalence interval [-0.2, 0.2].

4 General discussion

We conducted two studies. In the preparatory study (Study 1), we developed and validated measures of participants' opinion on a certain topic and two manipulation treatments for biasing participants' opinion on the topic and inducing belief perseverance. In the main study (Study 2), we developed debiasing techniques to mitigate the belief perseverance bias after the retraction of misinformation and compared them in terms of their effectiveness. In this section, we review the findings and suggest directions for future research.

4.1 Topic of experiments

A prerequisite for studying the effectiveness of techniques to mitigate the belief perseverance bias in our study was that we could induce the belief perseverance bias by the participants. We succeeded in manipulating participants' opinion and inducing the belief perseverance bias in our study. Namely, 85% of the participants got biased by the biasing treatment, and 68.5% showed belief perseverance. These high numbers demonstrate how easily individuals' opinion can be manipulated and emphasize the importance of techniques for mitigating the belief perseverance bias.

The topic we used to manipulate participants' opinion in our study has two features. First, we assume that the vast majority of the participants are not involved with this topic as it is supposed to be of low relevance for their lives or decisions. Second, we assume that the vast majority of the participants have no pre-formed opinion on this topic. That is, we assume that the participants have not been actively thinking about it before the experiment and "form" their opinion on the topic first at the beginning of the experiment as they are asked for their initial opinion. These two features might have made biasing participants' opinion and inducing the belief perseverance bias in an experimental setting easier than it would have been with other topics.

While belief perseverance caused by misinformation on topics of low relevance has only limited negative implications, belief perseverance caused by misinformation on topics of high relevance (such

as topics concerning health, money, safety, or politics) can have serious implications for individuals, organizations, and society. Therefore, further research on the belief perseverance bias should focus on topics of high relevance for individuals, organizations, and society.

4.2 Debiasing techniques

The focus of the paper was on developing debiasing techniques suitable for mitigating the belief perseverance bias after the retraction of misinformation and comparing them in terms of their effectiveness. We adopted (with modifications) the counter-explanation technique (CE) proposed by Anderson (1982) and developed two new debiasing techniques – counter-speech (CS) and awareness training (AT). All three debiasing techniques proved to mitigate the belief perseverance bias (the hypotheses H_A, H_B, and H_C were supported). The CE and AT debiasing techniques had a medium-sized effect on mitigating the belief perseverance bias and were close to being equivalent in terms of their effectiveness. The CS debiasing technique had a very large-sized effect on mitigating the belief perseverance and proved to be the most effective in mitigating the belief perseverance bias among the three debiasing techniques. The data suggested that the CS technique could even fully eliminate the belief perseverance bias.

However, the conclusions about the effectiveness of the debiasing techniques in mitigating the belief perseverance bias after retraction of misinformation should be generalized to other topics with care. It is likely that the effectiveness of the debiasing techniques changes with the topic. Moreover, we used a topic of low relevance for the participants in our experiment. Therefore, it is unclear how effective the techniques are with topics of high relevance. Future research should, therefore, examine the effectiveness of the debiasing techniques in mitigating the belief perseverance bias on other topics, especially on topics of high relevance to individuals, organizations, and society.

The debiasing techniques vary not only in terms of their effectiveness in mitigating the belief perseverance bias after the retraction of misinformation but also in terms of their practical applicability and the effort related to applying these techniques in praxis. A brief overview of the performance of the debiasing techniques in terms of effectiveness, practical applicability, and effort is provided in Table 4.

Table 4: Comparison of the debiasing techniques in terms of effectiveness, practical applicability, and effort.

Debiasing technique	Effectiveness	Practical applicability	Effort of the recipients of misinformation	Effort of the providers of the debiasing treatment
CE	moderate	limited	High	moderate
CS	high	high	low	high
AT	moderate	high	low	low

In the CE debiasing technique, the recipients of misinformation are asked to actively think of and write down counter-arguments. Thus, this technique requires active participation from the recipients of misinformation associated with high cognitive and time effort. Moreover, a moderate effort is required from the providers of the debiasing treatment who have to formulate the text of the CE treatment for every single piece of misinformation. Thus, the practical applicability of this technique in the context of

misinformation in the general public (such as with fake news or fake research) is very limited. The technique could, however, be applied to particular one-time decision situations of high relevance to individuals, organizations, or society in which the individuals are willing or motivated to undergo the required effort. Thus, the technique could be applied, for example, in a court setting when asking jurors to disregard a piece of information they have heard. An up-to-date example of a personal decision situation possibly influenced by misinformation to which the CE technique could be applied is deciding whether to get a COVID-19 vaccination or which COVID-19 vaccine to choose. Another example is selecting one of several available treatments to handle a life-threatening disease when finding out that a piece of information playing an essential role in the decision situation is actually misinformation.

In the AT debiasing technique, the recipients of misinformation are supposed to read a short text explaining and illustrating the effect of the belief perseverance bias. This technique, therefore, requires only passive participation of the recipients of misinformation associated with low cognitive and time effort. Additionally, also the effort of the providers of the debiasing treatment is low as the general text of the AT treatment does not need to be adapted to a particular piece of misinformation. Therefore, this technique is well applicable in practice, particularly in the context of misinformation in the general public. In our study, the AT technique was applied after the retraction of misinformation. However, the AT technique could also be used independently of particular misinformation to prevent the belief perseverance bias even before misinformation occurs. This could increase the effectiveness of the retraction of misinformation on mitigating the belief perseverance bias. Future research should therefore examine whether a general awareness training on the belief perseverance bias in the context of misinformation, applied, for example, as a part of an initiative to raise awareness and improve societal resilience to misinformation, increases the effectiveness of the retraction of misinformation on mitigating the belief perseverance bias. There is already some evidence that this might work. Indeed, Ecker et al. (2010) showed in an experiment that awareness training applied up-front reduces the continued influence effect of misinformation after retraction.

In the CS debiasing technique, the recipients of misinformation are supposed to read a short text with counter-arguments. This technique, therefore, requires only passive participation of the recipients of misinformation associated with low cognitive and time effort. Contrarily, the CS technique requires high effort from the providers of the debiasing treatment who need to formulate the counter-arguments. The text with the arguments for the CS treatment has to be designed for every single piece of misinformation or, in an ideal case, for every topic susceptible to misinformation. For example, one standardized CS text containing arguments for the COVID-19 vaccination might be used to counter any fake news containing arguments against the vaccination. Thus, the CS technique is applicable in practice, but its application is associated with the effort of the providers of the debiasing treatment.

The CS treatment in our study included arguments for an alternative (or opposite) hypothesis. However, the CS treatment in this form can be applied only when there exists an alternative hypothesis or explanation. There are, however, many situations in which an alternative hypothesis is unknown, even

when it is clear that the initial information was not correct (Lewandowsky & van der Linden, 2021). The CS treatment could also be adapted to such cases. Namely, instead of providing arguments why an alternative hypothesis is true, arguments, why the retracted hypothesis is not true, could be provided. In future research, the effectiveness of this version of the CS treatment should be tested and both versions compared in terms of their effectiveness.

As our study examined the effectiveness of single debiasing techniques on mitigating the belief perseverance bias after the retraction of misinformation, an interesting question for future research is whether the effectiveness could be increased by combining various debiasing techniques. For example, awareness training and counter-speech could be well combined. A general awareness training could be applied before the misinformation to increase the effectiveness of retraction, and the counter-speech could be then applied after the retraction of misinformation.

5 Conclusions

This paper was concerned with bias mitigation in the phases of gathering relevant information and forming preferences in the presence of misinformation. In particular, we proposed the counter-speech and awareness-training debiasing techniques for mitigating the belief perseverance bias after the retraction of misinformation and compared them in an experiment with the counter-explanation technique proposed by Anderson (1982). In the experiment, we manipulated participants' opinion on a topic of low relevance adopted from previous experiments on the belief perseverance bias (such as Anderson et al., 1980; Anderson, 1982, 1983). All three debiasing techniques proved to mitigate the belief perseverance bias. The counter-speech technique was highly effective in mitigating the belief perseverance bias, while the awareness-training and the counter-explanation techniques were moderately effective. Moreover, the counter-speech and awareness-training techniques have a high potential for practical applicability in the context of misinformation in the general public (such as with fake news), mainly because they require only low effort from the recipients of misinformation.

The study has some limitations. The retraction of misinformation and the debiasing were done shortly after the biasing treatment in our experiment, as it is common in experiments on reducing the effects of misinformation. However, in practice, it usually takes days between reading a piece of information and finding out that it was actually misinformation. It is, therefore, unclear whether or how the effectiveness of the debiasing techniques would change in practice. Moreover, the effectiveness of the debiasing techniques was studied on one particular topic of low relevance to individuals. Thus, the conclusions about their effectiveness should be generalized to other topics, particularly to topics of high relevance to individuals, organizations, and society, with care.

The paper provides several directions for future research. First, the effectiveness of the debiasing techniques in mitigating the belief perseverance bias should be examined on topics of high relevance to individuals, organizations, and society. Second, the focus should be on enhancing the applicability of the debiasing techniques in practice, particularly with fake news and fake research. Third, future

research should focus on enhancing the effectiveness of the debiasing treatments, e.g., by combining various debiasing techniques. Fourth, the impact of the belief perseverance bias and the debiasing techniques on the preferences articulated in real-world decision problems should be analyzed. Fifth, the focus should be put on integrating the debiasing techniques into standard decision processes and adapting the preference elicitation methods accordingly.

Funding: This work was supported by the Czech Ministry of Education, Youth and Sports [grant number CZ.02.2.69/0.0/0.0/18_070/0010285].

References

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Is it time for studying real-life debiasing? Evaluation of the effectiveness of an analogical intervention technique. *Frontiers in Psychology*, *6*, 1–13.

Anderson, C. A. (1982). Inoculation and counterexplanation: Debiasing techniques in the perseverance of social theories. *Social Cognition*, *1*(2), 126–139.

Anderson, C. A. (1983). Abstract and concrete data in the perseverance of social theories: When weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology*, *19*(2), 93–108.

Anderson, C. A. (1989). Causal reasoning and belief perseverance. *Proceedings of the Society for Consumer Psychology*.

Anderson, C. A. (2007). Belief Perseverance. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of social psychology*. Los Angeles, London: Sage.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: the role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*(6), 1037–1049.

Anderson, C. A., & Lindsay, J. J. (1998). The development, perseverance, and change of naive theories. *Social Cognition*, *16*(1), 8–30.

Anderson, C. A., New, B. L., & Speer, J. R. (1985). Argument Availability as a Mediator of Social Theory Perseverance. *Social Cognition*, *3*(3), 235–249.

Anderson, C. A., & Sechler, E. S. (1986). Effects of explanation and counterexplanation on the development and use of social theories. *Journal of Personality and Social Psychology*, *50*(1), 24–34.

Anglin, S. M. (2019). Do beliefs yield to evidence? Examining belief perseverance vs. Change in response to congruent empirical findings. *Journal of Experimental Social Psychology*, *82*, 176–199.

Badunenko, O., Kumbhakar, S. C., & Lozano-Vivas, A. (2021). Achieving a sustainable cost-efficient business model in banking: The case of European commercial banks. *European Journal of Operational Research*, *293*(2), 773–785.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7.

Bush, J. G., Johnson, H. M., & Seifert, C. M. (1994). The Implications of Corrections: Then Why Did You Mention It? In A. Ram, K. Eiselt, & C. S. S. (. S. [No last name!] (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society: August 13 to 16, 1994, Georgia Institute of Technology* (1st ed., pp. 112–117). London: Routledge.

Connor Desai, S. A., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453.

Del Vicario, M., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and Fake News: Early warning of potential misinformation targets. *ACM Transactions on the Web*, *13*(2), 1–22.

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570–578.

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100.

European Commission (2018a). A Europe that Protects: The EU steps up action against disinformation. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/IP_18_6647

European Commission (2018b). Final results of the Eurobarometer on fake news and online disinformation - Shaping Europe's digital future. Retrieved from https://ec.europa.eu/digital-single-market/en/news/final-results-eurobarometer-fake-news-and-online-disinformation

Farrell, J. (2019). The growth of climate change misinformation in US philanthropy: Evidence from natural language processing. *Environmental Research Letters*, *14*(3), 34013.

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, *33*(2), 263–282. Retrieved from https://www.sciencedirect.com/science/article/pii/0030507384900242

Gordon, A., Brooks, J. C. W., Quadflieg, S., Ecker, U. K. H., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia*, *106*, 216–224.

Green, M. C., & Donahue, J. K. (2011). Persistence of Belief Change in the Face of Deception: The Effect of Factual Stories Revealed to Be False. *Media Psychology*, *14*(3), 312–331.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, *363*(6425), 374–378.

Hammond, J. S., Keeney, R. L., & Raiffa, H. (1998). The hidden traps in decision making. *Harvard Business Review*, *76*(5), 47–58.

Hübner, A., Amorim, P., Fransoo, J., Honhon, D., Kuhn, H., Martinez de Albeniz, V., & Robb, D. (2021). Digitalization and omnichannel retailing: Innovative OR approaches for retail operations. *European Journal of Operational Research.* Advance online publication.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY, United States of America: Farrar Straus and Giroux.

Kai Shu, Suhang Wang, Dongwon Lee, & Huan Liu (2020). Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. In *Disinformation, Misinformation, and Fake News in Social Media* (pp. 1–19). Springer, Cham.

Kassebaum, U. B. (2004). *Interpersonelles Vertrauen: Entwicklung eines Inventars zur Erfassung spezifischer Aspekte des Konstrukts* (Doctoral dissertation), Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky. Retrieved from https://ediss.sub.uni-hamburg.de/bitstream/ediss/618/2/zusammen.pdf

Keeney, R. L. (1996). *Value-focused thinking: A path to creative decisionmaking*. Cambridge, Mass.: Harvard Universit Press.

Lahtinen, T. J., Hämäläinen, R. P., & Jenytin, C. (2020). On preference elicitation processes which mitigate the accumulation of biases in multi-criteria decision analysis. *European Journal of Operational Research*, *282*(1), 201–210.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Los Angeles, London: Sage.

Lawrence, E. K., & Estow, S. (2017). Responding to misinformation about climate change. *Applied Environmental Education & Communication*, *16*(2), 117–128.

Lewandowsky, S., Ecker, U. K., & 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.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest : A Journal of the American Psychological Society*, *13*(3), 106–131.

Lewandowsky, S., & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 1–38.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243.

Maegherman, E., Ask, K., Horselenberg, R., & van Koppen, P. J. (2021). Law and order effects: On cognitive dissonance and belief perseverance. *Psychiatry, Psychology and Law*, 1–20.

Marcus, A., & Oransky, I. (2014). What studies of retractions tell us. *Journal of Microbiology & Biology Education*, *15*(2), 151–154.

Meixler, E. (2017, December 22). Facebook Is Dropping Its Fake News Red Flag Warning After Finding It Had the Opposite Effect. *Time*. Retrieved from https://time.com/5077002/facebook-fake-news-articles/

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*(4), 379–383.

Montibeller, G., & Winterfeldt, D. von (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis*, *35*(7), 1230–1251.

Moravec, P., Minas, R., & Dennis, A. R. (2018). *Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All.*

Mowen, J. C., & Gaeth, G. J. (1992). The evaluation stage in marketing decision making. *Journal of the Academy of Marketing Science*, *20*(2), 177–187.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021). Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, *290*(1), 99–115.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780.

Perkins David (1989). Reasoning as it is and could be: An empirical perspective. In D. M. Topping, D. C. Crowell, & V. N. Kobayashi (Eds.), *Thinking across cultures: 3rd International conference on thinking : Papers / edited by Donald M. Topping, Doris C. Crowell, Victor N. Kobayashi* (pp. 175–194). Hillsdale, N.J.: L. Erlbaum Associates.

Perkins David (2019). Learning to reason: The influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues. *Judgment and Decision Making*, *14*(6), 624–643.

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906–917.

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited. *Educational and Psychological Measurement*, *81*(2), 340–362.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., . . . van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), 201199.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 497–502.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns. In M. P. Zanna (Ed.), *Advances in experimental social psychology: v.39. Advances in experimental social psychology* (pp. 127–161). Amsterdam, Netherlands, Boston, MA: Elsevier Academic Press.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In B. H. Ross (Ed.), *Psychology of Learning and Motivation: Vol. 41: Advances in Research and Theory* (pp. 265–292). San Diego: Academic Press [Imprint]; Elsevier Science & Technology Books.

Siebert, J. U., & Keeney, R. L. (2015). Creating more and better alternatives for decisions using objectives. *Operations Research*, *63*(5), 1144–1158.

Siebert, J. U., & Kunz, R. (2016). Developing and validating the multidimensional proactive decision-making scale. *European Journal of Operational Research*, *249*(3), 864–877.

Siebert, J. U., Kunz, R. E., & Rolf, P. (2020). Effects of proactive decision making on life satisfaction. *European Journal of Operational Research*, *280*(3), 1171–1187.

Sosu, E. M. (2013). The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking Skills and Creativity*, *9*, 107–119.

Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, *53*(3), 171–174.

Treen, K. M. d., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(5), e665.

Van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating Against Fake News About COVID-19. *Frontiers in Psychology*, *11*, 566790.

Watson, L. (2018). Systematic Epistemic Rights Violations in the Media: A Brexit Case Study. *Social Epistemology*, *32*(2), 88–102.

Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research*, *279*(3), 1036–1052.