Journal of Experimental Psychology: General
2014, Vol. 143, No. 1, 425–433

0096-3445/14/$12.00 DOI: 10.1037/a0032234

The Illusion of Argument Justification

Matthew Fisher and Frank C. Keil
Yale University

Argumentation is an important way to reach a new understanding. Strongly caring about an issue, which is often evident when dealing with controversial issues, has been shown to lead to biases in argumentation. We suggest that people are not well calibrated in assessing their ability to justify a position through argumentation, an effect we call the illusion of argument justification. Furthermore, we find that caring about the issue further clouds this introspection. We first show this illusion by measuring the difference between ratings before and after producing an argument for one's own position. The strength of the illusion is predicted by the strength of care for a given issue (Study 1). The tacit influences of framing and priming do not override the effects of emotional investment in a topic (Study 2). However, explicitly considering counterarguments removes the effect of care when initially assessing the ability to justify a position (Study 3). Finally, we consider our findings in light of other recent research and discuss the potential benefits of group reasoning.

Keywords: argumentation, meta-cognition, overconfidence

Disagreement is an inevitable part of our daily lives. We form opinions, take sides, and argue for our point of view. Whether it is a young child talking about which toy is best or a leading intellectual justifying a complex technical position, arguments are a key part of human interactions. To know if the arguments in support of opinions on important issues are sound, one important factor is the quality of argument that can be produced. Here, we propose an "illusion of argument justification" in which people overrate the quality of the justification that they can provide for their positions on controversial topics. We also provide evidence that emotional investment leads to a greater difference between initial appraisals of argument strength and actual persuasive force. We attempt to counteract the illusion through tacit interventions and then show that explicit consideration of alternative perspectives leads to debiasing.

We are interested in whether accurate judgments of argument quality for personal positions are readily accessible. This metacognitive ability could provide insight into how arguments are understood and engaged. If people are poor predictors of the ability to justify their views, it would point to the concern that they are ill-equipped when they enter arguments. We do not aim to cover all possible factors on an argument, but for the purposes of these studies, we focus on the introspective accuracy of the ability to justify to an audience through argument. Other research on self-evaluation and metacognition suggests that this kind of self-assessment may pose a serious challenge. People overestimate the affective impact of future events (Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000), perceive themselves as above average (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Headey & Wearing, 1988), and inaccurately predict how long it will take to complete a task (Buehler, Griffin, & Ross, 1994; Kahneman & Tversky, 1979). More generally, people have a meta-bias, known as the "bias blind spot," where they readily see the effects of cognitive and motivational biases in others but not in themselves (Pronin, Lin, & Ross, 2002). The blind spot is supported by valuing introspection (Pronin & Kugler, 2007) and by naïve realism, the tendency to view one's own subjective interpretation as the truth about reality (Griffin & Ross, 1991). Folk theories explaining phenomena are surprisingly incomplete (Wilson & Keil, 1998), and people are often unaware of the lack of depth in their understanding. In the context of reasoning about argumentation, we believe emotional investment could influence self-assessment about the ability to justify a position through argumentation.

Self-assessment for argumentation might be especially difficult because emotional investment prevents objective self-examination. Arguments, especially those dealing with controversial issues, are qualitatively different from explanations. Arguments often come with deep ties to emotions and values. Thus, we may expect emotions and values to modulate self-assessments for arguments more than those for explanations even though both may have explanatory components.

The perceived strength of an argument changes according to a variety of factors. One's level of involvement leads to stronger attitudes. For example, opinions become more extreme after participants commit to express their opinions publicly (Jellison & Mills, 1969). Attitude commitment, as measured by certainty, likelihood of change, and extremity, is associated with increased intentions to act on that attitude, attitude polarization, resistance to opposing arguments, and biased elaboration (Pomerantz, Chaiken, & Tordesillas, 1995). Similarly, attitude embeddedness, which includes measures of importance of attitude, the attitude's representativeness of values, and relevance to concept of self, is signif-

This article was published Online First March 18, 2013.

Matthew Fisher and Frank C. Keil, Department of Psychology, Yale University.

Preparation of this report and the research described herein was supported by National Institutes of Health Grant R37HD023922 to Frank C. Keil. We would like to thank Joshua Knobe, Yale's Cognition and Development Lab, and Yale's Thinking Lab for helpful comments on this article.

Correspondence concerning this article should be addressed to Matthew Fisher, Department of Psychology, Yale University, Box 208205, New Haven, CT 06520-8205. E-mail: matthew.fisher@yale.edu

icantly associated with objective elaboration but also with biased memory, especially for arguments opposing one's own attitude. These strong attitudes also tend to be stable: The greater the importance of an attitude, the less likely it is to change over time (Krosnick, 1988). Commitment to attitudes may also be strengthened through direct connection to values (Kristiansen & Zanna, 1988). Furthermore, consideration of normatively important values causes more processing, strengthened attitudes, and resistance to opposing arguments even on issues of little personal relevance (Blankenship & Wegener, 2008). Thus, strongly caring about an issue can polarize attitudes in several ways.

Emotional investment also affects how people view arguments. In one study, participants first rated the strength of pro and con arguments for seven issues and then participated in a thought-listing task by responding to a randomized subset of the arguments in short sentences (Edwards & Smith, 1996). Based on previous ratings of feelings toward the issues, the participants were placed in either a low or high emotional conviction group. Participants with high emotional conviction on an issue rated opposing arguments significantly weaker, generated more overall arguments, and generated more redundant arguments to undermine the opposing view than did participants with low emotional conviction on an issue. These biases from emotional investment could inflate people's assessment of their ability to actually articulate a coherent argument, especially if they are not fully aware of the redundancies in their own arguments and falsely see more detail and support where there is none. We therefore hypothesize that the more one cares about an issue, the stronger the illusion of argument justification. Ironically, while care and investment might be thought to lead to greater intellectual investment and understanding, the more prominent effect may merely be an illusion of having such a competence.

Overview of Present Studies

Self-testing the ability to justify through an argument requires sufficient time. One way to measure a possible illusion of argument justification is to observe how participants' assessments of their own abilities change over time. To accomplish this, we adopted a time-sensitive measure that has been used in previous research on the illusion of explanatory depth (IOED), where people overestimate their ability to produce explanatory knowledge for mechanical objects and natural processes (Rozenblit & Keil, 2002). In the IOED research procedure, participants provided initial ratings for their level of understanding for a series of devices and phenomena without pausing excessively on any item. After writing out explanations for a subset of these items, participants rated their depth of understanding as significantly lower. This effect across time was observed for explanations and not for other types of knowledge, suggesting that the illusion is not due to general overconfidence. Moreover, the IOED is even stronger in young children and therefore seems to be a foundational cognitive bias (Mills & Keil, 2004). Others have used the methodology of IOED research to investigate common controversial topics like political disagreements (Alter, Oppenheimer, & Zemla, 2010; Fernbach, Rogers, Fox, & Sloman, in press), but we applied it to people's judgments of their ability to justify through argumentation their position on these sorts of issues.

We are interested in interventions that could reduce the illusion of argument justification and the influence of emotional investment, and we explored both influences that could be considered more tacit and those that more explicitly focused on taking another coherent perspective. Tacit influences can drastically influence a wide variety of tasks such as evaluating contingencies (Tversky & Kahneman, 1981) and reporting belief in God (Shenhav, Rand, & Greene, 2012). In Study 2, we use tacit influences that both reframed the task and encouraged reflective thought. In Study 3, we investigate the impact of considering alternative perspectives on objective self-assessment. Unlike a typical explanation, arguments contain competing positions on the same issue, often creating multiple irreconcilable perspectives. Only considering one hypothesis leads to overconfidence. Explicitly instructing participants to be "objective and unbiased" does not eliminate this bias (Lord, Lepper, & Preston, 1984). However, taking into account and articulating other possible hypotheses prevents overreliance on one explanation (Brem & Rips, 2000; Gettys, Mehle, & Fisher, 1986). Thus, a potential source of greater objectivity is "considering the opposite" (Lord et al., 1984), and we therefore predict that the illusion of argument justification may only occur when people fail to actively engage alternative points of view.

Although people may be adept at generating arguments and ignoring disconfirming evidence, the degree to which people overestimate their ability to justify through arguments has not been investigated. This error could be costly. When contemplating important issues, self-reflection may lead to a belief that one has an adequate grasp of the underlying arguments, when in fact one has only a superficial understanding. In an area like politics, where arguments can eventually lead to important public policy, it seems especially important to know how well an argument is understood.

It is an open question whether we accurately assess the quality of arguments that we are about to make; we often engage in arguments and through experience we could become properly calibrated. However, based on the other metacognitive inaccuracies, we predict that there will be an illusion of justification: people will overestimate their ability to justify to others through arguments before actually having articulated them. Furthermore, we predict that deeper emotional involvement in an issue will correspond with a larger illusion of argument justification.

Study 1A

We first tested for an illusion of argument justification for common controversial topics. We also examined whether higher ratings of care for a topic corresponded with differences in predictions and self-evaluations of argument quality.

Method

Participants. One hundred eighteen adult participants (62 female, 56 male; M_{age} = 35.58, SD = 12.70) completed the survey online through Amazon's Mechanical Turk, a highly effective participant pool for research in the social sciences (Rand, 2012). All participants lived in the United States.

Procedure. Although previous work shows people are proficient at evaluating arguments (Mercier & Sperber, 2011), participants additionally received a brief training on using a 7-point scale to rate the quality of an argument based on how well it justifies the position on an issue through an understanding of the basis of the arguments. Participants read and rated three randomly ordered arguments for the use of nuclear power in the United States. The relative strength of these arguments had been determined through previous norming.

Participants received feedback if they misjudged any of the arguments and could not continue until they had correctly rated each argument. They next read and rated three randomly ordered arguments against the decision of the United States to drop atomic bombs on Japan in World War II. Only participants who were sufficiently accurate at judging the second set of arguments were included in the final analysis. Participants qualified if their three ratings fell within one standard deviation of the arguments' mean rating. The screening procedure ensured that only the participants who demonstrated an adequate understanding of how to properly use the scale were included in the analyses. This qualification criterion presumably retained only those participants who could correctly identify arguments' strength. Since we designed the experiment to detect an inability to accurately assess argument strength, excluding those who already showed this tendency only worked against our hypothesis. Implementing less strict requirements for analysis yields a similar pattern of results across all three studies.

After the training, participants considered 20 controversial topics and rated how well they could justify their position through an understanding of the basis of the arguments (Time 1 ratings). Participants used a 7-point scale anchored at 1 (*Very poorly*) and 7 (*Very well*). The order of the presentation of the issues was randomized. They received instructions to not pause excessively on any item. The instruction not to pause was designed to elicit only initial impressions and to discourage any attempt to generate arguments at that point in the procedure.

They next were asked to write out an argument of the highest quality they could for their position on four of the issues they have previously rated. Additionally, they were asked not to use any outside resources to answer the questions and informed that copying information from websites or other sources would disqualify them from the experiment. We are confident participants did not copy online material because their responses did not match arguments that could be found on top sites using an Internet search engine. Furthermore, it would take much more time to find coherent arguments of appropriate length and detail for the specific positions online than simply writing one's own take on a common controversial topic. Mechanical Turk participants are strongly motivated to progress through their tasks in the most time efficient manner possible (Rand, 2012); they were strongly incentivized not to spend the much longer time to look up answers and construct summaries from them. Each participant received a random subset of four of the following topics: human activity as a significant contributor to global warming, universal health care in America, the existence of God, and the use of cell phones leading to risk of cancer, marijuana legalization, the use of hydraulic fracturing to extract oil and gas, stem cell research, and capital punishment. After writing out each of the four arguments for their position, participants rerated how well they could justify their position through an understanding of the basis of the arguments (Time 2 ratings). The order of the topics was randomized. It is important to note that the Time 1 question and the Time 2 question used the exact same wording.

As a measure of emotional investment, participants next rated how strongly they cared about each of the 20 controversial issues on a 7-point scale anchored at 1 (*Not at all*) and 7 (*Very much*). Finally, participants reported age, gender, level of education, and college major (if applicable).

Results and Discussion

Twenty-three participants were excluded based on their response to the training items, so only responses of the remaining 95 participants were analyzed. When arguing for their own position, as predicted, participants inaccurately assessed their ability to present quality arguments. Averaging across all topics for each participant, a paired samples t test revealed a significant drop from initial ratings ($M = 4.40$ $SD = 1.20$) to ratings after writing out the argument ($M = 3.72$ $SD = 1.36$), $t(94) = 5.95$, $p < .001$. Thus, there is an illusion of argument justification when predicting the ability to justify positions on controversial topics. We next analyzed ratings of strength of care to determine if emotional investment influenced the strength of the illusion.

Reports of care consistently correlated with the participants' initial ratings. We found a significant relationship between average care and average prediction ratings (Time 1), $r(93) = .45$, $p < .001$. This indicates that participants who cared about the topics initially rated their ability to justify arguments higher than those participants who cared little about the topics. Participants completed all 20 Time 1 judgments in an average of 114.77 s and thus could not have explicitly simulated their arguments.

Strength of care also consistently correlated positively with the evaluation rating (Time 2), $r(93) = .38$ $p < .001$. As shown in Figure 1, participants were less critical of their own arguments when they were more personally invested in the issue. So even after writing out their own arguments, those who cared about the issues judged their arguments more favorably than those who cared little.

Study 1B

It could be argued that participants low in care reported lower ratings because they actually produced weaker arguments for their own position. Study 1B examines the accuracy of ratings in Study 1A by having independent raters assess the quality of arguments. If the independent raters show that the quality of arguments lowered when participants cared less, then participants had accurately reported the quality of their arguments. But if the raters judge argument quality as consistent across levels of care, we can then interpret the care and rating correlations from Study 1A as showing that those who cared strongly about the topics had the

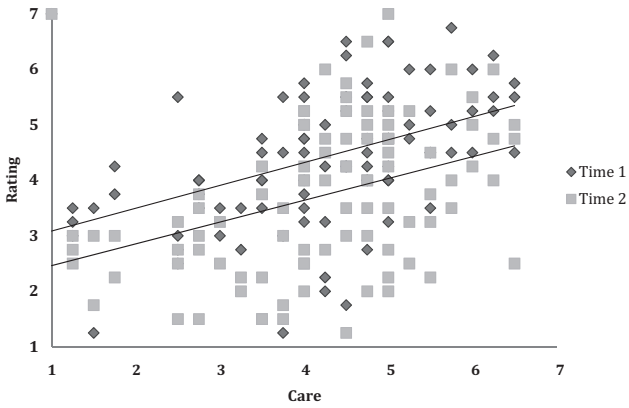

Figure 1. The bias of emotional investment in Study 1A.

largest illusions of argument justifications and were the least accurate in their self-assessments by failing to admit the weakness of their own arguments.

Method

Participants. Twenty-seven adult participants (16 female, 11 male; M_{age} = 35.81, SD = 13.42) completed the survey online through Amazon's Mechanical Turk. All participants lived in the United States. The evaluators were recruited from the same pool as the original writers, so any difference in ratings is not due to special qualifications or expert knowledge.

Procedure. Participants first received a version of the scale training used in previous studies to ensure that they had the same understanding of how to use the scale as the participants in Study 1A. Participants saw a randomized set of 36 arguments produced by an independent sample and were asked to rate the arguments according to how well it justified the position on an issue through an understanding of the basis of the arguments. This was the exact same question that participants in Study 1A responded to at both Time 1 and Time 2. Each written argument from the participants received ratings from at least seven independent judges.

Results and Discussion

An analysis of the arguments produced in Study 1 using a paired samples t test revealed that independent judges rated the arguments lower (M = 3.31, SD = 1.06) than the participants (M = 4.03, SD = 1.52), $t(35) = 3.89$, $p < .05$. Although there was a significant drop between Time 1 and Time 2 ratings in Study 1A, participants were still inaccurately rating their own arguments. Participants' ratings for their arguments were significantly higher than the ratings assigned by independent raters even after realizing some of the limitations of their initial assessments.

We next analyzed the relationship between scores of the independent raters and participants' care ratings to determine if participants who cared more about the topics produced higher quality arguments. There was no significant correlation, $r(34) = .12$, $p = .49$, suggesting that caring about the issues does not lead to better arguments. So the largest difference between self and other evaluation was for the writers who cared most about the topics. This evidence confirms that emotionally invested participants gave inaccurately high self-assessment, and they thus showed a stronger illusion of argument justification.

Study 1C

The instructions in Study 1A did not precisely specify the intended audience of participants' arguments. Thus, there was some potential for ambiguity about whether participants thought they should simply judge the validity of their arguments or whether they should assess the persuasive force of their arguments to a heterogeneous group (which could conceivably be quite different from how valid they thought their arguments were). We therefore ran an additional study with instructions that were explicitly worded so as to address this potential problem by ensuring that they were rating the arguments on the same basis used by the independent raters.

Method

Participants. One hundred nine participants (60 female, 49 male; M_{age} = 36.79, SD = 13.51) completed the survey online through Amazon's Mechanical Turk. All participants lived in the United States.

Procedure. The experiment followed the same procedure as Study 1A except for changes to the instructions. When making their Time 1 ratings, participants were asked,

> Consider your stance on the issue *X*. If you were to write about your position to a group of other Amazon Turk workers, what rating would your argument be given according to how well it justifies the position on an issue through an understanding of the basis of the arguments?

Before writing each of the 4 arguments, participants read the instructions, "Please justify your position on *X* through an understanding of the basis of the arguments." And at Time 2, participants answered the exact same question as had been posed at Time 1. Other than the changes to the instructions, the procedure was identical to Study 1A.

Results and Discussion

Twenty-four participants failed the training, so only the remaining 85 participants' responses were analyzed. As in Study1A, participants ratings dropped from Time 1 (M = 4.08, SD = 1.25) to Time 2 (M = 3.35, SD = 1.42), $t(84) = 4.80$, $p < .001$. Neither Time 1 nor Time 2 ratings were significantly different from Study 1A to Study 1C. We again found emotional investment significantly correlated with Time 1, $r(83) = .42$, $p < .001$, and Time 2 ratings, $r(83) = .25$, $p < .05$. This study provides strong evidence that removing ambiguity about the audience does not affect the results.

In summary, Study 1 demonstrated two distinct effects. First, regardless of the level of emotional investment, people are not well calibrated to the quality of arguments they are able to produce, even when their audience is fully specified. Second, emotional investment corresponded with high prediction and self-evaluation of argumentative justification, and through the ratings of independent judges, we found that those who cared the most also had the biggest illusion.

Study 2A

Study 2 attempted to blunt the illusion and the effect of care by introducing tacit influences, namely manipulations that arise from task framing or priming in ways that would not lead participants to explicitly engage other possible perspectives. Through a slight change in phrasing we framed the task to neutralize the emotional charge that had been used in Study 1A. Instead of generating arguments, participants were instructed to create lists of the pros and cons for each topic. This modification led to the generation of very similar content but eliminated the pressure of defending a personal position on a contentious issue. If the original effects were due to simply the adversarial context of the task, then they should be eliminated in this altered framing.

Method

Participants. Seventy-three adult participants (42 female, 31 male; M_{age} = 35.27, SD = 11.55) completed the survey online

through Amazon's Mechanical Turk. All participants lived in the United States.

Procedure. The Study 2 procedure was identical to Study 1, except participants rated and produced "lists of pros and cons" instead of arguments. Participants received a similar training as in Study 1, but the content of the training was restructured into lists of pros and cons. After training, participants considered 20 controversial issues and rated how thoroughly they could list the pros and cons for each topic. They received instructions to not pause excessively on any item. Participants next wrote out lists of pros and cons on one of the subsets of four issues used in Study 1. After writing out each list, they rerated how thoroughly they could list the pros and cons for those topics. Finally, participants rated how strongly they cared about each of the 20 controversial issues and reported demographic information.

Results and Discussion

Nineteen participants failed the training, so only the remaining 54 participants' responses were analyzed. As in Study 1A, there was a drop from Time 1 to Time 2. A paired samples *t* test showed that Time 1 ratings were higher ($M = 3.78$, $SD = 1.36$) than Time 2 ratings ($M = 2.86$, $SD = 1.38$), $t(53) = 4.73$, $p < .001$. Furthermore, the frame of pro and con lists did not eliminate the systematic effects of care. Care correlated strongly with both T1, $r(52) = .42$, $p = .001$, and T2 ratings, $r(52) = .49$, $p < .001$ (see Figure 2). A comparison between the items in Study 2 and the same items written as arguments in Study 1 shows that overall levels of care were no different across studies, $t(146) = .04$, $p = .97$. These correlations show that participants low in care gave low ratings for their ability to write out pros and cons and participants high in care gave high ratings for their ability to write out pros and cons.

Study 2B

We next attempted to neutralize the intuitive appeal of providing inaccurately high initial judgments at Time 1 by using a tacit prime. The modified procedure adopted a writing exercise that either promoted an intuitive or reflective mindset (Shenhav et al., 2012). If the effects were eliminated when in a reflective mindset, it would suggest that the illusion of justification and the bias of emotional investment can be overcome through subtle primes without directly comparing coherent alternative position.

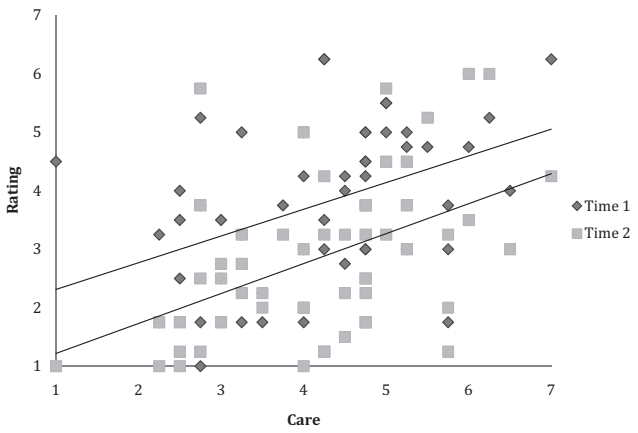

Figure 2. The bias of emotional investment in Study 2A.

Method

Participants. One hundred thirty-three adult participants (61 female, 72 male; $M_{age} = 32.45$, $SD = 10.92$) completed the survey online. All participants lived in the United States.

Procedure. The Study 2 procedure was identical to Study 1A, except participants were assigned to one of two conditions and began by answering one of four prompts. In the Intuition Good condition, participants received a prompt designed to promote intuitive thinking by instructing them to describe "a time *your intuition/first instinct* led you in the *right* direction and resulted in a *good* outcome" or "a time *carefully reasoning through a situation* led you in the *wrong* direction and resulted in a *bad* outcome." In the Intuition Bad condition, the valence of the instructions was reversed, and participants saw one of two prompts promoting reflective thinking. In both conditions, participants were required to write approximately 8–10 sentences. Following the intervention, participants completed the same procedure as reported in Study 1A.

Results and Discussion

Forty participants failed the training, so only the remaining 93 participants' responses were analyzed. In the Intuition Bad condition there was a significant drop between Time 1 ($M = 4.36$, $SD = 1.08$) and Time 2 ($M = 3.76$, $SD = 1.34$) ratings, $t(46) = 3.22$, $p = .002$. Care correlated strongly with both T1, $r(45) = .63$, $p < .001$, and T2 ratings, $r(45) = .52$, $p < .001$ (see Figure 3). These correlations suggest that participants low in care gave low ratings for their ability to justify their position and participants high in care gave high ratings for their ability to justify their position.

In the Intuition Good condition there was also a drop from Time 1 to Time 2. A paired samples *t* test showed that Time 1 ratings were higher ($M = 4.12$, $SD = 1.33$) than Time 2 ratings ($M = 3.64$, $SD = 1.30$), $t(45) = 3.15$, $p = .003$. Again, care correlated strongly with both T1, $r(44) = .48$, $p = .001$, and T2 ratings, $r(44) = .49$, $p = .001$ (see Figure 4). There were no significant differences between conditions.

These results suggest that the illusion of argument justification cannot be overcome using tacit primes. Simply entering into a more reflective mindset did not sufficiently counteract the illusion or effects of emotional investment. Next, we examined whether more explicit interventions can have a stronger impact.

Study 3A

In Study 3A, we tested the effect of considering alternative perspectives on rating the ability to justify with arguments. Unlike Study 2A and 2B, participants would now be actively considering and articulating other points of view. Previous research on the "saying is believing" effect has shown that repeating a message can influence later evaluations (Higgins & Rholes, 1978). Work on argumentation shows that when participants are randomly assigned to one side of an argument, those assigned to argue their actual position rate arguments that support their side as more acceptable than opposing arguments (Greenwald, 1969). Interestingly, those assigned to the opposing side of their position later accept an equal

number of arguments from both sides of the issue. Thus, there is some evidence that "role playing" the opposing position leads to less biased evaluation of arguments. Could assessing understanding of the opposing position instead of one's own position eliminate the effect of care on illusion of argument justification? This is the main question addressed in Study 3A.

Emotional investment was operationalized as strength of care as in Study 1. Even if emotional investment in the argument in high, there is not necessarily also personal investment. This is the case when considering the opposing position on an issue. There is a high degree of care for the issue at hand, but there is no personal investment in the particular side that is being articulated. If care prevents accurate assessment, then when considering the ability to articulate a rival view, participants should be equally incapable of accurately predicting their own performance. However, if personal investment is also necessary to produce the illusion, then participants will accurately assess how well they can justify opposing views.

Method

Participants. Ninety adult participants (52 female, 38 male; M_{age} = 34.64, SD = 11.92) completed the survey online through Amazon's Mechanical Turk. All participants lived in the United States.

Procedure. Participants received the same training as in Study 1. Participants then considered the opposing view to their stance on 20 controversial issues and rated how well they could justify the opposing position through an understanding of the basis of the arguments. They received instructions to not pause excessively on any item. Participants next wrote arguments for the opposing view on one of the subsets of four issues used in Study 1. After writing out each argument, they rerated how well they could justify the opposing position through an understanding of the basis of the arguments. Finally, participants rated how strongly they cared about each of the 20 controversial issues and reported demographic information.

Results and Discussion

Twenty-two participants failed the training, so only the remaining 68 participants' response were analyzed. When arguing the opposing view, participants again showed a significant difference between initial ratings and ratings after writing out the argument. Averaging across all items, a paired samples t test revealed a significant difference between Time 1 ratings (M = 3.78 SD = 1.20) and Time 2 ratings (M = 3.06 SD = 1.12), $t(67) = .48$, $p < .001$. These results indicate participants also could not accurately assess their ability to articulate the opposing position.

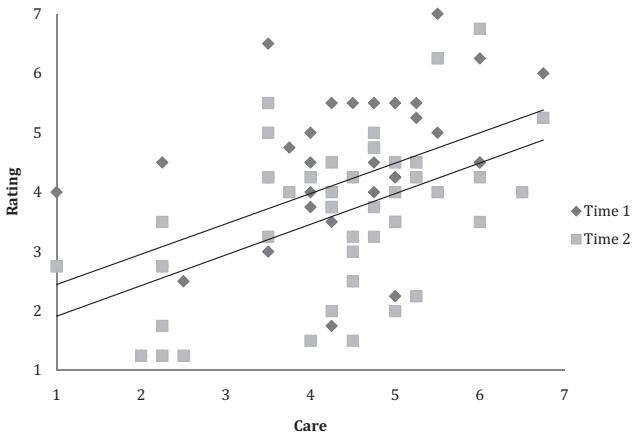

Figure 3. The bias of emotional investment in the Intuition Bad condition of Study 2B.

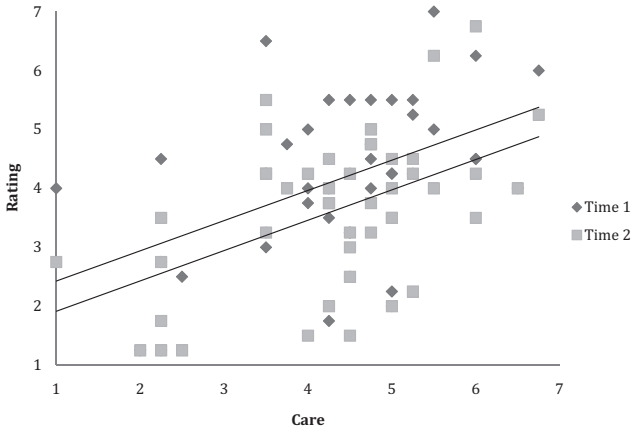

Figure 4. The bias of emotional investment in the Intuition Good condition of Study 2B.

Although there was still a significant drop from Time 1 to Time 2, unlike the previous studies, the level of care no longer influenced the initial ratings. We found that considering the opposite perspective significantly reduced the relationship between care and Time 1 ratings. There was no correlation between care and Time 1 ratings, $r(66) = .15$, $p = .22$, and the size of the correlation was significantly reduced from Study 1, $r(93) = .45$, $Z_{diff} = -2.05$, $p < .05$ (Preacher, 2002)[1]. There was also no correlation between care and Time 2 ratings, $r(66) = .22$, $p = .07$ (see Figure 5), but the size of the correlation was not significantly reduced from the care and Time 2 correlation in Study 1, $r(93) = .38$, $Z_{diff} = -1.08$, $p = .14$. This result confirms that both care and personal investment are needed to produce a significant link between care and Time 1 ratings as in Study 1A. Arguing a position without actually believing that position does not produce a systematic relationship between strength of care and positive self-assessment.

Importantly, this change in the relationship between care and ratings when articulating an opposing position is not due to a decrease in level of care for the topics. Comparing across studies, there was no significant change in total amount of care for the issues between Study 1 and Study 2, $t(161) = .39$ $p = .70$. Participants showed no less confidence when considering opposing positions as compared to considering their own positions. The average Time 1 rating did not drop from Study 1 to Study 3, $t(161) = 1.60$, $p = .11$, and the Time 2 rating also did not dropped from Study 1 to Study 3, $t(161) = 1.59$, $p = .113$.

Although previous research has found that arguments can become more convincing through "role playing," we found evidence that even in these cases where there is no personal investment in

[1] For comparisons between correlations, we report one-tailed p-values since the interventions were predicted to lessen the effect of emotional investment.

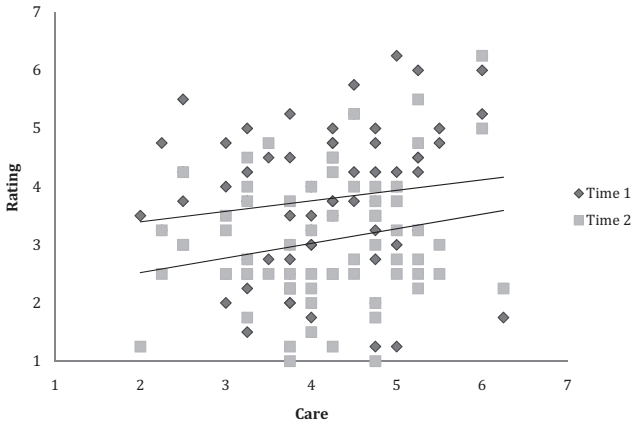

Figure 5. The removal of the emotional investment bias in Study 3A.

the argument being made, the illusion of argument justification persists. However, in line with Greenwald (1969) we found that considering the opposing position at least partially eliminates the bias of emotional investment. While this intervention was effective, it was quite heavy handed in that participants only considered the opposing viewpoints and were not queried about their own positions. We therefore next tested an intervention in which multiple perspectives on an issue were taken into consideration.

Study 3B

Entertaining a view contrary to one's own led to a reduction in the bias of emotional investment; next, we examine if an intervention causing consideration of multiple perspectives can have a similar influence. Perhaps first considering other perspectives before predicting the ability to produce arguments in support of one's own position could reduce the effects of care. Such an intervention would first introduce opposing views and then require participants to rate their ability to justify their own views. In the area of evaluating explanations, considering alternative hypotheses has been shown to mitigate overreliance on a singular explanation (Brem & Rips, 2000; Gettys et al., 1986). Furthermore, previous work on cognitive biases like the hindsight bias show that self-generating counter-arguments can help to gain a more objective perspective and have a corrective influence of social biases (Arkes, Faust, Guilmette, & Hart, 1988; Davies, 1992; Hirt & Markman, 1995; Koriat, Lichtenstein, & Fischhoff, 1980; Lord et al., 1984).

Study 3B aims to debias by introducing a new element to the original task; we ask participants to articulate an argument for the *opposing* position before predicting, writing, and assessing their *own* arguments. Importantly, this manipulation differs from the pro–con frame of Study 2A because participants never had to consider how their own justifications could be countered. Participants must now consider two fully elaborated coherent positions. We predict that this addition will reduce the effects of care as in Study 3A and will maintain high confidence for justifying one's own position.

Method

Participants. Sixty-nine adult participants (41 female, 28 male; M_{age} = 35.49, *SD* = 13.39) completed the survey online through Amazon's Mechanical Turk. All participants lived in the United States.

Procedure. The procedure was identical to Study 1A except for one addition. After the training portion, participants wrote out opposing arguments for four issues. Participants wrote arguments for one of the two subsets used in the previous studies. After Time 1 ratings of 20 items, participants then wrote out arguments for their own position on the same four issues they had previously considered. They then rerated their ability to justify their own position after each argument, rated strength of care for all 20 items, and reported demographic information.

Results and Discussion

Sixteen participants failed the training, so only the remaining 53 participants' response were analyzed. After writing an argument for the opposing view, participants again showed a significant difference between initial ratings and ratings after writing out the argument. Averaging across all items, a paired samples *t* test revealed significant difference between initial ratings (M = 4.20 SD = 1.09) and ratings after writing out an argument for their own position, (M = 3.47 SD = 1.35), $t(52) = 5.35$, $p < .001$.

We next analyzed the effect of care on ratings. In line with predictions, the intervention eliminated systematic relationships between care and ratings. There was no correlation between care and Time 1 ratings $r(51) = .16$, $p = .24$, and the size of the correlation was significantly reduced from Study 1, $r(93) = .45$, $Z_{diff} = -1.84$, $p < .05$ (Preacher, 2002). There was also no correlation between care and Time 2 ratings, $r(51) = .20$, $p = .24$ (see Figure 6), but the size of the correlation was not significantly reduced from the care and Time 2 correlation in Study 1, $r(93) = .38$, $Z_{diff} = -1.12$, $p = .13$.

The intervention managed to partially eliminate the bias of care while still maintaining the same level of confidence as the ratings without a pre-task intervention (Study 1). There was no drop in the average T1 ratings, $p = .53$, or the average T2 rating, $p = .33$. It is also important to note that the intervention did not cause participants to simply care less about the issues. An independent samples *t* test revealed no difference in total care for the participants with the intervention and those without the intervention, $p =$

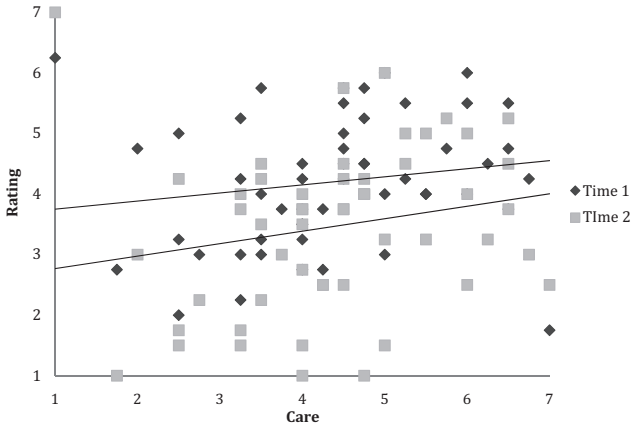

Figure 6. The removal of the emotional investment bias in Study 3B.

.70. This intervention provides evidence that a shift in perspective before considering personally held positions helps overcome biased predictions of the quality of argumentative justifications.

General Discussion

These experiments demonstrate a consistent failure to accurately assess the ability to produce justification through arguments. In particular, when exclusively articulating their own point of view, participants in our studies both over-predicted their abilities to justify their positions and rated their own arguments as better than when rated by independent judges. Furthermore, the strength of the illusion increased for those who were emotionally invested in the topic. Study 1 showed that when articulating arguments for one's own positions, caring about an issue leads to reluctance in admitting the weakness of an argument. Study 2 suggested that tacit influences are not effective correctives. Study 3 demonstrated that explicitly considering opposing points of view eliminates the effects of emotional investment on the initial assessment of the ability to offer justification.

On the surface, the illusion of argument justification may seem at odds with recent work showing there is an illusion of explanatory depth for mechanistic explanations of public policies but not when generating reasons for a position on the issue (Fernbach et al., in press). But a key difference between these studies may be in the sorts of topics considered by participants. Fernbach et al. (in press) used more technical issues such as "a cap-and-trade system for carbon emission," while the present research used more familiar topics such as abortion for which there are no straightforward mechanistic accounts. It may be the case that for less familiar, more mechanistic topics there is an illusion of causal understanding, but for more familiar, less mechanistic topics, there is an illusion of argument justification. Further research could investigate these fine-grained distinctions in more detail.

Although some of the various manipulations affected the relationship between care and ratings, it is striking that none of these manipulations were successful in removing the difference between initial and follow-up ratings. In every study participants lowered the rating of their ability after actually articulating the arguments. Given that these topics are highly controversial issues, people may believe they have more information available no matter how the task is framed. A possible criticism of the method used in our studies is that no matter what sort of ability people assess, they are always more confident in their ability before completing the task than after completing the task. While this may be true in the specific domain of articulating positions on controversial topics, it is not true of all tasks, so the results are not due simply to general overconfidence. Rozenblit and Keil (2002) specifically ruled out the general overconfidence explanation for the rate, write, re-rate task by showing that there is often no drop from Time 1 ratings to Time 2 ratings when participants consider factual knowledge, knowledge of procedures, and knowledge of narratives.

The most effective interventions were those that asked participants to consider alternative positions. Considering the opposite is a useful strategy to gain greater objectivity, but it has its limitations as well. A person is confined to expressing only as much of the opposing position as is understood. And since there is no personal investment in the opposing position on the issue, there will most often be glaring gaps and misunderstandings. One potential way around this problem is group reasoning. Reasoning in collaboration with others can serve as a corrective for individual blind spots, and it is often through argumentation that groups can reach correct answers (Moshman & Geil, 1998; Trognon, 1993).

In fact argumentation, a social device, may be the primary functional role of reasoning (Mercier, 2011; Mercier & Sperber, 2011). This claim, known as the argumentative theory of reasoning, suggests that the role of reasoning is to produce and evaluate arguments in order to persuade and be persuaded by others. For example under this view, one puzzle of reasoning, the confirmation bias (Nickerson, 1998), is viewed as a feature of reasoning, rather than a flaw, because a tendency to ignore disconfirming evidence and identify supporting arguments is useful if argumentation is the purpose of reasoning. Human reasoning might therefore not be deeply flawed and irrational but rather well suited for argumentation. Indeed, the ability to rapidly produce arguments, especially in a social context (Kuhn, Shaw, & Felton, 1997; Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993), provides support for the argumentative theory of reasoning.

Our findings fit well with this view. If reasoning is optimized through social interaction, then isolated introspection would be prone to inaccuracies such as the illusion of argument justification. But one might ask, if the purpose of reasoning is argumentation, how could people be so out of touch with their ability to produce arguments? Perhaps the illusion of argument justification removes hesitation and doubt and actually promotes argumentative engagement. If people had accurate understandings of their grasp on an argument, they would realize their gaps and inconsistencies, be less prone to argue with others, and be less able to effect changes in others' views through a forceful stance. Thus, this illusion could enable effective social interactions or at least help one achieve social goals that otherwise would not take place. When these interactions take place, ideally they bring the benefits of group reasoning: more humility and truth.

Regardless of the possible effects of group reasoning, we demonstrated that introspecting about the ability to justify through arguments has at least two built-in biases. One bias occurs when people are unaware of their actual ability to articulate arguments, and the other occurs when caring about an issue clouds accurate assessment even further. These results do not imply that people are poor arguers; just because there is poor introspective access to a cognitive mechanism does not mean it is failing to function properly. However, we have demonstrated the inaccuracy of the metacognitive judgment of argument quality and the role of emotional investment in this process, which does make it more plausible that the quality of arguments themselves would be worse that their arguers presume.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the above-average effect. *Journal of Personality and Social Psychology, 68,* 804–825. doi:10.1037/0022-3514.68.5.804

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99,* 436–451. doi:10.1037/a0020218

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology, 73,* 305–307. doi:10.1037/0021-9010.73.2.305

Blankenship, K. L., & Wegener, D. T. (2008). Opening the mind to close it: Considering a message in light of important values increases message processing and later resistance to change. *Journal of Personality and Social Psychology, 94,* 196–213. doi:10.1037/0022-3514.94.2.94.2.196

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24,* 573–604. doi:10.1207/s15516709cog2404_2

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67,* 366–381. doi:10.1037/0022-3514.67.3.366

Davies, M. F. (1992). Field dependence and hindsight bias: Cognitive restructuring and the generation of reasons. *Journal of Research in Personality, 26,* 58–74. doi:10.1016/0092-6566(92)90059-D

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71,* 5–24. doi:10.1037/0022-3514.71.1.5

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (in press). Political extremism is supported by an illusion of understanding. *Psychological Science*.

Gettys, C. F., Mehle, T., & Fisher, S. (1986). Plausibility assessment in hypothesis generation. *Organizational Behavior and Human Decision Processes, 37,* 14–33. doi:10.1016/0749-5978(86)90042-7

Greenwald, A. G. (1969). The open-mindedness of the counterattitudinal role player. *Journal of Experimental Social Psychology, 5,* 375–388. doi:10.1016/0022-1031(69)90031-6

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. *Advances in Experimental Social Psychology, 24,* 319–359. doi:10.1016/S0065-2601(08)60333-0

Headey, B., & Wearing, A. (1988). The sense of relative superiority—Central to well-being. *Social Indicators Research, 20,* 497–516.

Higgins, E. T., & Rholes, W. S. (1978). Saying is believing: Effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology, 14,* 363–378. doi:10.1016/0022-1031(78)90032-X

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69,* 1069–1086. doi:10.1037/0022-3514.69.6.1069

Jellison, J. M., & Mills, J. (1969). Effect of public commitment upon opinions. *Journal of Experimental Social Psychology, 5,* 340–346. doi:10.1016/0022-1031(69)90058-4

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science, 12,* 313–327.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118. doi:10.1037/0278-7393.6.2.107

Kristiansen, C. M., & Zanna, M. P. (1988). Justifying attitudes by appealing to values: A functional perspective. *British Journal of Social Psychology, 27,* 247–256. doi:10.1111/j.2044-8309.1988.tb00826.x

Krosnick, J. A. (1988). Attitude importance and attitude change. *Journal of Experimental Social Psychology, 24,* 240–255. doi:10.1016/0022-1031(88)90038-8

Kuhn, D., Shaw, V. F., & Felton, M. (1997). Effects of dyadic interaction on argumentative reasoning. *Cognition and Instruction, 15,* 287–315. doi:10.1207/s1532690xci1503_1

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,* 1231–1243. doi:10.1037/0022-3514.47.6.1231

Mercier, H. (2011). Reasoning serves argumentation in children. *Cognitive Development, 26,* 177–191. doi:10.1016/j.cogdev.2010.12.001

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34,* 57–74. doi:10.1017/S0140525X10000968

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology, 87,* 1–32. doi:10.1016/j.jecp.2003.09.003

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning, 4,* 231–248. doi:10.1080/135467898394148

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology, 2,* 175–220. doi:10.1037/1089-2680.2.2.175

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69,* 408–419. doi:10.1037/0022-3514.69.3.408

Preacher, K. J. (2002). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Retrieved from http://quantpsy.org

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology, 43,* 565–578. doi:10.1016/j.jesp.2006.05.011

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369–381. doi:10.1177/0146167202286008

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299,* 172–179. doi:10.1016/j.jtbi.2011.03.004

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction, 11,* 347–364. doi:10.1080/07370008.1993.9649029

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521–562. doi:10.1207/s15516709cog2605_1

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General, 141,* 423–428. doi:10.1037/a0025391

Trognon, A. (1993). How does the process of interaction work when two interlocutors try to resolve a logical problem? *Cognition and Instruction, 11,* 325–345.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458. doi:10.1126/science.7455683

Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines, 8,* 137–159. doi:10.1023/A:1008259020140

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78,* 821–836. doi:10.1037/0022-3514.78.5.821

Received August 30, 2012
Revision received February 7, 2013
Accepted February 8, 2013 ■