Recognizing Biased Reasoning: Conflict Detection during Decision-Making and Decision-Evaluation

Eva M. Janssen[a], Samuël B. Velinga[a], Wim De Neys[b], & Tamara van Gog[a]
[a] Utrecht University
[b] Université de Paris

Abstract

Although it is well established that our thinking can often be biased, the precise cognitive mechanisms underlying these biases are still debated. The present study builds on recent research showing that biased reasoners often seem aware that their reasoning is incorrect; they show signs of conflict detection. One important shortcoming in this research is that the conflict detection effect has only been studied with classic problem-solving tasks, requiring people to make a decision themselves. However, in many reasoning situations people are confronted with decisions already made by others. Therefore, the present study (N = 159) investigated whether conflict detection occurs not only during reasoning on problem-solving tasks (i.e., decision-making), but also on vignette tasks, requiring participants to evaluate decisions made by others. We analyzed participants' conflict detection sensitivity on confidence and response time measures. Results showed that conflict detection occurred during both decision-making and decision-evaluation, as indicated by a decreased confidence. The response time index appeared to be a less reliable measure of conflict detection on the novel tasks. These findings are very relevant for studying reasoning in contexts in which recognizing reasoning errors is important; for instance, in education where teachers have to give feedback on students' reasoning.

Keywords: reasoning and decision-making; decision-evaluation; heuristics and biases; conflict detection.

Introduction

Every day, people make countless decisions, and the vast majority is made effortlessly, without deliberate thought. This is highly adaptive because we would be exhausted if we had to think through each and every decision and, moreover, it usually yields good decisions. Yet, it can also lead to biases in reasoning (Kahneman, 2011; Stanovich et al., 2016). Biases are systematic errors in people's thinking and violate the normative rules of rationality as set for instance by logic or probability (Stanovich et al., 2016; Tversky & Kahneman, 1974). For example, consider the following reasoning task:

> In a study 1000 people were tested. Among the participants there were 5 dentists and 995 rock singers. Stan is a randomly chosen participant of the study. Stan is 36.
>
> He married his college sweetheart after graduating and has two kids. He doesn't drink or smoke but works long hours.

This work was funded by the Netherlands Organization for Scientific Research under project number 409-15-203.

All data and the analysis script are stored on an Open Science Framework (OSF) page for this project, see osf.io/k7uhs.

Address correspondence to Eva Janssen e.m.janssen@uu.nl.

What is most likely?

Stan is a dentist
Stan is a rock singer

Because the description of Stan fits with people's stereotype of a dentist, most people indicate that Stan is most likely a dentist (cf. 80% in a university student sample, see De Neys et al., 2011; and 60% in a North-American Mechanical Turk sample, see Frey et al., 2018). According to principles of statistical probability, however, this conclusion is not correct. The description of Stan indeed fits the image of a dentist, but could also apply to a rock singer. Importantly, since the large majority of the study's participants are rock singers it is much more likely that Stan is a rock singer than a dentist. The bias in this conclusion is referred to as "base-rate neglect", and base-rate neglect tasks such as this one are illustrative of the classic "heuristics-and-biases tasks". These tasks are widely used to demonstrate that human judgment is often based on fast intuitions or "heuristic" thinking rather than on more deliberate reasoning (Kahneman, 2011). In the example, people tend to make a probability estimation based on a representativeness heuristic telling them whether the description is more representative of a dentist or rock singer, which leads to a statistical base-rate neglect bias in their estimation.

Decades of reasoning and decision-making studies have proven that people typically perform very poorly on a wide range of heuristics-and-biases tasks (Evans & Over, 1996; Kahneman, 2011). Biases are inherent to human cognition and often relatively innocent. However, there are also many situations in which biased decisions can have serious consequences. For example, when a judge misinterprets evidence based on intuitive stereotypical associations (Eberhardt et al., 2006; Thompson & Schumann, 1987), when a doctor makes a diagnostic error due to exposure to popular media information about a disease (Schmidt et al., 2014), when investors make bad investment decisions based on the mere familiarity of a stock (Oster & Koesterich, 2013), or when parents decide not to vaccinate their children because of rare but highly publicized instances in which vaccines have failed (Smith, 2017). Therefore, it is important to understand when and why our reasoning is biased.

Although it is well established that our thinking can often be biased, the precise cognitive mechanisms underlying these biases are still debated. Until recently, influential scholars in the field suggested that most people perform poorly on heuristics-and-biases tasks because they do not recognize that their intuitive heuristic response is at conflict with logical or probabilistic principles (Evans & Stanovich, 2013; Kahneman, 2011). Put differently, it was assumed that biased reasoners are completely unaware of the error in their reasoning. Interestingly, however, recent studies have started to show that, even though they make a biased decision, most biased reasoners do show at least some sensitivity to the conflict between their heuristic response and logical considerations (De Neys & Pennycook, 2019). These studies typically compared participants' responses on reasoning tasks that – as in the "Stan" task above – prime a heuristic response which is *incongruent* with logical principles (i.e., conflict tasks) to participants' responses on tasks that prime a heuristic response which is *congruent* with the logical principles (i.e., no-conflict tasks). A no-conflict version of the "Stan" task above would refer to a study sample of 995 dentists and 5 rock singers, so that the most likely option is congruent with the prompted stereotype. In other words, conflict and no-conflict tasks trigger the exact same heuristic response, namely that Stan is a dentist, but only on the no-conflict task is this heuristic response also the correct response. Not surprisingly, almost everyone solves no-conflict tasks correctly (De Neys et al., 2011; Frey et al., 2018).

Interestingly, even though most people give the same heuristic responses to both conflict tasks and no-conflict tasks, they process the two tasks differently. People take significantly longer to enter their incorrect heuristic response on conflict tasks than they do to enter their correct heuristic response on no-conflict tasks (e.g., Bonner & Newell, 2010; De Neys & Glumicic, 2008). They are also less confident about their incorrect responses to conflict tasks, compared to their correct responses to no-conflict tasks (e.g., De Neys et al., 2011; Gangemi et al., 2015). In other words, biased reasoners show sensitivity to the logical conflict. This conflict detection effect, as indicated by confidence ratings and response times, has been found across a wide variety of classic heuristics-and-biases tasks (Bago & De Neys, 2017; De Neys, 2014; Frey et al., 2018; Mevel et al., 2015; Pennycook et al., 2015;

Stupple et al., 2013), although there are also studies that found no evidence for conflict detection (Ferreira et al., 2016; Mata et al., 2017; Pennycook et al., 2012).

Despite the increasing number of studies showing that biased reasoners often show sensitivity to their reasoning errors, research on the conflict detection effect is still in its formative stages and the effect requires further investigation (De Neys, 2012, 2014; De Neys & Pennycook, 2019). One important shortcoming is that the conflict detection effect has only been studied with classic heuristics-and-biases tasks, like the base-rate neglect task above. This is problematic because, in the end, we want to know how biased reasoning occurs in everyday situations and – while effective for demonstrating bias – these classic tasks are arguably rather artificial (Politzer et al., 2017; Prado et al., 2020). For example, judging whether a person is most likely a dentist or rock singer is quite far removed from important real-world decisions with far-reaching consequences.

Moreover, in classic heuristics-and-biases tasks, participants are always instructed to make a particular decision themselves, whereas, in everyday situations, we are also confronted quite often with biased conclusions or decisions made by others. For example, when reading news articles, people are not asked to actively reason about the likelihood of a particular situation, but are confronted with a likelihood estimation made by someone else. When that estimation confirms a reader's own intuitive ideas, *recognizing* that it is biased is arguably just as difficult as making the estimation yourself. This ability to detect biases in texts reflecting the reasoning of others is important in daily life. For example, when interpreting and analyzing interpreting arguments from activists or politicians on societal issues such as vaccines or climate change. Also, many professional contexts require people to be able to detect biases in reasoning of others. For instance, in medicine where physicians often see patients after a referral and initial diagnosis by another doctor (Van den Berge et al., 2012), in education where teachers have to detect and give feedback on biases in their students' reasoning (Janssen et al., 2019), or in justice where judges and lawyers have to interpret and weigh arguments of the prosecutors and the accused (Thompson & Schumann, 1987).

Thus, to improve our understanding of biased reasoning, it is important to establish whether people would detect biased reasoning in decisions of others, and if not, whether they show signs of conflict detection. Detecting a conflict in your own versus another person's decision might involve similar cognitive mechanisms. In this case, failing to detect bias in reasoning of others would occur as frequently as failing to avoid bias in people's own reasoning, and, moreover, a similar conflict detection effect might apply. However, it could also be the case that the underlying mechanisms differ. For instance, research into argumentation suggests that people become more deliberative and critical to biases when they have to judge the argumentation of others than when they themselves have to make a judgment (Mercier & Sperber, 2011; Trouche et al., 2016). Furthermore, Mata et al. (2013) showed that some people become better at detecting biases when they are judging others' reasoning than when they are judging reasoning without any reference to another person. If this is the case, then people would be more likely to detect biases in reasoning of others than in their own, and possibly show stronger signs of conflict sensitivity in case they do not accurately detect others' bias. On the other hand, people typically agree with conclusions confirming their own ideas and beliefs (Markovits & Nantel, 1989; Thompson & Evans, 2012). Thus, if someone else's conclusion is in line with their own intuitive ideas and the related decision does not directly affect them, people might be less motivated to pay attention to someone else's reasoning. In that case, people would be less likely to detect biases in reasoning of others than in their own or to show signs of conflict detection.

1.1. The Present Study

In sum, many previous studies have shown that people not only make biased decisions on classic heuristics-and-biases problems, but also in a wide range of other, more realistic, reasoning scenarios (e.g., Janssen et al., 2019; Mata et al., 2013; Mercier & Sperber, 2011; Schmidt et al., 2014; Thompson & Schumann, 1987; Trouche et al., 2016). It has not yet been investigated whether biased reasoners also show signs of conflict detection in reasoning scenarios other than the classic heuristics-and-biases problems. Therefore, it is both theoretically and practically

relevant to also start investigating conflict detection processes in a broader range of reasoning scenarios. The present study served as a first step in this direction by investigating reasoning accuracy and the conflict detection effect not only in decision-making but also in decision-evaluation tasks. Similar to the classic heuristics-and-bias-tasks, our problem-solving tasks required participants to make a decision about the probability of an event themselves, whereas our novel vignette tasks required participants to evaluate decisions on probability made by others that were described in short texts. The context or framing of both the problem-solving tasks and the vignette tasks differed from the classic heuristics-and-biases tasks in the sense that they described longer and more complex situations, in which the required reasoning was always relevant for achieving a particular goal. The study was explorative in nature; as mentioned earlier, it is hard to make a priori predictions on whether reasoning accuracy and conflict detection would differ or not between decision-making and decision-evaluation. Also note that our main goal was not to draw a direct comparison between conflict detection during decision-making versus decision-evaluation. Given that the conflict detection effect has already been demonstrated convincingly for decision-making on problem-solving tasks, the main goal of this study was to establish whether the conflict detection effect is also observed during decision-evaluation on vignette tasks. We used confidence ratings and response times as indices of conflict detection (e.g., De Neys, 2014; Frey et al., 2018; Pennycook et al., 2015). A lower confidence and longer response time on incorrectly performed conflict tasks relative to correctly performed no-conflict tasks would point to conflict detection.

Methods

2.1. Participants

In total, 160 native Dutch-speaking participants were recruited on Prolific Academic (www.prolific.ac) and paid £7.75 for participation. One participant had to be excluded due to a technical error, leaving a final sample of 159 participants (108 males) with an average age of 26.9 years (*SD* = 9.2). In terms of educational background, 73.0% of the participants reported having obtained a higher education degree or being enrolled to obtain this degree, 9.4% a vocational education degree, and 17.6% a secondary education degree.

2.1.1. Data statement. All data and the analysis script are stored on an Open Science Framework (OSF) page for this project, see osf.io/k7uhs.

2.2. Materials

We designed a total of 24 new reasoning tasks in Dutch, based on classic base-rate and conjunction tasks (De Neys et al., 2011; Frey et al., 2018). Section 1 in the Supplementary Materials provides an example and explanation of a classic conjunction task. Reasoning in a decision-making format was measured with six base-rate problem-solving tasks and six conjunction problem-solving tasks. From now on, we refer to these tasks as "base-rate problems" and "conjunction problems", respectively. Reasoning in a decision-evaluation format was measured with six base-rate vignette tasks and six conjunction vignette tasks. From now on, we refer to these tasks as "base-rate vignettes" and "conjunction vignettes", respectively. For the base-rate and conjunction problems, participants had to reason about probability estimation themselves (i.e., decision-making). For the base-rate and conjunction vignettes, on the other hand, the participants' job was to evaluate the probability estimation made by someone else (i.e., decision-evaluation). In addition, both the problems and vignettes differed on other aspects from the classic heuristics-and-biases tasks. Whereas classic tasks typically described short and simple situations in which the reasoning was quite far removed from real-world decisions (e.g., deciding whether Stan is a dentist or whether Jon plays in a rock band), the current tasks described longer and more complex situations in which the required reasoning was always relevant for achieving a particular goal (e.g., tackling companies committing fraud or deciding whether soups are likely to contain dangerous additives).

2.2.1. Base-rate problems. Three out of the six base-rate problems were conflict problems: the description and base-rates cued conflicting responses. The other three were no-conflict problems in which the description and base-rates cued the same response. A translated example of a base-rate problem in conflict version is:

> The Dutch government has recently made tackling fraud by companies one of the police's priorities. The police have received a list of 1000 companies that may be committing fraud. Further investigation has shown that 8 of these companies have committed fraud and that the remaining 992 companies have not committed fraud. However, certain information was lost during a reorganization. The police no longer know which companies have committed fraud. Van Been Ltd is a randomly chosen company that is on the police's list.
>
> Van Been Ltd has a closed and competitive corporate culture. Its employees put a lot of effort into making big profits. The annual report also shows that the company has made a remarkably high profit in the past year. There is also a strikingly high number of fines that employees have received in company cars.
>
> What is most likely?
>
> Van Been Ltd committed fraud
> Van Been Ltd did not commit fraud

As in the classic base-rate tasks, the narrative description was designed to cue an intuitive response based on a stereotype that is at odds with the base-rate information. All base-rate problems had the same underlying structure of about the same word length, but a different cover story. Each problem started with a sentence that introduced a particular situation, followed by two sentences including base-rate information, a sentence with additional information explaining the current situation, and a sentence introducing a randomly selected individual case. In the next paragraph, specific information about the selected individual case was presented, after which the participant had to indicate which of two possible situations was most likely[1]. To construct a no-conflict version, we simply changed the sentence including the base-rate information, so that the intuitively cued response was in line with the statistically most likely option (e.g., "Further investigation has shown that 992 of these companies have committed fraud and that the remaining 8 companies have not committed fraud).

2.2.2. Base-rate vignettes. Three out of the six base-rate vignettes were conflict vignettes, meaning that the heuristic decision by the other was at conflict with the base-rate mentioned in the task. The other three were no-conflict vignettes, meaning that the heuristic decision by the other was in line with base-rate mentioned in the task. Here is an example of the earlier base-rate conflict problem in vignette format:

> The Dutch government has recently made tackling fraud by companies one of the police's priorities. The police have received a list of 1000 companies that may be committing fraud. Further investigation has shown that 8 of these companies have committed fraud and that the remaining 992 companies have not committed fraud. However, certain information was lost during a reorganization. The police no longer know which companies have committed fraud. One of the companies on the list, Van Been Ltd, stands out for the police because of a strikingly high number of fines that employees have received in company cars. Van Been Ltd has a closed and competitive corporate culture. Its employees put a lot of effort into making big profits. The annual report also shows that the company has made a remarkably high profit in the past year. The police have decided to start an official investigation into the company, because they estimate it more likely that Van Been bv has committed fraud than that Van Been bv has not committed fraud.

[1] Responses that were in line with the base-rates (i.e., selection of the largest group as most likely answer) were labeled as correct answers. In line with Frey et al. (2018), we used extreme base-rates (variations around 995 and 5) and moderate cues to minimize the concern developed by Gigerenzer et al. (1988) that when relying on a formal Bayesian approach, selection of the heuristic response should be considered normatively correct (see De Neys, 2014).

Is the estimation on which the police have based its decision correct?

Yes
No

As the example indicates, the base-rate vignettes were very similar to the base-rate problems, but differed on three aspects. First, instead of just presenting information about a randomly chosen individual case, the story explained that one individual case had caught the attention of one of the actors in the story. Second, a sentence was added in which the actor estimated the likelihood of two possible situations, on which a specific decision was based. Third, instead of indicating which of two possible situations was most likely, participants had to indicate whether the estimation on which the actor's decision was based, was correct. No-conflict versions were again constructed by switching the base-rate information.

2.2.3. Conjunction problems. Of the six conjunction problems, three were again conflict problems and three were no-conflict problems. An example of a conjunction problem in conflict version is:

> In the past year, the Dutch Food and Consumer Product Safety Authority has investigated 10 brands of tomato soup to determine whether these contained dangerous additives or not. Immediately after the investigation, Heinz removed all its tomato soups from the store shelves, according to the company itself in order to improve the taste of the soup.
>
> What is most likely?
>
> Heinz wanted to improve the taste of the soup.
> Heinz wanted to improve the taste of the soup and the soup contained dangerous additives.

The conflict above emerges because the cued stereotype, the soup contained dangerous additives, is in the conjunctive answer option. Yet logically, the conjunction of any two probabilities can never be more likely than either of the conjuncts in isolation, formally: $p(A\&B) \leq p(A), p(B)$. In other words, the probability of Heinz wanting to improve the taste *plus* the soup containing dangerous additives can never be greater than merely the probability of Heinz wanting to improve the taste of the soup. Each problem had about the same word length and was structured as follows: It started with a sentence that introduced a particular situation. Next, an action by a person or institution was described. The person or institution always provided an unlikely explanation for this action, after which participants had to indicate which of two possible situations in the answering option was most likely[2]. Following Frey et al. (2018), to construct a no-conflict version we changed the person's or institution's provided unlikely explanation into a likely explanation. For example: "Immediately after the investigation, Heinz removed all its tomato soups from the store shelves, according to the company itself because the soup contained dangerous additives". Next, we replaced the unlikely explanation in the non-conjunctive answering option with the likely explanation. For example:

> What is most likely?
>
> The soup contained dangerous additives.
> The soup contained dangerous additives and Heinz wanted to improve the taste of the soup.

[2] People have the tendency to choose the answer that contains the stereotypical description, irrespective of whether this is the conjunctive or non-conjunctive answer option (Tversky & Kahneman, 1983). Although, it is possible that the conjunction of two probabilities is equally large as one of the two in isolation, it can never exceed the probability of either one in isolation. Therefore, the conjunctive answering option can never be more likely than the non-conjunctive one. Hence, in this reasoning situation one should normatively always choose the non-conjunctive statement.

2.2.4. Conjunction vignettes. Three out of the six conjunction vignettes were conflict vignettes and three were non-conflict vignettes. The vignette format of the conflict problem above is:

> In the past year, the Dutch Food and Consumer Product Safety Authority has investigated 10 brands of tomato soup to determine whether these contained dangerous additives or not. Immediately after the investigation, Heinz removed all its tomato soups from the store shelves, according to the company itself in order to improve the taste of the soup. However, according to an investigative journalist of the Volkskrant [Dutch news paper], it is more likely that Heinz not only wanted to improve the taste of the soup but that the soup also contained dangerous additives.
>
> Is the estimation of the investigative journalist of the Volkskrant correct?
>
> Yes
> No

As the example indicates, the conjunction vignettes differed from the conjunction problems on two aspects. First, a sentence was added in which a new actor was introduced (e.g., an investigative journalist), who made a decision about the likelihood of two possible situations. Second, instead of indicating which of two possible situations was most likely (cf. problem-solving tasks), participants had to evaluate whether the decision of the actor was correct. The no-conflict versions were created by changing the unlikely explanation provided by a person or institution into a likely explanation, and by changing the decision of the new actor into a probability estimation of a conjunctive situation in which an unlikely explanation was added to the likely explanation. For example, "However, according to an investigative journalist of the Volkskrant, it is more likely that the soup not only contained dangerous additives but that Heinz also wanted to improve the taste of the soup". In each vignette, the actor judged the conjunctive situation as more likely than the non-conjunctive situation. Hence, the actor was always incorrect.

2.2.5. Filler tasks. In addition to the 12 problem-solving tasks and 12 vignette tasks, four filler tasks were presented about halfway through to make the tasks of interest less repetitive and predictable. These were problem-solving tasks in which participants had to find the correct day of the week (cf. Schmeck et al., 2015; Van Gog et al., 2012). For example:

> Suppose today is Friday.
> What day is it the day after the day before yesterday?

2.2.6.Task sequence. Participants completed a total of 28 tasks grouped in five blocks. The first two blocks were always vignette tasks: a block of six base-rate vignettes (three conflict, three no-conflict) and a block of six conjunction vignettes (three conflict, three no-conflict). The order of these two blocks was randomized and the order of the six vignettes within each block was also randomized. Hereafter, participants completed a block with the four filler tasks. The final two blocks were always problem-solving tasks: a block of six base-rate problems (three conflict, three no-conflict) and a block of six conjunction problems (three conflict, three no-conflict). Again, the order of the two blocks and of the six problems within each block was randomized. The vignette tasks were administered first because our main goal was to establish whether conflict detection would occur during decision-evaluation. Therefore, we wanted to ensure that participants' reasoning evaluation processes were not influenced by prior exposure to problem-solving tasks. We counterbalanced the content of the reasoning tasks across task format and conflict version[3].

[3] Note each task had four versions: a conflict problem-solving version, a no-conflict problem-solving version, a conflict vignette version, and a no-conflict vignette version. Participants completed 24 tasks, hence, there were 24 × 4 = 96 task versions in total.

2.2.7. Response time. On each task, participants' response time was logged from the moment the task was presented on the screen until the participant clicked on one of the two multiple-choice answering options.

2.2.8. Confidence. Immediately after submitting their task responses, participants had to indicate how confident they were that their answer to the reasoning task was correct. Their confidence was measured in percentages from 0% (not at all confident) to 100% (completely confident) that increased in steps of 5%.

2.2.9. Confidence response time. Note that when initially designing our study, in line with Johnson et al. (2016), we also aimed to measure participants' confidence response times. For each confidence rating, we logged the time it took participants to rate their confidence (i.e., the interval between the presentation of the scale and the moment they clicked a percentage point). However, our results on this conflict-detection index appeared unreliable. Since two recent studies also found this index to be unreliable and cautioned against its use (Frey et al., 2018; Šrol & De Neys, 2019) we decided to refrain from basing any conclusions on it. For completeness and parsimony, the analyses of this index are presented in the Supplementary Materials in Section 3.

2.3. Procedure

The experiment was run online. All materials were presented in Gorilla software (Anwyl-Irvine et al., 2019). Participants were instructed that the study would take up to 45 minutes and demanded their full attention. After giving informed consent, participants were presented with general instructions on how the experiment should be displayed (full screen and notifications off). Next, an attention check was conducted to see whether the participants had read the full instruction[4], followed by some demographic questions (age, gender, and educational background). Hereafter, a short reading test was implemented to check for anomalies in reading speed or reading comprehension (adopted from Taalblad.be, Van Kelecom, 2017). None of the participants was excluded based on the reading test. To familiarize participants with the confidence measure, they were given three weekday problems (cf. filler tasks) as practice tasks. By varying the complexity on these tasks, we also got an indication of whether participants varied their confidence ratings accordingly, which was the case. Then, participants could start with the actual reasoning tasks. After finishing all blocks, one final attention check was administered to determine whether participants still answered the confidence measure attentively. Participants were presented a clearly false statement ("München is the capital of Germany") and had to indicate whether this statement was correct or incorrect and give their confidence in their answer. Ninety-six percent answered correctly with an average confidence of 97.6%, *SD* = 7.6. Four percent answered incorrectly with an average confidence of 45.0%, *SD* = 49.7.

2.4. Data Analysis

All analyses were performed using R version 4.0.0. and run separately for the base-rate and conjunction tasks. As outlined below, we fitted several mixed effects models to the trial-level data. Mixed effect models can specify fixed and random effects. Fixed effects concern the variables of theoretical interest. Random effects define the assumptions that one makes about how sampling units vary (participants and test items), and the structure of dependency that this variation creates in one's data (Barr et al., 2013). In contrast to ANOVA, mixed effects models allow for defining multiple sources of clustering in the data. This advantage allowed us to account not only for participant variation but also for item variability in each model testing our research questions.

2.4.1. Item-level check. Because the tasks were new, we checked whether the content of the items' cover stories influenced participants' accuracy. We conducted mixed-effects logistic regression models on the base-rate and conjunction tasks with response accuracy (incorrect = 0; correct = 1) as dependent variable, with item-content

[4] The final sentence of the general instruction was: "On the next page you will be asked which button you have to press. Then press space bar." On the next page, a next-button appeared along with the question "Which button do you have to press?". Participants who incorrectly clicked the next-button instead of pressing the spacebar were prompted to the read the general instructions again.

number as fixed effect. Participant number, task format (problem-solving = 0; vignette = 1), and conflict version (conflict = 0; no conflict = 1) were specified as random effects (random intercepts). Item content did not tend to affect accuracy on the tasks (see Supplementary Materials, Section 2).

2.4.2. Accuracy. To get an overview of the overall performance, we calculated participants' proportion of correct responses per task format and per conflict version. To test whether accuracy differed between task formats and conflict version, we conducted mixed-effects logistic regression models with response accuracy as dependent variable (incorrect = 0; correct = 1). Task format (problem-solving = 0; vignette = 1), conflict version (conflict = 0; no conflict = 1), and the interaction between these two were specified as fixed effects. Participant number and item-content number were specified as random effects (random intercepts).

2.4.3. Conflict detection. To provide an overview of the conflict-detection indices, we calculated participants' average confidence (%) and response time (s) across their correctly and incorrectly performed trials, per task format and per conflict version. For both task formats, we tested for conflict detection effects using the conflict-detection indices. For these analyses, we followed the standard practice to only include participants who gave at least one biased (i.e., incorrect) response on conflict tasks (e.g., De Neys et al., 2011; Frey et al., 2018). We did not analyze the correctly performed conflict trials, as conflict detection measures on correctly performed conflict trials do not provide a pure indication of conflict detection efficiency per se (De Neys & Bonnefon, 2013). The few incorrectly answered no-conflict trials were also discarded from further analyses (i.e., it is hard to interpret these trials, since no-conflict trials cue heuristic responses which are congruent with correct performance).

Per task format, we conducted linear mixed-effect models on each conflict-detection index. Conflict version (conflict = 0; no conflict = 1) was entered as fixed effect and participant number and item-content number were entered as random effects (random intercepts)[5]. In all analyses using response times, we used log-transformed values. For ease of interpretation we report the raw response time values in the tables and the text. Finally, to see how large the conflict detection effects were, we calculated the difference between participants' confidence ratings or response times on incorrect responses to conflict tasks and on correct responses to no-conflict tasks. The reported group-level conflict detection effect sizes were calculated following a standard procedure (e.g., De Neys et al., 2011; Frey et al., 2018): we subtracted the average confidence/response times on biased participants' correctly solved no-conflict trials from the average confidence/response times on biased participants' incorrectly solved conflict trials.

3. Results

3.1. Reasoning Accuracy

Table 1 presents an overview of participants' average reasoning accuracy on the base-rate and conjunction tasks. The table shows that, as expected, most participants performed poorly on the conflict tasks, whereas they performed well on the no-conflict tasks. This pattern applied to both bias tasks (conjunction and base-rate) and to both task formats (problems and vignettes). Correct solution rates were comparable to those obtained in previous studies (e.g., Frey et al. 2018).

[5] For reasons of sample size, we did not first test for effects of task format in the main analyses (i.e., then only participants who were biased on all conflict-detection indices and on both the problem-solving tasks and the vignette tasks could be included). However, we additionally ran these analyses on the smaller sample and report significant effects in the results section.

Table 1: Average Accuracy Proportion (SD) on the Base-rate and Conjunction Tasks

	Base-rate tasks	Conjunction tasks
Problems		
Conflict	0.39 (0.35)	0.26 (0.32)
No-conflict	0.90 (0.17)	0.89 (0.19)
Vignettes		
Conflict	0.35 (0.32)	0.34 (0.29)
No-conflict	0.83 (0.24)	0.75 (0.27)

The mixed-effects logistic regression models yielded a significant interaction effect between task format and conflict version on both bias tasks, base-rate: B = -0.50, SE= 0.24, W = -2.05, p = .040; conjunction: B = -1.49, SE= 0.24, W = -6.18, p < .001. The follow-up analyses reported in Table 2 show the effects of task format (decision-making versus decision-evaluation) on participants' reasoning accuracy. Task format effects differed per conflict version and per bias task. For base-rate tasks, task format did not affect performance on conflict tasks. For conjunction tasks, on the other hand, results showed that participants performed conflict tasks significantly better in vignette format than in problem-solving format. Interestingly, for both bias tasks, no-conflict versions were performed significantly better in problem-solving format than in vignette format.

Thus, with these novel reasoning tasks that described more complex and longer reasoning scenarios and included not only decision-making but also decision-evaluation, we found a similar performance pattern on conflict and no-conflict tasks as previously obtained on classic heuristic-and-biases tasks. With regard to the two reasoning formats, no-conflict tasks were performed better in problem-solving format than in vignette format. Conflict tasks, on the other hand, were either performed better in vignette than in problem-solving format (conjunction tasks) or performance did not significantly differ across task formats (base-rate tasks).

Table 2: Mixed-Effects Logistic Regression Models on Reasoning Accuracy for the Base-rate and Conjunction Tasks

	Base-rate tasks B	Conjunction tasks B
Conflict tasks		
Fixed effects		
Intercept (SE)	-0.60 (0.21)**	-1.23 (0.19)***
Task format (SE)	-0.20 (0.15)	0.44 (0.15)**
Random effects		
Item content variance (SD)	0.30 (0.55)	0.21 (0.46)
Participant variance (SD)	1.12 (1.06)	0.64 (0.80)
No-conflict tasks		
Fixed effects		
Intercept (SE)	2.46 (0.21)***	2.55 (0.28)***
Task format (SE)	-0.70 (0.20)***	-1.18 (0.20)***
Random effects		
Item content variance (SD)	0.06 (0.25)	0.40 (0.63)
Participant variance (SD)	0.56 (0.75)	0.77 (0.88)

Note. Task format: 0 = problems, 1 = vignettes. * p < .05, ** p < .01, *** p <.001.

3.2. Conflict Detection

Table 3 provides an overview of the average scores on the conflict-detection indices for correctly and incorrectly performed trials. The table shows that 135 out of the 159 participants gave at least one biased (incorrect) response to one of the conflict tasks. Furthermore, all 159 participants gave at least one correct response to one of the no-conflict tasks. To investigate whether the biased participants showed signs of conflict detection, we contrasted their average confidence and response time on incorrectly performed conflict trials with that on correctly performed no-conflict trials. As the total number of biased participants differed per task format and per bias task (see Table 3), the sample sizes differed per analysis.

Table 3: Group-Level Averages (SD) on Each of the Three Conflict-detection indices as a Function of Response Accuracy

Conflict-detection index	Conflict: correct	Conflict: incorrect	No-conflict: correct	No-conflict: incorrect
Base-rate problems				
Participants by group	*n* = 104	*n* =135	*n* = 159	*n* = 43
Average confidence (%)	66.7 (17.8)	66.8 (18.1)	77.1 (15.4)	63.5 (18.8)
Average response time (s)	39.9 (21.9)	38.3 (25.0)	36.5 (15.2)	65.6 (14.8)
Base-rate vignettes				
Participants by group	*n* = 102	*n* = 147	*n* = 156	*n* = 64
Average confidence (%)	64.6 (17.5)	68.9 (17.6)	77.7 (13.8)	62.9 (19.1)
Average response time (s)	49.8 (24.2)	47.2 (23.3)	47.7 (24.2)	51.8 (26.2)
Conjunction problems				
Participants by group	*n* = 78	*n* = 145	*n* = 157	*n* = 45
Average confidence (%)	69.1 (16.8)	70.7 (16.7)	81.7 (14.3)	67.7 (17.9)
Average response time (s)	27.4 (25.0)	24.0 (20.8)	22.0 (16.5)	25.9 (16.6)
Conjunction vignettes				
Participants by group	*n* = 112	*n* = 149	*n* = 155	*n* = 88
Average confidence (%)	58.6 (19.4)	65.4 (17.0)	71.0 (16.8)	65.5 (16.8)
Average response time (s)	30.3 (16.6)	29.1 (15.3)	28.3 (17.6)	31.5 (19.0)

3.2.1. Confidence (%). For the confidence conflict-detection index, we found that task format did not affect conflict detection effects on base-rate tasks, but it did on conjunction tasks. For both the base-rate problems and the base-rate vignettes, results showed that participants were significantly less confident about their performance on incorrectly performed conflict tasks than about their performance on correctly performed no-conflict tasks, problems: β = 0.26, *SE* = 0.03, *t*(585.70) = 9.39, $p < .001$; vignettes: β = 0.25, *SE* = 0.03, *t*(575.02)= 8.11, $p < .001$. They showed an average confidence decrease of 9.4 percentage points (*SD* = 18.6) on the problems and of 8.9 percentage points (*SD* = 18.4) on the vignettes. We will refer to this difference as the size of the conflict detection effect (De Neys et al., 2011; Frey et al., 2018). The additional model testing the effects of task format on the smaller sample (see footnote 5) suggested that these conflict effect detection effect sizes did not differ significantly, β = -0.05, *SE* = 0.04, *t*(1272.86) = -1.45, $p = .147$. However, a significant main effect of task format did reveal that participants were significantly more confident about their performance on vignettes than on problems on both conflict and no-conflict tasks, β = 0.08, *SE*= 0.03, *t*(1277.64) = 2.40, $p = .017$. For the conjunction tasks, we also found significant conflict detection effects for both task formats, problems: β = 0.27, *SE* = 0.02, *t*(631.94) = 11.31, $p < .001$; vignettes: β = 0.13, *SE* = 0.03, *t*(535.63) = 4.34, $p < .001$. However, the additional model including task format as predictor showed that the size of these conflict detection effects differed significantly across the two

formats, β = -0.13, SE = 0.03, t(1288.58) = -3.81, p < .001. The average conflict detection effect on conjunction problems was -10.1% (SD = 14.3), whereas it was -5.2 (SD = 16.7) on conjunction vignettes.

Overall, the conflict detection findings on the new tasks in problem-solving format were fully consistent with previous studies using classic heuristics-and-biases tasks (e.g., Frey et al., 2018, who also found significant conflict detection effects with average sizes of -12.3% for base-rate tasks and of -12.5% for conjunction tasks). For the vignette format, we found a similar conflict detection effect on the base-rate vignettes and a smaller but significant effect on the conjunction vignettes.

3.2.2. Response time (s). Results on the response time conflict-detection index were quite consistent across the two task formats and the two bias tasks. To all tasks applied that participants' average response time on incorrectly performed conflict tasks was not significantly longer than on correctly performed no-conflict tasks, base-rate problems: β = -0.01, SE= 0.03, t(558.51)= -0.43, p = .668; base-rate vignettes: β = -0.001, SE= 0.03, t(557.82)= -0.05, p = .960; conjunction problems: β = -0.03, SE = 0.03, t(626.31)= -1.17, p = .243; conjunction vignettes: β = -0.02, SE = 0.03, t(511.78) = -0.54, p = .592. In other words, we found no significant conflict detection effects. The average difference in response times ranged from -0.6 s (SD = 19.9) to 1.7 s (SD = 15.4). The additional models including task format (footnote 5) did not reveal any significant differences in conflict detection across tasks formats, only a main effect of task format for both the base-rate and conjunction tasks: participants took significantly longer to complete the vignettes than the problems (independent of conflict version), base-rate: β = 0.20, SE= 0.03, t(1260.30) = 6.81; p < .001 ; conjunction: β = 0.24, SE = 0.03, t(1281.59) = 8.71, p < .001. Note that this latter finding could be expected given that the vignettes were about twenty words longer than the problems.

These conflict detection results were not in line with previous studies (e.g., Frey et al. 2018, who did find significant conflict detection effects, with an average effect size of 1.3 s and 1.2 s for the base-rate and conjunction tasks, respectively).

3.3. Individual Differences

Next to investigating whether conflict detection takes place at the averaged group level (cf. the analyses above), we also explored potential individual differences in conflict detection. First, we analyzed how many individuals actually showed the conflict detection effect (for a discussion on this, see Frey et al., 2018). Second, we tested whether the size of participants' conflict detection effect correlated with their reasoning accuracy (cf. Mevel et al., 2015; Pennycook et al., 2015). Third, we analyzed whether participants were consistent conflict detectors in the two studied task formats. Below, we summarize the results. The interested reader can find a complete overview of these results in the Supplementary Materials, Section 3.

3.3.1. Number of detectors. First, we analyzed how many of the biased reasoners showed conflict detection. Per conflict-detection index, on each task format of both bias tasks, we tallied the percentage of the biased reasoners showing the conflict detection effect, a reversed conflict detection effect, or no effect (i.e., no difference between conflict indices on conflict and no-conflict trials). Results on the confidence conflict-detection index showed that the vast majority of the biased reasoners showed conflict detection at the individual level too. This was the case for both tasks formats and both bias tasks (between 57.9% and 72.0% of the biased responders). About half of the biased reasoners (between 50.7% and 55.9%) showed conflict detection on the response time conflict-detection index. Interestingly, for the conjunction tasks, we additionally observed a difference between the two task formats with regard to the total number of conflict detectors. The percentage of conflict detectors on the confidence index was lower on the vignette format (57.9%) than on the problem-solving format (72.0%). The average effect size of both detection groups, however, did not differ between the two task formats (16.0% vs. 16.2%).

3.3.2. Accuracy correlations. Some previous studies found correlations between the conflict detection effect size and performance accuracy on conflict problems (Mevel et al., 2015; Pennycook et al., 2015). In line with those previous studies, we also calculated correlations between each individual's conflict detection effect size and their total accuracy on the conflict tasks. The correlation analyses indicated that conflict detectors with larger

confidence and response time effect sizes (i.e., larger difference on these measures between the incorrect conflict and correct no-conflict trials) were more likely to be correct on subsequent conflict tasks in that same block. These effects applied to both the problems and the vignettes of the conjunction tasks. For the base-rate tasks, however, we only found these effects on the vignettes, not on the problems.

3.3.3. Conflict detection consistency. Finally, given the similarity of conflict detection patterns across both task formats, one would expect that individuals who detected conflict on problem-solving tasks would also detect conflict on vignette tasks. To test this assumption, we used cross-tables and counted how many of the biased participants showed conflict detection across both task formats. According to all conflict-detection indices, there was a group of consistent detectors, who showed conflict detection on both the problem-solving tasks and on the vignette tasks, and a relatively small group of consistent non-detectors, who showed no sign of conflict detection in either of the two task formats. Surprisingly, most participants were inconsistent detectors (between 43.5% and 52.0% of biased responders), showing conflict detection on only one of the two task formats. For all conflict-detection indices on both bias tasks, there were more participants who detected conflict on the problem-solving tasks than on vignette tasks, although the differences were small.

4. Discussion

Thus far, research on conflict detection has studied this effect only during reasoning on classic problem-solving tasks, requiring people to make a decision themselves. However, people are also confronted quite often with decisions already made by others, requiring them to correctly evaluate these decisions. The aim of this study was to start investigating the conflict detection effect in a broader range of reasoning scenarios than studied before. To this end, we investigated reasoning accuracy and conflict detection not only in decision-making (problem-solving tasks) but also in decision-evaluation (vignette tasks).

4.1. Accuracy

In line with previously studied classic heuristics-and-biases tasks (e.g., Frey et al., 2018; Pennycook et al., 2015; Raoelison & De Neys, 2019; Thompson et al., 2011), participants performed very well on the no-conflict versions of our tasks and performed quite poorly on the conflict versions. This applied to both bias tasks (base-rate and conjunction) and both task formats (problems and vignettes). Yet, there was an additional effect of task format on reasoning accuracy. No-conflict tasks were performed better when presented in problem-solving format than in vignette format; this applied to both bias task types. Conflict tasks were performed equally well in both task formats of the base-rate tasks, but, for the conjunction tasks, we found that conflict tasks were performed slightly better when presented in vignette format.

The higher performance on the no-conflict problem-solving tasks may simply indicate that participants improved due to the repeated task presentation, as the problem-solving tasks were always performed *after* the vignette tasks. Given that the vignette tasks were always completed *before* the problem-solving tasks, it is not possible that the better performance on these tasks resulted from a general repeated task presentation effect. Hence, this could indicate that, for the conjunction tasks, participants were better at recognizing someone else's biased decision (vignette tasks) than at making an unbiased decision themselves (problem-solving tasks). This would align with the suggestion that people become more deliberative and critical when they have to judge the argumentation of others than when they themselves have to make a judgment (Mercier & Sperber, 2011; Trouche et al., 2016) or have to judge reasoning without specific reference to another person (Mata et al. , 2013). Note, however, that the average difference in correct solution rates between the two task formats was not that large (i.e., 12%).

4.2. Conflict Detection

With regard to conflict detection, the confidence index showed clear and consistent conflict detection effects whereas the response time index did not show any effects. For the confidence index, we found significant conflict detection effects on both base-rate and conjunction tasks in both problem-solving and vignette format. In line with many previous studies (Bago & De Neys, 2017; De Neys et al., 2011, 2013; De Neys & Feremans, 2013; Gangemi et al., 2015; Thompson & Johnson, 2014), participants were, on average, less confident about their incorrect performances on conflict tasks than about their correct performances on no-conflict tasks. For the conjunction tasks, the results additionally showed that the conflict detection effect size on the vignette task format was significantly smaller than on the problem-solving task format. Interestingly, further individual differences analyses revealed that it was not so much the size of the conflict detection effect, but the total percentage of biased participants showing conflict detection, that differed between tasks formats. That is, the confidence effect size for both subgroups of conflict detectors was quite similar, but the percentage of conflict detectors on vignette tasks was smaller than on problem-solving tasks. In other words, fewer participants seemed aware of their errors when evaluating decisions of others than when making decisions themselves. Given that the vignette tasks were presented first, this could imply that participants needed multiple trials before they started to show conflict detection. However, it could also imply that – in addition to a group of participants who become more deliberate when evaluating other people's decisions (cf. accuracy results) – there is another group of reasoners who become less motivated to pay attention to other people's decisions when these do not directly affect them and are in line with their intuitive ideas. The latter implication seems to corroborate with findings by Mata et al. (2013), who found that a subgroup of their participants detected more biases and reasoned better when judging others' responses compared to judging responses without reference to another person. Interestingly, however, another subgroup became worse when judging others' reasoning (Mata et al., 2013). Only participants prone to the bias blind spot, which is tendency to believe that others are more prone to bias than they themselves (Pronin et al., 2002), were better at judging others' reasoning.

Looking at participants' response times, we found no significant conflict detection effects. The lack of response time effects is in stark contrast with many previous studies (Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Trippas, et al., 2014; Stupple & Ball, 2008). The most likely explanation for this seems to lie in the longer, more complex reasoning scenarios used in our tasks. In comparison with classic problems, average response times on the current tasks were very long and the variances were rather large. Hence, subtle differences in task processing could probably not reliably be captured with such response times. A potential solution might be to design shorter tasks or to apply a rapid-response paradigm in which the descriptive information is presented serially to obtain less noisy reasoning time measures (cf. Pennycook et al., 2014). Note, however, that both solutions would render the tasks less similar to real-world reasoning situations, which was also of interest here. Future studies that consider conflict detection with longer or more complex tasks can best rely on confidence measures or investigate other potential measures of conflict detection (e.g., reasoning effort, or process measures obtained through eye-tracking).

4.3. Consistency in Conflict Detection

Taken together, the current results on the confidence measures suggest that conflict detection also occurs during reasoning on longer, more complex, and realistic reasoning tasks than studied before. In addition, the results indicated that the conflict detection effect was very similar during decision-making (problem-solving tasks) and decision-evaluation (vignette tasks), except for the finding that somewhat fewer participants were conflict detectors on the conjunction vignette tasks. The individual differences analyses also pointed to another potential effect of task format. Namely, the results on conflict detection consistency showed that most biased reasoners were inconsistent conflict detectors, that is, they detected conflict in only one of the two task formats. Both conflict-detection indices indicated that slightly more participants detected conflict on the problem-solving

tasks than on the vignette tasks. Although the differences were small, these results could imply that conflict detection is fairly task-format or domain-specific (cf. Frey & De Neys, 2017; Šrol & De Neys, 2019) and that it was slightly more challenging to detect conflict in vignette tasks. Alternatively, it could imply that some people are better conflict detectors during reasoning on their own decisions, whereas other people are better detectors during reasoning on others' decisions. For instance, Mata et al. (2013) found that individual differences in bias blind spot played a role in whether or not it was easier to evaluate another person's reasoning. Future research could investigate whether and how such individual differences play a role in conflict-detection with different task formats.

4.4. Limitations

This study took a first step towards investigating conflict detection in more realistic scenarios and in evaluating other people's decisions. However, some limitations need to be taken in to account. First, since our main interest was to establish whether conflict detection occurs during decision-evaluation, all participants started with the vignette tasks and then completed the problem-solving tasks. Consequently, our findings concerning the direct comparison between the vignette tasks and the problem-solving tasks need to be interpreted with caution as we cannot rule out the effects of task sequence here. Second, there were multiple differences between our two task format conditions, again hindering a direct comparison between the two. Our goal was to make the decision-evaluation tasks as realistic and ecologically valid as possible. This necessarily implied making some changes. For example, the texts in our vignette tasks were somewhat longer because we had to add the description of someone else's decision in each vignette task. In addition, instead of introducing a randomly chosen individual (cf. classical base-rate problems) the base-rate vignettes always explained that one individual case had caught the actor's attention (i.e., more realistic reason to start a decision-making process). In order to draw a more direct comparison between conflict detection during decision-making and decision-evaluation, future studies could randomize the order of the vignette tasks and the problem-solving tasks. Furthermore, one could increase experimental control by reducing the multiple differences between the two task formats (e.g., equal text length, fully similar cover stories etc.). Note, however, that while a focus on minimizing such condition differences may be positive for experimental control, it may not always be fruitful for gaining more insight into real-world reasoning processes. This brings us to a third potential limitation that also applies to the current study. That is, even though they differed in length somewhat, for reasons of experimental rigor all tasks were still similarly structured, well-structured, and included all necessary information to make an adequate probability estimation. In addition, on the vignette tasks, participants' attention was always directed explicitly to the relevant reasoning part (i.e., they were explicitly asked whether a specific estimation in the text was correct or not). It would be fruitful for future research to start addressing conflict-detection indices in (gradually) more realistic (and therefore less structured) reasoning contexts (e.g., evaluating decisions of two people engaged in a dialogue without pointing to the relevant reasoning parts explicitly).

4.5. Conclusion

In conclusion, the present study suggests that conflict detection also occurs on longer, more complex reasoning tasks than the classic heuristic-and-biases problems studied before. Moreover, conflict detection occurs not only when making a decision oneself, but also when evaluating decisions of others (as described in a text). This is relevant because there are many everyday situations in which we are confronted with biased conclusions or decisions made by others and have to evaluate or form our own opinion on those decisions. Hence, these findings indicate that even if people may not detect biased reasoning in decisions of others, they often do show signs of conflict detection. The current findings are very relevant for studying reasoning in contexts in which recognizing errors is important; for instance, in medicine, where doctors often have to evaluate initial diagnoses of others, or in education, where teachers have to detect and give feedback on biases in their students' reasoning.

Even though people may err when evaluating others' reasoning, there seems to be some error or conflict detection going on. One may envisage how future training could try to build on the currently demonstrated error signal to de-bias people's evaluative reasoning.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*. https://doi.org/10/gfz97m

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*, 186–196. https://doi.org/10.3758/MC.38.2.186

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*, 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*, 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W., & Bonnefon, J.-F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*, 172–178. https://doi.org/10.1016/j.tics.2013.02.001

De Neys, W., Cromheeke, S., & Osman, M. (2011). Correction: Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*. https://doi.org/10.1371/annotation/1ebd8050-5513-426f-8399-201773755683

De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, *49*, 258–269. https://doi.org/10.1037/a0028320

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*, 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*, 503–509. https://doi.org/10.1177/0963721419855658

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269–273. https://doi.org/10.3758/s13423-013-0384-5

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Psychology.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241. https://doi.org/10.1177/1745691612460685

Ferreira, M. B., Mata, A., Donkin, C., Sherman, S. J., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Memory & Cognition*, *44*, 1050–1063. https://doi.org/10.3758/s13421-016-0618-7

Frey, D., & De Neys, W. (2017). Is conflict detection in reasoning domain general? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *39*, 391–396. https://pdfs.semanticscholar.org/aaec/4079baaeba9ba75c6cf816874e5cc2b9a201.pdf

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, 1–52. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*, 383–396. https://doi.org/10.1080/13546783.2014.980755

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 513–525. https://doi.org/10.1037/0096-1523.14.3.513

Janssen, E. M., Mainhard, T., Buisman, R. S. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Van Peppen, L. M., & Van Gog, T. (2019). Training higher education teachers' critical thinking and attitudes towards teaching it. *Contemporary Educational Psychology*, *58*, 310–322. https://doi.org/10.1016/j.cedpsych.2019.03.007

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2011). *Thinking, fast and slow*. Lane.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, *17*, 11–17. https://doi.org/10.3758/BF03199552

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*, 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*, 486–491. https://doi.org/10.1016/j.jesp.2013.01.010

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

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*, 227–237. https://doi.org/10.1080/20445911.2014.986487

Oster, N., & Koesterich, R. (2013). Breaking bad behaviors: Understanding investing biases and how to overcome them. *IShares Market Perspectives*, 1–9.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*, 1–10. https://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*, 101–106. https://doi.org/10.1016/j.cognition.2012.04.004

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 544–554. https://doi.org/10.1037/a0034887

Politzer, G., Bosc-Miné, C., & Sander, E. (2017). Preadolescents solve natural syllogisms proficiently. *Cognitive Science*, *41*, 1031–1061. https://doi.org/10.1111/cogs.12365

Prado, J., Léone, J., Epinat-Duclos, J., Trouche, E., & Mercier, H. (2020). The neural bases of argumentative reasoning. *Brain and Language*, *208*, 104827. https://doi.org/10.1016/j.bandl.2020.104827

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. https://doi.org/10.1177/0146167202286008

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*, 178–178.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*, 93–114. https://doi.org/10.1007/s11251-014-9328-3

Schmidt, H. G., Mamede, S., Van den Berge, K., Van Gog, T., Van Saase, J. L. C. M., & Rikers, R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine*, *89*, 285–291. https://doi.org/10.1097/ACM.0000000000000107

Smith, T. C. (2017). Vaccine rejection and hesitancy: A review and call to action. *Open Forum Infectious Diseases*, *4*(3). https://doi.org/10.1093/ofid/ofx146

Šrol, J., & De Neys, W. (2019). *Predicting individual differences in conflict detection and bias susceptibility during reasoning* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/2uf6g

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Stupple, E. J. N., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning*, *14*, 168–181. https://doi.org/10.1080/13546780701739782

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*, 54–77. https://doi.org/10.1080/13546783.2012.735622

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

Thompson, V A, & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*, 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, Valerie A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, *11*, 167–187. https://doi.org/10.1007/BF01044641

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2016). The selective laziness of reasoning. *Cognitive Science*, *40*, 2122–2136. https://doi.org/10.1111/cogs.12303

Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*, 293–315. https://doi.org/10.1016/B978-1-4832-1446-7.50038-8

Van den Berge, K., Mamede, S., Van Gog, T., Romijn, J. A., Van Guldener, C., Van Saase, J. L., & Rikers, R. M. (2012). Accepting diagnostic suggestions by residents: A potential cause of diagnostic error in medicine. *Teaching and Learning in Medicine*, *24*, 149–154. https://doi.org/10.1080/10401334.2012.664970

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*, 833–839. https://doi.org/10.1002/acp.2883

Van Kelecom, K. (2017, February). *We zeggen te vaak sorry (én we doen het verkeerd)* [Taalblad.be]. https://web.archive.org/web/20181209123856/http:/taalblad.be/e-zine/kort-en-klein/we-zeggen-te-vaak-sorry-en-we-doen-het-verkeerd-/1314.html

Supplementary Materials

1. Classical conjunction problem

An example of a classical conjunction problem (see e.g., De Neys et al., 2011; Frey et al., 2018):

> Jon is 32. He is intelligent and punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in languages and art.
>
> Which one of the following statements is most likely?
>
> Jon plays in a rock band
> Jon plays in a rock band and is an accountant

Because Jon's description is more representative of an accountant, most people incorrectly indicate that it is more likely that Jon plays in a rock band and is accountant than that Jon is a rock band player only. However, the conjunction of any two probabilities is always less likely than either of the conjuncts in isolation.

2. Item-level check

In order to check whether the content of the items' cover stories influenced the accuracy of the items, we conducted a logistic mixed effect model, with response accuracy as dependent variable (incorrect = 0; incorrect =1), item-number content as fixed effect and task format (problem solving = 0; vignette = 1), conflict version (conflict = 0; no conflict = 1), and participants' identification as random effects. Results of the model for the base-rate tasks, indicated that – after controlling for variance due to task format or conflict version – two items were performed significantly better compared to the other test items, $B = 0.61$ $SE = 0.28$, $W = 2.31$, $p = .021$; $B = 0.61$ $SE = 0.28$, $W = 2.17$, $p = .030$. However, post-hoc pairwise comparisons between all items did not yield significant accuracy differences between any of the items (*p*s ≥ .144). Thus, base-rate accuracy did not seem to depend on the content of the cover stories.

For the conjunction items, we found that six items were performed significantly better compared to the other test items (*ps* ≤ .019), and that one item was performed significantly worse, $B = 0.70$ $SE = 0.28$, $W = -2.52$, $p = .012$. Post-hoc pairwise comparisons between all items showed no significant accuracy differences between items (*ps* ≥ .095), except for the one item that was performed worse compared to other items (*ps* ≤ .010). Based on the content of the cover story, we could not directly explain the obtained accuracy difference. Because we only found one potentially different item, we decided not to exclude any items. Note that we counterbalanced the content of the reasoning tasks across formats and conflict versions and always included item-content number as random effect in our statistical models to account for potential variance due to the item's content. Hence, a difference in one item is unlikely to affect our conclusions.

3. Supplementary results

3.1. Conflict detection indicated by participants' confidence response time. Table S1 (extended version of Table 3 in the manuscript) provides an overview of the average scores on all three conflict-detection indices for correctly and incorrectly performed trials. Results on the confidence and response time index are reported in the manuscript. Below, we additionally report the results on the unreliable confidence response time index.

Overall, the conflict detection results on the confidence response time index were mixed. On both task formats of the base-rate tasks, participants took on average slightly longer to enter their confidence judgements for incorrectly performed conflict tasks than to enter their confidence judgements for correctly performed no-conflict tasks. However, this difference was only significant for the problems and not for the vignettes, problems: $\beta = -0.11$, $SE = 0.03$, $t(594.44) = -3.80$, $p < .001$; vignettes: $\beta = 0.05$, $SE = 0.03$, $t(598.20) = -1.65$, $p = .099$. In other words, we found a significant conflict detection effect on the base-rate problems, but only a trend towards detection on the base-rate vignettes. The average increase in confidence response times was only 0.3 s ($SD = 1.3$) and 0.2 ($SD = 1.6$) for the problems and vignettes, respectively. For the conjunction tasks, on the other hand, we found a reversed conflict detection pattern. That is, participants took slightly longer to enter their confidence judgements on their correctly performed no-conflict tasks than on their incorrectly performed conflict tasks. This reversed effect was not significant on the problems but it was significant on the vignettes, problems: $\beta = 0.02$, $SE = 0.03$, $t(651.81) = 0.63$, $p = .527$ (average difference -0.1 s, $SD = 1.0$); vignettes: $\beta = 0.08$, $SE = 0.03$, $t(545.98) = 2.52$, $p = .012$ (average effect size of -0.3s, $SD = 2.8$). The additional models did not yield any significant effects of task format on participants' confidence response times for either the base-rate tasks or the conjunction tasks.

These mixed findings were more or less in line with a previous study by Frey et al. (2018), who found reversed (non-significant) trends with very small effect sizes of -0.3 s and -0.2 s for base-rate and conjunction tasks, respectively.

Table S1: Group-Level Averages (SD) on Each of the Three Conflict Detection Indices as a Function of Response Accuracy

Conflict detection index	Conflict: correct	Conflict: incorrect	No-conflict: correct	No-conflict: incorrect
Base-rate problems				
Participants by group	*n* = 104	*n* =135	*n* = 159	*n* = 43
Average confidence (%)	66.7 (17.8)	66.8 (18.1)	77.1 (15.4)	63.5 (18.8)
Average RT (s)	39.9 (21.9)	38.3 (25.0)	36.5 (15.2)	65.6 (14.8)
Average confidence RT (s)	2.2 (1.5)	2.2 (1.9)	1.9 (1.2)	1.8 (1.5)
Base-rate vignettes				
Participants by group	*n* = 102	*n* = 147	*n* = 156	*n* = 64
Average confidence (%)	64.6 (17.5)	68.9 (17.6)	77.7 (13.8)	62.9 (19.1)
Average RT (s)	49.8 (24.1)	47.2 (23.3)	24.2 (0.7)	26.2 (0.1)
Average confidence RT (s)	2.5 (1.7)	2.3 (1.5)	1.3 (0.7)	2.4 (2.2)
Conjunction problems				
Participants by group	*n* = 78	*n* = 145	*n* = 157	*n* = 45
Average confidence (%)	69.1 (16.8)	70.7 (16.7)	81.7 (14.3)	67.8 (17.9)
Average RT (s)	27.4 (25.0)	24.0 (20.9)	22 (16.5)	25.9 (16.6)
Average confidence RT (s)	2.0 (1.3)	1.8 (0.8)	2.0 (1.1)	2.1 (1.8)
Conjunction vignettes				
Participants by group	*n* = 112	*n* = 149	*n* = 155	*n* = 88
Average confidence (%)	58.6 (19.4)	65.4 (17.0)	71.0 (16.8)	65.5 (16.8)
Average RT (s)	30.3 (16.6)	29.1 (15.3)	28.3 (17.6)	31.5 (19.0)
Average confidence RT (s)	2.0 (1.2)	2.1 (1.4)	2.4 (2.9)	1.8 (1.2)

3.2. Individual differences. Next to investigating whether conflict detection takes place at the group level, we explored potential individual differences in conflict detection and tested the consistency of conflict detection across task formats.

3.2.1. Base-rate tasks. We first report all results for the base-rate tasks and then for the conjunction tasks.

Number of detectors. We analyzed how many of the biased reasoners showed conflict detection. For both task formats, we tallied per conflict detection index which percentage of the biased reasoners showed the conflict detection effect, a reversed conflict detection effect, or no effect (i.e., same; no difference between conflict indices on conflict and no-conflict trials). The results are shown in Table S1 (top two panels).

Confidence (%). For the problem-solving tasks, we found that 65.9% of the biased reasoners showed conflict detection as indexed by their confidence ratings, with an average effect size of –18.1% (*SD* = 16.2). This was rather similar on the vignette tasks, with the majority (65.3%) of the biased reasoners showing conflict detection, with an average effect size of -17.5% (*SD* =16.6). Both the percentage of conflict detectors and the size of the effect were comparable to previous findings of Frey et al. (2018) on classical base-rate problems, who found that 72% of the biased participants were conflict detectors with a confidence effect size of -20.0%.

Response time (s). For the response time index, we found that 51.1% of the biased participants showed conflict detection on the problem-solving tasks, with an average conflict detection effect size of 9.9 s (*SD* = 18.7). Likewise, 50.7% of the biased participants showed conflict detection on the vignette tasks, with an average size of

12.5 s (*SD* = 15.8). These findings deviated from Frey et al. (2018) who found a slightly larger group of conflict detectors (64%), with a smaller effect size (*M* = 4.2 s).

Confidence response time (s). For the problem-solving tasks, we found that 60.7% of the biased participants showed conflict detection on the confidence response time index, with an average effect size 0.9 s (*SD* =1.4). For the vignette tasks we found that 54.9% showed conflict detection, with an effect size of 1.1 s (*SD* = 1.5). These findings were more or less in line with Frey et al. (2018), who found that somewhat smaller group of biased participants showed conflict detection (43%), yet with a comparable effect size (*M* =1.3 s).

In sum, even when taking individual differences into account, we found very similar conflict detection patterns across both base-rate task formats (problem-solving tasks and vignette tasks). Furthermore, apart from some small differences, we found similar conflict detection patterns on the current, more realistic versions of base-rate tasks as Frey et al. (2018) found with classical base-rate problems.

Accuracy correlations. We calculated the correlation between each individual's conflict detection effect size on the three conflict detection indices and their total accuracy on the conflict tasks. The results are also included in Table S1 (top two panels). Results showed only one significant correlation for the base-rate problem-solving tasks. Within the reversed detection group, a larger response time effect size was related to lower accuracy on conflict problems, r = -0.32, p = .009. Hence, the longer participants took to answer correctly solved no-conflict problems (relative to their incorrectly solved conflict problems), the more likely it was that they solved other conflict problems incorrectly. For the vignette tasks, the analyses yielded two significant correlations, both within the detection group. A larger confidence effect size, r = -0.24, p = .022, and a larger response-time effect size, r = 0.30, p =.009, was related to higher total accuracy on conflict vignettes. Hence, a lower confidence and longer response time on incorrectly evaluated decisions on conflict vignettes (relative to correctly evaluated decision on no-conflict vignettes) was associated with better accuracy on other conflict vignettes. For the remaining conflict detection indices and subgroups the results are less clear, with only small and non-significant correlations. Hence, in these cases there is no clear evidence that the size of the conflict detection effect reflects individual differences in the quality of the detection process among biased reasoners.

In sum, for the vignette tasks we found indications that the size of the conflict detection effects reflected individual differences in the quality of the detection process among biased reasoners. In line with Frey et al. (2018) for classical base-rate problems, we obtained significant correlations on the confidence and response time indices. For problem-solving tasks, on the other hand, we found no such correlations.

Table S2: Individual-Level Findings for Different Subgroups of Biased Reasoners and for the Whole Group of Biased Reasoners

	Subgroup: Conflict detection	Subgroup: Reversed detection	Subgroup: Same	Whole biased group
Base-rate problems				
Confidence				
% of biased group	65.9 (*n* = 89)	25.9 (*n* = 35)	8.1 (*n* = 11)	100 (*n* = 135)
Confidence effect size (*SD*)	-18.1 (16.2)	9.9 (8.6)	-	-9.4 (18.6)
Accuracy correlation *r* (*p*)	-0.12 (.247)	0.19 (.273)	-	0.04 (.678)
Response time				
% of biased group	51.1 (*n* =69)	48.9 (*n* = 66)	-	100 (*n* =135)
Response time effect size (*SD*)	9.9 (18.7)	-8.8 (13.0)	-	0.7 (18.6)
Accuracy correlation *r* (*p*)	0.23 (.054)	-0.32 (.009)		.008 (.924)
Confidence response time				
% of biased group	61.5 (*n* = 83)	38.5 (*n* = 52)	-	100 (*n* = 135)

Confidence RT effect size (*SD*)	0.9 (1.4)	-0.5 (0.6)	-	0.3 (1.3)
Accuracy correlation *r* (*p*)	0.10 (.386)	-0.05 (.704)		-0.02 (.849)
Base-rate vignettes				
Confidence				
% of biased group	65.3 (*n* = 94)	25.7 (*n* = 37)	9 .0 (*n* = 13)	100 (*n* = 144)
Confidence effect size (*SD*)	-17.5 (16.6)	9.8 (7.5)	0	-8.9 (18.4)
Accuracy correlation *r* (*p*)	-0.24 (.022)	0.19 (.267)		-0.11 (.204)
Response time				
% of biased group	50.7 (*n* = 70)	49.3 (*n* = 74)	-	100 (*n* = 144)
Response time effect size (*SD*)	12.5 (15.8)	-12.9 (15.0)	-	-0.6 (19.9)
Accuracy correlation *r* (*p*)	0.30 (.009)	-0.21 (.085)		0.11 (.175)
Confidence response time				
% of biased group	51.4 (*n* = 74)	47.9 (*n* = 69)	0.7 (*n* = 1)	100 (*n* = 144)
Confidence RT effect size (*SD*)	1.1 (1.5)	-0.7 (1.0)	0	0.2 (1.6)
Accuracy correlation *r* (*p*)	0.08 (.474)	-0.08 (.548)		0.10 (.240)
Conjunction problems				
Confidence				
% of biased group	72.0 (*n* = 103)	22.4 (*n* = 32)	5.6 (*n* = 8)	100 (*n* = 143)
Confidence effect size (*SD*)	-16.0 (12.3)	6.4 (4.8)	0	-10.1 (14.3)
Accuracy correlation *r* (*p*)	-0.35 (<.001)	0.09 (.610)		-0.30 (<.001)
Response time				
% of biased group	55.9 (*n* = 80)	44.1 (*n* = 63)	-	100 (*n* = 143)
Response time effect size (*SD*)	9.9 (14.1)	-8.7 (9.7)	-	1.7 (15.4)
Accuracy correlation *r* (*p*)	0.30 (.008)	0.01 (.925)		0.24 (.004)
Confidence response time				
% of biased group	44.8 (*n* = 62)	55.2 (*n* = 81)	-	100 (*n* = 143)
Confidence RT effect size (*SD*)	0.6 (0.7)	-0.7 (0.9)	-	-0.1 (1.0)
Accuracy correlation *r* (*p*)	0.16 (.201)	0.08 (.490)		0.04 (.600)
Conjunction vignettes				
Confidence				
% of biased group	57.9 (*n* = 84)	37.2 (*n* = 54)	4.8 (*n* = 7)	100 (*n* = 145)
Confidence effect size (*SD*)	-16.2 (11.9)	11.1 (8.4)		-5.2 (16.7)
Accuracy correlation *r* (*p*)	-0.29 (.008)	0.28 (.041)		-0.16 (.055)
Response time				
% of biased group	54.5 (*n* = 79)	45.5 (*n* = 66)	-	100 (*n* = 145)
Response time effect size (*SD*)	9.2 (7.3)	-10.3 (16.8)		0.4 (15.8)
Accuracy correlation *r* (*p*)	0.07 (.547)	-0.13 (.291)		0.04 (.606)
Confidence response time				
% of biased group	41.4 (*n* = 60)	58.6 (*n* = 85)	-	100 (*n* = 145)
Confidence RT effect size (*SD*)	0.8 (1.2)	-1.1 (3.4)		-0.3 (2.8)
Accuracy correlation *r* (*p*)	0.29 (.022)	0.02 (.886)		0.17 (.047)

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

Conflict detection consistency. Given the similarity of conflict detection patterns across both task formats, one would expect that individuals who detected conflict in problem-solving tasks, would also detect conflict in vignette tasks. Therefore, we used cross-tables to count how many of the biased participants showed conflict detection across both task formats. Results are shown in Tables S2-S4.

Confidence. Table S2 shows that 78.6% of all participants (125 out of 159) were consistent biased reasoners (i.e., entering at least one biased response on one of the three conflict tasks in both task formats). Of these consistent biased reasoners, 42.4% was also a consistent conflict detector, as indexed by confidence. That is, this group showed conflict detection in both task formats. There were also two groups of inconsistent detectors: first, 23.2% of the biased reasoners were conflict detectors only in the problem-solving format; second, 20.8% detected conflict, only in the vignette format. The remaining 13.6% were consistent non-detectors.

Response time. For the response time index (Table S3), 22.4% of the consistent biased reasoners (n =125), were also consistent conflict detectors (i.e., increased response time) in both problem-solving and vignette tasks. Furthermore, 28.8% detected conflict only in the problem-solving tasks, and 23.2% detected conflict only in vignette tasks. The remaining 25.6% were consistent non-detectors.

Confidence response time. Finally, 32.0% of the consistent biased reasoners (n = 125) showed conflict detection on the confidence response time index (i.e., increased confidence response time) in both task formats (Table S4). Furthermore, 28.8% showed conflict detection in the problem-solving tasks only, and 17.6% in the vignette tasks only. The other 21.6% did not detect conflict in any task format.

In sum, all three conflict detection indices indicated there was a group of consistent conflict detectors and a group of consistent non-detectors. The confidence index yielded the largest group of consistent detectors, followed by the confidence response time, and response time indices, respectively. There were also two groups of inconsistent detectors, which showed conflict detection in only one of the two task formats. All three indices indicated that more participants detected conflict in the problem-solving tasks (i.e., decision-making) than in the vignette tasks (i.e., decision-evaluation), although it were small differences.

Table S3: Cross Table Showing the Number of Individuals Who Detected Conflict on the Base-Rate Tasks across both Task Formats, as Indexed by Confidence

	Vignettes			
	Detection	Reverse	Same	All correct
Problems				
Detection	53	21	8	7
Reverse	17	12	3	3
Same	9	1	1	0
All correct	15	3	1	5

Table S4: Cross Table Showing the Number of Individuals Who Detected Conflict on the Base-Rate Tasks across both Task Formats, as Indexed by Response Time

	Vignettes		
	Detection	Reverse	All correct
Problems			
Detection	28	36	5
Reverse	29	32	5
All correct	13	6	5

Table S5: Cross Table Showing the Number of Individuals Who Detected Conflict on the Base-Rate Tasks across both Task Formats, as Indexed by Confidence Response Time

	Vignettes			
	Detection	Reverse	Same	All correct
Problems				
Detection	40	36	0	7
Reverse	22	27	0	3
All correct	12	6	1	5

3.2.2. Conjunction tasks. We conducted the same individual difference analyses for the conjunction tasks.

Number of detectors. As for the base-rate tasks, we tallied per conflict detection index how many of the biased reasoners showed the conflict detection effect. Results are shown in Table S1 (bottom two panels).

Confidence (%). For the problem-solving tasks, we found that 72.0% of the biased reasoners showed conflict detection on the confidence index, with an average effect size of -16.0% (*SD* =12.3). For the vignette tasks, we found that a somewhat smaller majority of the biased reasoners (57.9%) showed conflict detection, yet with a similar effect size (-16.2%, *SD* =11.9). The percentage of conflict detectors on the problem-solving tasks was in line with Frey et al. (2018),who found a percentage of 79%, yet the effect sizes in both task formats were smaller than the effect size (- 27.6%) in Frey et al. (2018).

Response time (s). For the response time index, we found that 55.9% of the biased participants showed conflict detection on the problem-solving tasks, with an average detection effect size of 9.9 s (*SD* =14.1). Similarly, 54.5% of the biased participants showed conflict detection on the vignette tasks, with an average size of 9.2 s (*SD* = 7.3). These findings deviated from Frey et al., (2018) who found a larger group of conflict detectors (71%), with a smaller effect size (3.0 s).

Confidence response time (s). For the problem-solving tasks, we found that 44.8% of the biased participants showed conflict detection on the confidence response time index, with an average effect size 0.6 s (*SD* = 0.7). Likewise, 42.1% of the biased participants showed conflict detection on the vignette tasks with an average effect size of 0.8 s (*SD* = 1.2). These findings were in line with Frey et al. (2018), who found that 48% of the biased participants showed conflict detection with a comparable effect size 1.1 s.

In sum, also when taking individual differences into account, we found very similar conflict detection patterns for both problem-solving (i.e., decision-making) and vignette (i.e., decision-evaluation) tasks. Furthermore, apart from the response time index, the results on the current, more realistic versions of the conjunction tasks were very similar to what Frey et al. (2018) obtained on classical conjunction problems.

Accuracy correlations. We calculated the correlation between individuals' conflict detection effect size on the three conflict detection indices and their total accuracy on the conflict tasks (see Table S1, bottom two panels). The analyses yielded four significant correlations for the problem-solving tasks. Within the conflict detection group (and also in the whole biased group), a larger confidence effect size, r = -0.35, p < .001, and a larger response time effect size, r = 0.30, p =.008, was related to higher accuracy on conflict problems. Hence, a lower confidence and longer response time on incorrectly solved conflict problems (relative to correctly solved no-conflict problems) was associated with higher accuracy on other conflict problems. For the vignette tasks, we obtained three significant correlations. Within the conflict detection group, a larger confidence effect size, r = -0.29, p =.008, and confidence response time effect size, r = 0.29, p =.022, was related to higher accuracy on conflict vignettes. Within the reversed detection group, a larger reversed effect size was related to lower accuracy, r = 0.28, p =.041.

In sum, the correlations indicated that, for both for tasks formats, the size of the conflict detection effects could reflect individual differences in the quality of the detection process among biased reasoners. This was in

contrast to Frey et al. (2018), who, for the classical conjunction problems, obtained only one significant correlation (on the response time index) in this direction.

Conflict detection consistency. We used cross-tables to test to what extent reasoners were consistent conflict detectors across the two task formats (Tables S5-S7).

Confidence. Of all participants (n = 159), 82.4% (n = 131) responded biased at least once to the conflict tasks in both task formats (Table S5). Of these consistent biased reasoners, 43.5% was also a consistent conflict detector in both task formats, as indicated by confidence ratings. There were again two groups of inconsistent detectors: 29.0% detected conflict in the problem-solving tasks only, and 16.8% detected conflict in the vignette tasks only. The remaining 10.7% were consistent non-detectors.

Response time. For the response time index (Table S6), 33.6% of the consistent biased reasoners (n =131), was also consistent conflict detector in both task formats. Furthermore, 22.9% only detected conflict on the problem-solving tasks, whereas 20.6% only detected conflict on the vignette tasks. The remaining 22.9% were consistent non-detectors.

Confidence response time. Finally, 17.6% of the consistent biased reasoners (n =131). showed conflict detection on the confidence response time index in both task formats, 26.7% of the biased reasoners only showed conflict detection in the problem-solving format, and 22.1% only in the vignette format (Table S7). The final 33.6% did not detect conflict in any task format.

In sum, all three conflict detection indices showed there was one group of consistent conflict detectors and a group of consistent non-detectors. The confidence index yielded the largest group of consistent conflict detectors, followed by the response time index and confidence response time index, respectively. There was also a relatively large group of inconsistent detectors, meaning that they detected conflict either in the problem-solving format or in the vignette format. All three indices pointed out that slightly more participants detected conflict in the problem-solving format (i.e., decision-making) than in the vignette format (i.e., decision-evaluation).

Table S6: Cross Table Showing the Number of Individuals Who Detected Conflict on the Conjunction Tasks across both Task Formats, as Indexed by Confidence

	Vignettes			
	Detection	Reverse	Same	All correct
Problems				
Detection	57	35	3	8
Reverse	16	11	1	4
Same	6	1	1	0
All correct	5	7	2	2

Table S7: Cross Table Showing the Number of Individuals Who Detected Conflict on the Conjunction Tasks across both Task Formats, as Indexed by Response Time

	Vignettes		
	Detection	Reverse	All correct
Problems			
Detection	44	30	6
Reverse	27	30	6
All correct	8	6	2

Table S8: Cross Table Showing the Number of Individuals Who Detected Conflict on the Conjunction Tasks across both Task Formats, as Indexed by Confidence Response Time

	Vignettes		
	Detection	Reverse	All correct
Problems			
Detection	23	35	4
Reverse	29	44	8
All correct	8	6	2